System Reliability Theory

System Reliability Theory

Models, Statistical Methods, and Applications

Third Edition

Marvin Rausand
Norwegian University of Science & Technology
Trondheim, Norway

Anne Barros
CentraleSupélec, Paris-Saclay University
Paris, France

Arnljot Høyland[†]
Norwegian University of Science & Technology
Trondheim, Norway

WILEY

Edition History
John Wiley & Sons, Inc. (1e, 1994)
John Wiley & Sons, Inc. (2e, 2004)

Registered Office
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office
111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty
While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data

Names: Rausand, Marvin, author. | Barros, Anne, author. | Høyland, Arnljot,
1924- author.
Title: System reliability theory : models, statistical methods, and
applications / Marvin Rausand, Norwegian University of Science &
Technology, Trondheim, Norway, Anne Barros, CentraleSupélec, Paris-Saclay University,
Paris, France, Arnljot Høyland, Norwegian
University of Science & Technology, Trondheim, Norway.
Description: Third edition. | Hoboken, NJ : John Wiley & Sons, Inc., 2021.
| Series: Wiley series in probability and statistics | Includes
bibliographical references and index.
Identifiers: LCCN 2020016182 (print) | LCCN 2020016183 (ebook) | ISBN
9781119373520 (cloth) | ISBN 9781119374015 (adobe pdf) | ISBN
9781119373957 (epub)
Subjects: LCSH: Reliability (Engineering)–Statistical methods.
Classification: LCC TA169 .H68 2021 (print) | LCC TA169 (ebook) | DDC
620/.00452–dc23
LC record available at https://lccn.loc.gov/2020016182
LC ebook record available at https://lccn.loc.gov/2020016183

Cover Design: Wiley
Cover Images: System reliability theory Courtesy of Marvin Rausand,
Abstract purple and pink polygonal © Tuomas Lehtinen/Getty Images

Set in 9.5/12.5pt STIXTwoText by SPi Global, Chennai, India

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To Hella; Guro and Idunn; and Emil and Tiril
To: Nicolas; Penelope; and Garance

Contents

Preface

This book provides a basic, but rather comprehensive introduction to system reliability theory and the main methods used in reliability analyses. System reliability theory is used in many application areas. Some of these are illustrated in the book as examples and problems.

Main Changes from the Second Edition

Readers who are familiar with the second edition (Rausand and Høyland 2004) will find that the third edition is a major update and that most chapters have been rewritten. The most significant changes include:

- A new Chapter 2 defining the study object and its functions and operating context is included. System modeling by reliability block diagrams is introduced and the concept of complexity is discussed.
- A new Chapter 3 defining and discussing the concepts of failure and fault, together with several associated concepts is added. Two failure analysis techniques are presented.
- New component importance metrics are included.
- The treatment of dependent failures is significantly extended.
- Section 8.8 on complex systems in the second edition is removed from the chapter on Markov analysis where several new models are added.
- A new Chapter 12 on preventive maintenance is added. This chapter merges aspects from the previous edition with new models and methods. The presentation is supplemented by Python scripts that are found on the `book companion site`.
- Chapters 11 and 13 in the second edition on life data analysis and Bayesian reliability analysis are totally rewritten. The statistical program system R is extensively used in the presentation.

- Chapter 12 in the second edition on accelerated testing has been removed, but parts of the chapter are moved to the chapter on reliability data analysis.
- The end of chapter problems have been revised and new problems are added.
- Most of the appendices are removed. The content is partly integrated in the text and partly obsolete because of the use of R.
- An author index is provided.

Supplementary Information on the Internet

An immense amount of relevant information is today available on the Internet, and many of the topics in this book may be found as books, reports, lecture notes, or slides written by lecturers from many different universities. The quality of this information is varying and ranging from very high to rather low, the terminology is often not consistent, and it may sometimes be a challenge to read some of these Internet resources. The reader is encouraged to search the Internet for alternative presentations and compare with the book. This way, new ideas and increased insight may spring up.

With the abundance of free information on the Internet, it is pertinent to ask whether a traditional book is really needed. We strongly believe that a book may provide a more coherent knowledge and we have tried to write the book with this in mind.

Intended Audience

The book is written primarily for engineers and engineering students, and the examples and applications are related to technical systems. There are three groups that constitute our primary audience:

- The book was originally written as a textbook for university courses in system reliability at the Norwegian University of Science and Technology (NTNU) in Trondheim. This third edition is based on experience gained from use of the first two editions, at NTNU and many other universities, and also from using the book in a wide range of short courses for industry.
- The second is to be a guide for engineers and consultants who carry out practical system reliability analyses of technical systems.
- The third is to be a guide for engineers and consultants in areas where reliability is an important aspect. Such areas include risk assessment, systems engineering, maintenance planning and optimization, logistics, warranty engineering and

management, life cycle costing, quality engineering, and several more. It may be noted that several of the methods used in artificial intelligence and machine learning are treated in this book.

Readers should have a basic course in probability theory. If not, you should get hold of an introductory textbook in probability and statistics to study in parallel with reading this book. A multitude of relevant lecture notes, slides, and reports are also available on the Internet. Brief guidance to relevant sources is provided on the `book companion site`.

Aims and Delimitation

The book is intended to give a thorough introduction to system reliability. Detailed objectives and associated delimitations are found in Section 1.8. The study object may range from a single component up to a rather complicated technical system. The study object is delimited to items that are mainly based on mechanical, electrical, or electronic technology. An increasing number of modern items have a lot of embedded software. Functions that earlier were carried out by mechanical and electromechanical technology are today software-based functions. A family car that was built when the second edition was published is, for example, very different from a modern car, which is sometimes characterized as a "computer on wheels." Software reliability is different from hardware reliability in many ways and we, therefore, consider pure software reliability to be outside the scope of the book. Many software-based functions may, however, be treated with the methods presented.

Many modern systems are getting more and more complex. Chapter 2 introduces three categories of systems: simple, complicated, and complex systems. Complex systems are here defined to be systems that do not meet all the requirements of the Newtonian–Cartesian paradigm and therefore cannot be adequately analyzed with traditional methods. The complexity theory and the approaches to study complex systems is considered to be outside the scope of the book.

The objective of this book is to help the reader to *understand* the basic theory of system reliability and to become familiar with the most commonly used analytical methods. We have focused on producing reliability results by hand-calculation, sometimes assisted by simple R and Python programs. When you carry out practical reliability analyses of large systems, you usually need some special computer programs, such as fault tree analysis programs and simulation programs. A high number of programs are available on the market. We do not present any of these special programs in the book, but supply a list of the main vendors of such programs on the `book companion site`. To use a specific program, you need to study the user manual. This book should help you understand the content of such manuals and the sources of uncertainty of the results produced.

A wide range of theories and methods have been developed for system reliability analysis. All these cannot be covered in an introductory text. When selecting material to cover, we have focused on methods that:

- Are commonly used in industry or in other relevant application areas
- Give the analyst insights that increase her understanding of the system (such that system weaknesses can be identified at an early stage of the analysis)
- Provide the analyst with genuine insight into system behavior
- Can be used for hand-calculation (at least for small systems)
- Can be explained rather easily to, and be understood by nonreliability engineers and managers.

The authors have mainly been engaged in applications related to the offshore oil and gas industry and many examples therefore come from this industry. The methods described and many of the examples are equally suitable for other industries and application areas.

Authors

The first edition of the book (Høyland and Rausand 1994) was written with joint efforts from Arnljot Høyland and Marvin Rausand. Arnljot sorrily passed away in 2002. The second edition (Rausand and Høyland 2004), was therefore prepared by Marvin alone and represented a major update of the first edition. Marvin retired from his professorship at NTNU in 2015 and when Wiley wanted an updated version, he asked Anne Barros to help preparing this third edition. Due to unforeseen practical constraints, Anne could not devote as much time to this project as she wanted. Anne's contribution to this edition is mainly related to Chapters 11 and 12, the end of chapter problems, in addition to reviewing and proposing improvements to other chapters.

Acknowledgments

First of all, we express our deepest thanks to Professor Arnljot Høyland. Professor Høyland passed away in December 2002, 78 years old, and could not participate in writing any further editions of the book. We hope that he would have approved and appreciated the changes and additions we have made.

The authors sincerely thank a high number of students at NTNU, and lecturers and students at many other universities around the world for comments to the previous edition and for suggesting improvements. We have done our best to implement these suggestions. Special thanks go to Professor Bruno Castanier, Université d'Angers, for making significant contributions to Section 12.3, and to Per Hokstad, SINTEF, for many inputs to Chapter 8.

Many definitions used in the book are from, or are inspired by, the International Electrotechnical Vocabulary (IEV) www.electropedia.org. We appreciate the initiative of the International Electrotechnical Commission (IEC) to make this vocabulary freely available. References to the vocabulary are given in the text as the IEV ref. number (e.g. IEV 192-01-24 for the term reliability).

Last, but not least, we are grateful to the editorial and production staff at John Wiley & Sons for their careful, effective, and professional work. In particular, we would like to thank our main contacts in the final stages of preparing the book, Sarah Keegan, Kathleen Santoloci, and Viniprammia Premkumar.

Trondheim, 2020 *Marvin Rausand and Anne Barros*

References

Høyland, A. and Rausand, M. (1994). *System Reliability Theory: Models and Statistical Methods*. Hoboken, NJ: Wiley.

Rausand, M. and Høyland, A. (2004). *System Reliability Theory: Models, Statistical Methods, and Applications*, 2e. Hoboken, NJ: Wiley.

About the Companion Website

System Reliability Theory: Models, Statistical Methods, and Applications is accompanied by a companion website:

www.wiley.com/go/SystemReliabilityTheory3e

The book companion site is split into two sub-sites hosted by Wiley:

1. An *open site* that is accessible to all users of the book.
2. An *instructor site* for instructors/lecturers (i.e. not accessible for students and general readers of the book).

The two sites contain a number of PDF files. These files have version numbers and will be updated when required.

In addition to these two sites hosted by Wiley, we will maintain a GitHub site for the book.

Open Site

The open site contains:

1. A *supplement* to the book with comments to chapters, suggestions for further reading, and general information about the subject area.
2. Slides to the various chapters (made with LaTeX/Beamer).
3. Control questions to each chapter.
4. Errata (list of misprints and minor errors – a more frequently updated errata list may be found on the book's GitHub site).

Instructor Site

The instructor site contains:

1. Solutions to end of chapter problems.
2. Suggested lecturing plans (what to cover, which problems to use, etc.).
3. Additional problems with solutions.
4. FAQ list.

GitHub Site

The GitHub site is open to all users – and should have a clear link from the Wiley sites. The GitHub site will contain:

1. A brief description of the book.
2. Detailed R-scripts related to the book.
3. Detailed Python-scripts related to the book.
4. Errata list (see above under *Open site*).
5. FAQ related to the book – with our answers/comments.

The URL of the GitHub site is https://github.com/RausandBarros/ReliabilityBookScripts

Contact Person

The contact person for the book companion site and the GitHub site is
Anne Barros (anne.barros@centralesupelec.fr)

1

Introduction

1.1 What is Reliability?

Nowadays, nearly all of us depend on a wide range of technical products and services in our everyday life. We expect our electrical appliances, cars, computers, mobile phones, and so on, to function when we need them, and to be reliable for a rather long time. We expect services, such as electricity, computer networks, and transport, to be supplied without disruptions or delays. When a product, machinery, or service fails, the consequences may sometimes be catastrophic. More often, product flaws and service outages lead to customer dissatisfaction and expenses for the supplier through warranty costs and product recalls. For many suppliers, reliability has become *a matter of survival*.

There is no generally accepted definition of the *reliability* of a technical product. The definition and interpretation of the term vary from industry to industry and from user to user. For the purpose of this book, we choose a rather wide definition of the reliability of a technical item.

Definition 1.1 (Reliability)
The ability of an item to perform as required in a stated operating context and for a stated period of time. □

The term *item* is used to designate any technical system, subsystem, or component. The items studied in this book are built of hardware parts, and to an increasing degree, of software. When relevant, the user interface is part of the item, but operators and other humans are not part of the items studied here.

The reliability concept is illustrated in Figure 1.1. The *required performance* is determined by laws and regulations, standards, customer requirements and expectations, and supplier requirements, and is usually stated in a *specification document*, where delimitations of the operating context are stated. As long as

System Reliability Theory: Models, Statistical Methods, and Applications, Third Edition.
Marvin Rausand, Anne Barros, and Arnljot Høyland.

Companion Website: www.wiley.com/go/SystemReliabilityTheory3e

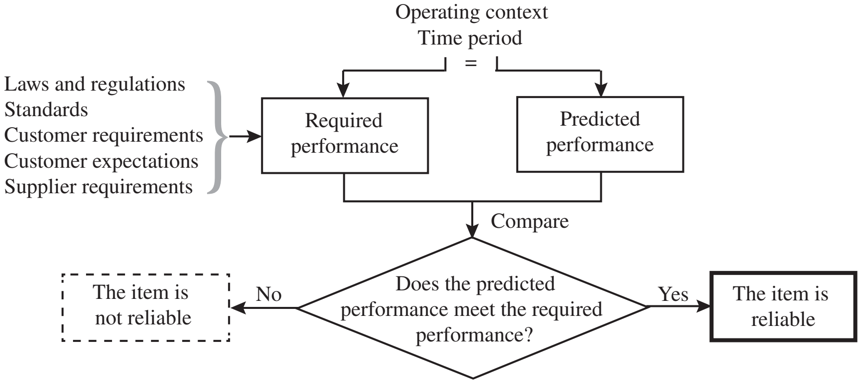


Figure 1.1 The reliability concept.

the predicted performance at least fulfills the required performance, the item is reliable – when it is used in the same operating context and for the period of time stated in the required performance.

By *operating context*, we mean the environmental conditions the item is used in, the usage patterns, and the loads it is subjected to, and how the item is serviced and maintained.

Definition 1.1 is not new and is not created by us. Several authors and organizations have used this, or a very similar definition of reliability, at least since the 1980s. A more thorough discussion of reliability and related concepts is given in Section 1.3.

1.1.1 Service Reliability

A *service* is provided by a person, an organization, or a technical item to a person or a technical item. The entity providing the service is called a *service provider*, and the entity receiving the service is called a *customer*. Services can be provided on a (i) continuous basis (e.g. electric power, computer networks), (ii) according to a timetable (e.g. bus, rail, and air transport), or (iii) on demand (e.g. payment by debit cards).

Many services are provided by a single service provider to a high number of customers. A customer considers the service to be reliable when she receives the service (e.g. electric power) with sufficient quality without outages. We define service reliability as follows:

Definition 1.2 (Service reliability)
The ability of the service to meet its supply function with the required quality under stated conditions for a specified period of time. □

Several quantitative service reliability metrics have been defined, but they vary between the different types of services.

1.1.2 Past and Future Reliability

In our daily language, the term "reliability" is used to describe both past and future behavior. We may, for example, say that (i) "my previous car was very reliable" and (ii) "I believe that my new car will be very reliable." These two statements are quite different. The first statement is based on experience with the car over a certain period, whereas the second statement is a *prediction* of what will happen in the future. We distinguish them by using two different terms.

Reliability (single word) is always used to describe the *future* performance of an item. Because we cannot predict the future with certainty, we need to use probabilistic statements when assessing the reliability.

Achieved reliability is used to describe the item's *past* performance, which is assumed to be known to the analyst. No probabilistic statements are therefore involved. The achieved reliability is also called *observed reliability*.

The focus of this book is on reliability and the future performance. The achieved reliability is most relevant in Chapter 14, where analysis of observed failure data is discussed.

1.2 The Importance of Reliability

Several producers of technical items have struggled and even collapsed because of item flaws and failures. To build a reputation for reliability is a long-term project, but it may take a short time to lose this reputation. The main drivers for high reliability are listed in Figure 1.2. Over the years, the reliability has improved for

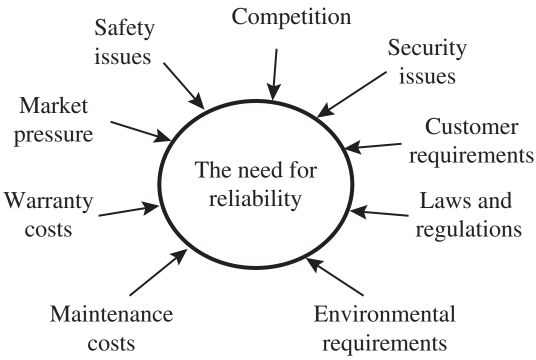


Figure 1.2 Main drivers for high reliability.

almost all types of items, but at the same time, customers expect a higher and higher reliability of the new items they buy. Current customers further expect that possible failures in the warranty period are rectified without any cost to the customer. To be attractive in the market, the suppliers have to offer a longer and longer warranty period.

If items have flaws that affect safety, safety regulations may require all the flawed items to be recalled for repair or modification. Such recalls are rather frequent in the car industry, but are also common in many other industries. In addition to excessive warranty costs and item recalls, flawed items lead to dissatisfied and nonreturning customers.

1.2.1 Related Applications

Reliability considerations and reliability studies are important inputs to a number of related applications. Several of these applications have adopted the basic terminology from reliability. Among the relevant applications are:

Risk analysis. The main steps of a *quantitative risk analysis* (QRA) are: (i) identification and description of potential *initiating events* that may lead to unwanted consequences, (ii) identification of the main causes of each initiating event and quantification of the frequency of the initiating events, and (iii) identification of the potential consequences of the initiating events and quantification of the probabilities of each consequence. The three steps are shown in the *bow-tie model* in Figure 1.3, where the main methods are indicated. The methods that are covered in this book are marked with an $*$.

Maintenance planning. Maintenance and reliability are closely interlinked. High-quality maintenance improves the operational reliability and high reliability gives few failures and low maintenance cost. The close link is also visible in the popular approach *reliability-centered maintenance* (RCM), which is discussed in Chapter 9.

Quality. Quality management is increasingly focused, stimulated by the ISO 9000 series of standards. The concepts of quality and reliability are closely connected. Reliability may in some respects be considered to be a quality characteristic.

Life cycle costing. The life cycle cost (LCC) may be split into three types: (i) capital expenditure (CAPEX), (ii) operational expenditure (OPEX), and (iii) risk expenditure (RISKEX). The main links to reliability are with types (ii) and (iii). The OPEX is influenced by how regular the function/service is and the cost of maintenance. The RISKEX covers the cost related to accidents, system failures, and insurance. LCC is also called *total ownership cost*.

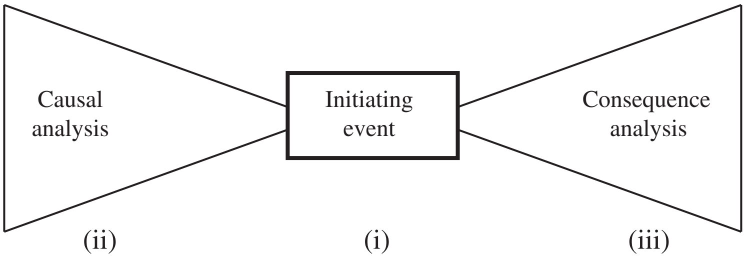
Production assurance. Failures in a production system lead to downtime and reduced production. To assure a regular production, the production system must have a high reliability. Production assurance is treated in the international standard ISO 20815 and discussed in Chapter 6.

Warranty planning. A warranty is a formal commitment to deliver reliable items. If failures and malfunctions are detected during a specified *warranty period*, the supplier has to repair and/or compensate the failure. Unreliable items may incur a high cost for the supplier.

Systems engineering. Reliability is one of the most important quality attributes of many technical systems. Reliability assurance is therefore an important topic during the systems engineering process. This is especially the case within the nuclear power, the aviation, the aerospace, the car, and the process industries.

Environmental protection. Reliability studies are used to improve the design and operational availability of many types of environmental protection systems. Many industries have realized that a main part of the pollution from their plants is caused by production irregularities and that consequently the reliability of the plant is an important factor in order to reduce pollution. Environmental risk analyses are carried out according to the procedure shown in Figure 1.3.

Technology qualification. Many customers require the producer of technical items to verify that the item satisfies the agreed requirements. The verification is carried out by following a *technology qualification program* (TQP) based on



Methods		
– Fault tree analysis* – Reliability block diagrams* – Bayesian networks* – FMECA*	– Checklists – Preliminary hazard analysis – FMECA* – HAZOP	– Event tree analysis* – Consequence models – Reliability assessment* – Evacuation models – Simulation
– Reliability data sources*	– Event data sources	

Figure 1.3 Main steps of risk analysis, with main methods. The methods covered in this book are marked with $*$.

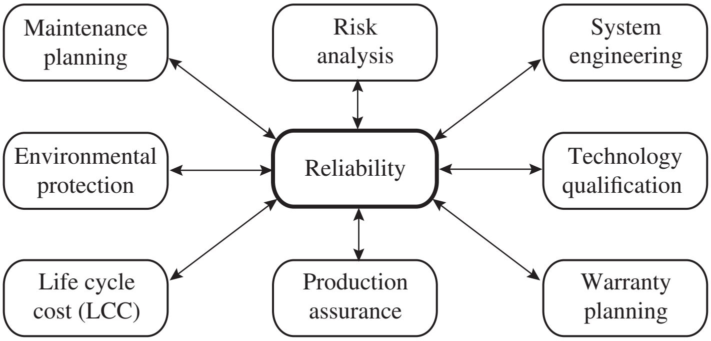


Figure 1.4 Reliability as basis of other applications.

analysis and testing. This is especially the case within the aerospace, defense, and petroleum industries (e.g. see DNV-RP-A203 2011).

Applications related to reliability are illustrated in Figure 1.4.

1.3 Basic Reliability Concepts

The main concept of this book is *reliability* as defined in Definition 1.1. The aim of this section is to discuss and clarify this definition and to define related terms, such as maintainability and maintenance, availability, quality, and dependability.

It is important that all main words are defined in an unambiguous way. We fully agree with Kaplan (1990) who states: “When the words are used sloppily, concepts become fuzzy, thinking is muddled, communication is ambiguous, and decisions and actions are suboptimal.”

1.3.1 Reliability

Definition 1.1 says that reliability expresses “the ability of an item to perform as required in a stated operating context and for a stated period of time.” We start by clarifying the main words in this definition.

(1) Reliability is defined by using the word *ability*, which is not directly measurable. A quantitative evaluation of the item’s ability to perform must therefore be based on one or more metrics, called *reliability metrics*. Several probabilistic reliability metrics are defined and discussed in Section 1.4.

(2) Some authors use the word *capability* instead of *ability* in the definition of reliability and claim that the term “capability” is more embracing, covering both ability and capacity. Most dictionaries list ability and capability as synonyms. We prefer the word “ability” because this is the word most commonly used.

(3) The statement *perform as required* means that the item must be able to perform one or more specified functions according to the performance criteria for these function(s). Functions and performance criteria are discussed in Section 2.5.

(4) Many items can perform a high number of functions. To assess the reliability (e.g. of a car), we must specify the required function(s) that are considered.

(5) To be reliable, the item must do more than meet an initial factory performance or quality specification – it must operate satisfactorily for a specified period of time in the actual operating context.

(6) The stated period of time may be a delimited time period, such as a mission time, the time of ownership, and several more.

(7) The time may be measured by many different time concepts, such as calendar time, time in operation, number of work cycles, and so on. For vehicles, the time is often measured as the number of kilometers driven. For items that are not operated continuously in the same mode, a more complicated time concept may be needed.

Inherent and Actual Reliability

It may be useful to qualify the reliability of an item by adding a word, such as inherent or actual. The inherent reliability is defined as follows:

Definition 1.3 (Inherent reliability)
The reliability of the item as designed and manufactured, which excludes effects of operation, environment, and support conditions other than those assumed and stated in the item requirements and specification. □

The inherent reliability is therefore the reliability of a brand new item that will be used and maintained exactly according to the conditions described in the item specification document or implicitly assumed. The inherent reliability is sometimes called *built reliability* or *built-in reliability* of the item.

The design and development team always attempts to adapt the item to the actual operating context, but it is difficult, if not impossible, to account for all the aspects in practical use. The actual reliability may consequently be different from the inherent reliability that was determined before the item was put into use. The actual reliability of an item is defined as follows:

Definition 1.4 (Actual reliability)
The reliability of the item in an actual operating context. □

The actual reliability is sometimes called *operational reliability* or *functional reliability*.

Software Reliability

Software reliability is different from hardware reliability. Hardware items generally deteriorate due to wear or other mechanisms and failures occur as a random process. Software, on the other hand, does not deteriorate and faults or *bugs* remain dormant and undetected until the software is modified or a specific condition or trigger activates the bug – leading to item failure. Software bugs are manifestations of mistakes done in specification, design, and/or implementation. Reliability analysis of a software program is done by checking the code syntax according to specific rules and by testing (debugging) the software for a variety of input data. This process is not discussed further in this book. Interested readers may consult ISO 25010.

1.3.2 Maintainability and Maintenance

Many items have to be maintained to perform as required. Two different concepts are important, maintainability, and maintenance. *Maintainability* is a design feature of the item and indicates how easy it is to get access to the parts that are to be maintained and how fast a specific maintenance task can be done. *Maintenance* describes the actual work that is done to maintain an item. Maintainability is defined as follows:

Definition 1.5 (Maintainability)

The ability of an item, under stated conditions of use, to be retained in, or restored to, a state in which it can perform as required, when maintenance is performed under stated conditions and using prescribed procedures and resources. □

Maintainability is further discussed in Chapter 9. Maintenance is defined as follows:

Definition 1.6 (Maintenance)

The combination of all technical and management actions during the life cycle of an item intended to retain the item in, or restore it to, a state in which it can perform as required (IEV 192-06-01). □

Hardware maintenance is discussed in more detail in Chapters 9 and 12. Software maintenance is not treated in this book.

1.3.3 Availability

Availability measures the degree to which an item is able to operate at some future time t or during a future time interval (t_1, t_2), and is in this book regarded as a reliability metric. The availability of an item depends on the reliability, recoverability, and maintainability of the item, and also on the maintenance support performance. Recoverability is the item's ability to recover from a failure, without repair. Maintenance support is the resources that are available for maintenance, such as workshops, qualified personnel, and tools. Availability is discussed in Chapters 6, 11, and 13.

1.3.4 Quality

The term "quality" is closely related to reliability and is defined as follows:

Definition 1.7 (Quality)
The totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs. □

Quality is sometimes defined as *conformity to specifications* and a quality defect is referred to as a nonconformity. According to common usage, quality denotes the conformity of the item to its specification as manufactured, whereas reliability denotes its ability to continue to comply with its specification over its useful life. With this interpretation, reliability may be considered as an extension of quality into the time domain.

1.3.5 Dependability

Dependability is a more recent concept that embraces the concepts of reliability, maintainability, and availability, and in some cases also safety and security. Dependability has, especially, become known through the important series of standards IEC 60300 "Dependability management." The IEV defines dependability as follows:

Definition 1.8 (Dependability)
The ability (of an item) to perform as and when required (IEV 192-01-01). □

Another commonly used definition is "Trustworthiness of a system such that reliance can justifiably be placed on the service it delivers" (Laprie 1992).

Remark 1.1 (Translating the word "dependability")
Many languages, such as Norwegian and Chinese, do not have words that can distinguish reliability and dependability, and reliability and dependability are therefore translated to the same word. □

1.3.6 Safety and Security

General safety is outside the scope of this book, and we deal only with the safety aspects of a specified technical item and define safety as follows:

Definition 1.9 (Safety)
Freedom from unacceptable risk caused by the technical item. □

This definition is a rephrasing of definition IEV 351-57-05. The concept *safety* is mainly used related to random hazards, whereas the concept *security* is used related to deliberate hostile actions. We define security as:

Definition 1.10 (Security)
Dependability with respect to prevention of deliberate hostile actions. □

The deliberate hostile action can be a physical attack (e.g. arson, sabotage, and theft) or a cyberattack. The generic categories of attacks are called *threats* and the entity using a threat is called a *threat actor*, a *threat agent*, or an *adversary*. Arson is therefore a threat, and an arsonist is a threat actor. The threat actor may be a disgruntled employee, a single criminal, a competitor, a group, or even a country. When a threat actor attacks, he seeks to exploit some weaknesses of the item. Such a weakness is called a *vulnerability* of the item.

Remark 1.2 (Natural threats)
The word "threat" is also used for natural events, such as avalanche, earthquake, flooding, landslide, lightning, tsunami, and volcano eruption. We may, for example, say that earthquake is a threat to our item. Threat actors are not involved for this type of threats. □

1.3.7 RAM and RAMS

RAM, as an acronym for reliability, availability, and maintainability, is often used, for example, in the annual RAM Symposium.[1] RAM is sometimes extended to RAMS where S is added to denote safety and/or security. The RAMS acronym is, for example, used in the railway standard IEC 62278.

Remark 1.3 (Broad interpretation of reliability)
In this book, the term "reliability" is used quite broadly, rather similar to RAM as defined above. The same interpretation is used by Birolini (2014). □

[1] RAM Symposium: www.rams.org.

1.4 Reliability Metrics

Throughout this book, it is assumed that the time-to-failure and the repair time of an item are *random variables* with *probability distributions* that describe the future behavior of the item. The future behavior may be evaluated based on one or more *reliability metrics*. A reliability metric is a "quantity" that is derived from the reliability model and is, as such, not directly measurable. When performance data become available, we may estimate or predict quantitative values for each reliability metric.

A single reliability metric is not able to tell the whole truth. Sometimes, we need to use several reliability metrics to get a sufficiently clear picture of how reliable an item is.

1.4.1 Reliability Metrics for a Technical Item

Common reliability metrics for an item include

(1) The mean time-to-failure (MTTF)
(2) The number of failures per time unit (*failure frequency*)
(3) The probability that the item does not fail in a time interval $(0, t]$ (*survivor probability*)
(4) The probability that the item is able to function at time t (*availability at time t*)

These and several other reliability metrics are given a mathematical precise definition in Chapter 5, and are discussed and exemplified in all the subsequent chapters.

Example 1.1 (Average availability and downtime)
Consider the electricity supply, which is supposed to be available at any time. The achieved average availability A_{av} of the supply is quantified as

$$A_{\text{av}} = \frac{\text{Uptime}}{\text{Total time}} = 1 - \frac{\text{Downtime}}{\text{Total time}}$$

If we consider a period of one year, the *total time* is approximately 8760 hours. The *downtime* is the time, during the specified time period, the service is not available. The relationship between the average availability and the length of the downtime is illustrated in Table 1.1. □

Table 1.1 Availability and downtime.

90	36.5 d
99	3.65 d
99.9	8.76 h
99.99	52 min
99.999	5 min

1.4.2 Reliability Metrics for a Service

A wide range of service reliability metrics have been defined, but these vary significantly between the application areas. The most detailed metrics are available for electric power supply (e.g. see IEEE Std. 1366 2012).

Example 1.2 (Airline reliability and availability)
Airline passengers are mainly concerned about whether the journey will be safe and whether the aircraft will take off and land on the scheduled times. The second concern is, by airlines, expressed by the *dispatch reliability*, which is defined as the probability that a scheduled departure takes place within a specified time after the scheduled departure time. Many airlines use a 15-minutes margin between actual and scheduled departure time for a flight to be considered as having departed on time. The achieved dispatch reliability indicator for a (past) period is reported as the percentage of all departures that departed on time.

$$\text{Dispatch reliability} = \frac{\text{No. of departures on time}}{\text{No. of departures} + \text{cancelations}}$$

For technical items, the airlines are mainly using the reliability metrics listed in Section 1.4.1 □

1.5 Approaches to Reliability Analysis

Three main branches of reliability can be distinguished:

- Hardware reliability
- Software reliability
- Human reliability

The present book is concerned with hardware items (existing or in design) that may or may not have embedded software. Within hardware reliability, two different approaches may be used: the *physical approach* and/or the *systems approach*.

1.5.1 The Physical Approach to Reliability

In the physical approach, the strength of a technical item is modeled as a random variable S. The item is exposed to a load L that is also modeled as a random variable. The distributions of the strength and the load at a specific time t are shown in Figure 1.5. A failure will occur as soon as the load is higher than the strength. The survival probability R of the item is defined as the probability that the strength is greater than the load,

$$R = \Pr(S > L)$$

where $\Pr(A)$ is the probability of event A.

The load may vary with time and be modeled as a time-dependent variable $L(t)$. The item may deteriorate with time, due to failure mechanisms, such as, corrosion, erosion, and fatigue. The strength of the item will therefore also be a function of time, $S(t)$. A possible realization of $S(t)$ and $L(t)$ is shown in Figure 1.6. The time-to-failure T of the item is the (shortest) time until $S(t) < L(t)$,

$$T = \min\{t; S(t) < L(t)\}$$

and the survivor probability $R(t)$ of the item may be defined as

$$R(t) = \Pr(T > t)$$

The physical approach is mainly used for reliability analyses of structural elements, such as beams and bridges. The approach is therefore often called *structural reliability analysis* (Melchers 1999). A structural element, such as a leg on an offshore platform, may be exposed to loads from waves, current, and wind. The loads may come from different directions, and the load must therefore be modeled as a vector $\boldsymbol{L}(t)$. In the same way, the strength will also depend on the direction and has to be modeled as a vector $\boldsymbol{S}(t)$. The models and the analysis therefore become complicated. The physical approach is not pursued further in this book.

1.5.2 Systems Approach to Reliability

By the systems approach, all our information about the operational loads and the strength of an item is incorporated in its probability distribution function

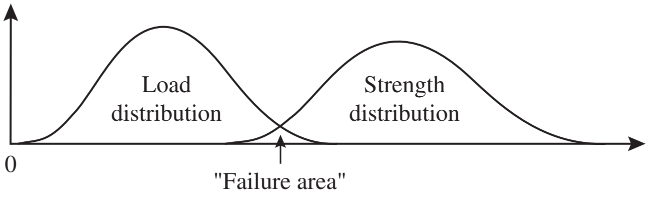


Figure 1.5 Load and the strength distributions at a specified time t.

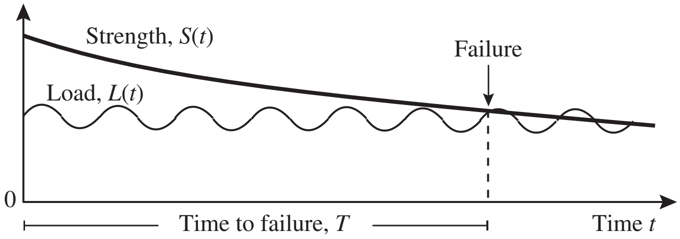


Figure 1.6 Possible realization of the load and the strength of an item.

$F(t)$ of the *time-to-failure* T. No explicit modeling of the loads and the strength is carried out. Reliability metrics, such as the *survivor probability* and the *mean time-to-failure* are deduced directly from the probability distribution function $F(t)$. Various approaches can be used to model the reliability of systems of several components and to include maintenance and replacement of components. When several components are combined into a system, the analysis is called a *system reliability analysis.*

Quantitative results are based on information about the reliability of the components. Such information comes from statistical data on past experience with the same or similar components, laboratory testing, or from expert judgments. This approach has similarities to actuarial assessments, and the systems approach to reliability is, therefore, sometimes referred to as an *actuarial approach*. This book is concerned with the systems approach to reliability.

System Models

In reliability studies of technical systems, we always have to work with models of the systems. These models may be graphical (networks of different types) or mathematical. A mathematical model is necessary in order to be able to bring in data and use mathematical and statistical methods to estimate reliability parameters. For such models, two conflicting interests always apply:

(1) The model should be sufficiently simple to be handled by available mathematical and statistical methods.
(2) The model should be sufficiently "realistic" such that the deducted results are of practical relevance.

We should always bear in mind that we are working with an idealized, simplified model of the system. Furthermore, the results we derive are, strictly speaking, valid only for the model, and are accordingly only "correct" to the extent that the model is realistic.

The modeling situation is illustrated in Figure 1.7. Before we start developing a model, we should clearly understand what type of decision the results from

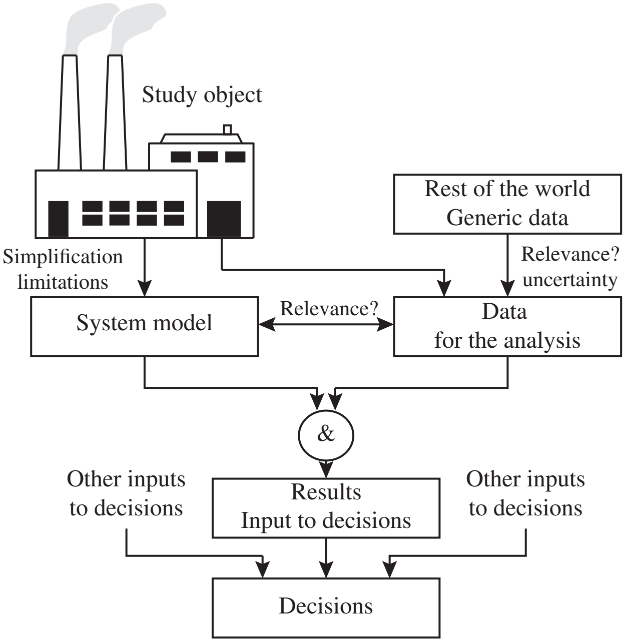


Figure 1.7 The system reliability analysis process.

our analysis should provide input to, and also the required format of the input to the decision. To estimate the system reliability from a model, we need input data. The data will usually come from generic data sources, as discussed in Chapter 16. The generic data may not be fully relevant for our system and may have to be adjusted by expert judgment. This is especially the case when we are introducing new technology. Some data may also come from the specific system. When establishing the system model, we have to consider the type, amount, and quality of the available input data. It has limited value to establish a very detailed model of the system if we cannot find the required input data.

1.6 Reliability Engineering

Engineering deals with the design, building, and use of technical items. *Reliability engineering* is an engineering discipline that provides support to the engineering process. To be successful, reliability engineering must be integrated in the engineering process and the reliability engineer(s) must take full part in the engineering team.

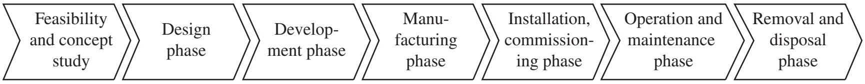

Feasibility and concept study
Design phase
Develop-ment phase
Manu-facturing phase
Installation, commission-ing phase
Operation and maintenance phase
Removal and disposal phase

Figure 1.8 The phases of a system development project (example).

An item development project is split into a number of phases. The number and the title of these phases vary from industry to industry and also between companies in the same industry. A typical set of phases is shown in Figure 1.8.

The phases in Figure 1.8 are arranged as a time axis, but iterations are usually required, for example, to make a redesign after a defect has been revealed in a later phase. Each phase is usually divided into stages, and many manufacturers have procedures describing in detail which reliability analyses to carry out in each stage together with procedures for the data flow.

Reliability engineering has its most important role in the three first phases in Figure 1.8, but should be integrated in all phases.

1.6.1 Roles of the Reliability Engineer

The objective of reliability engineering is to identify, analyze, and mitigate failures and operational problems during all phases of an item's life cycle. The reliability engineer has an important role in all these phases. Below, the roles of the reliability engineer are listed briefly in the design and development phases and in the operational phase.

Roles in Design and Development

A reliability engineer has her most important role in the specification, design, and development phases of a new item. In these phases, the reliability engineer helps the development team to

(1) Identify potential failures of suggested component and module concepts such that failures may be designed out.
(2) Quantify the reliability of suggested system concepts.
(3) Provide input to decisions about modularization, stacking, and system layout.
(4) Make tradeoffs between factors such as cost, functions, performance, reliability, time to market, safety, and security.
(5) Identify weaknesses of the system design such that they can be corrected before the system goes to manufacturing or to the customers.
(6) Clarify benefits and drawbacks related to redundancy of components and modules.
(7) Identify causes and effects of possible failure modes.

(8) Compare the LCC of design alternatives.
(9) Evaluate the cost of suggested warranty policies.
(10) Calculate the reliability of system options as input to choice between these.
(11) Plan and perform reliability acceptance or qualification testing (e.g. in a TQP framework).

Roles in Normal Operation

The main role of the reliability engineer in normal operation is to track items causing abnormally high maintenance cost and production losses or service outages, then find ways to reduce these losses or high costs. The role of a reliability engineer may vary from company to company, but the overall goal is always the same: reduce maintenance costs as much as possible without interrupting system operation.

Another main role of the reliability engineer in this phase is to collect, analyze, and present reliability data. This topic is treated in detail in Chapter 14.

Reliability has to be designed and manufactured into an item. It is too late and too costly to wait until the item is produced. Reliability considerations must be integrated into all steps of the development process. This book presents the main theory and many of the required methods and tools for reliability engineering, but reliability engineering also requires a number of methods that are outside the scope of this book. When to carry out an analysis, which data are available at this stage, and how to update and use the results are central questions in reliability engineering that are not covered in this book.

1.6.2 Timing of Reliability Studies

Reliability studies are carried out to provide input to *decision-making* related to an item. The objectives and the scope of the reliability study are dependent on the type of decision to be made. Before starting a reliability study, it is essential to have a clear understanding of the decision and the data needed as input to the decision-making. A reliability study to provide input to decisions on warranties may, for example, be quite different from a reliability study to provide input to decisions on safety barriers in a risk assessment.

It is very important that the reliability studies are planned and executed such that the required results are available before the decision-making takes place!

1.7 Objectives, Scope, and Delimitations of the Book

The overall objective of this book is to give a thorough introduction to component and system reliability analysis by the system reliability approach. More detailed objectives are

(1) To present and discuss the terminology and the main models used in system reliability studies.
(2) To present the main analytical methods used in reliability engineering and management.
(3) To present and discuss basic theory of maintenance and preventive maintenance modeling and illustrate how these can be applied.
(4) To present the main theory and a selection of methods for reliability data analysis, which is also called *survival analysis*.
(5) To give an introduction to Bayesian probability and Bayesian data analysis.

The book does not specifically deal with how to engineer and manage a reliable system. The main topics of the book are connected to how to define and quantify reliability metrics and to predict the reliability of a system. Our aim is that the book will be a valuable source as follows:

(a) A textbook for system reliability courses at university level.
(b) A handbook for reliability engineers in industry and consulting companies.
(c) A reference book for scientists and engineers in related disciplines.

The following delimitations apply:

- The study object is built of hardware parts based on mechanical, electrical, or electronic technology, and may or may not have embedded software and communication to/from the outside. In most cases, the study object has a human/operator interface. Operators and third-party personnel are outside the scope of the book. This means that human reliability, as such, is not covered. The prime focus of the book is on hardware items.
- The reliability of purely software items is outside the scope of this book.
- Structural reliability issues are not covered in this book.
- The focus of the book is on components and rather simple systems. The theory and methods presented may also be useful for analyzing complex systems, but we have to realize that they may not be sufficient.
- Failures caused by deliberate hostile actions is covered rather rudimentarily.
- In the main part of the book, we assume that each item can have only two states, functioning or failed. Multistate reliability is not covered properly.
- A general introduction to maintenance is not provided. The presentation is delimited to aspects of maintenance that are directly relevant for system reliability.
- The book provides a thorough introduction to system reliability analysis, but does not cover reliability engineering and reliability management in a sufficient way.

1.8 Trends and Challenges

System reliability has been around since the 1940s. The relevance of reliability has increased steadily and we clearly see trends and challenges that will increase the relevance in the years to come. In this section, we briefly mention some of these trends and challenges. An overall trend is that customers expect new items to be BETTER, FASTER, and CHEAPER than the items they replace. More specific challenges include

(1) Items get more and more complicated with a lot of embedded software. Hardware functions are replaced with software-based functions. Because the software-based functions are relatively cheap, many items are loaded with "nice-to-have" function that may also fail.
(2) Most producers meet fierce international competition. To survive, this requires reduced development costs, shorter time to market, and less time spent on analyses and testing. New items have to be sufficiently reliable in the first concept version.
(3) Customers require more and more of the items they purchase, related to functions, quality, and reliability. The requirements are often changing rapidly. Factors influencing item requirements are shown in Figure 1.9.
(4) There is an increasing focus on safety and environmental friendliness and an increasing risk of item call-back if the items should have safety-related defects.
(5) New items are increasingly made up of elements from a variety of subcontractors from many different countries, making it difficult for the main producer to verify the item reliability.

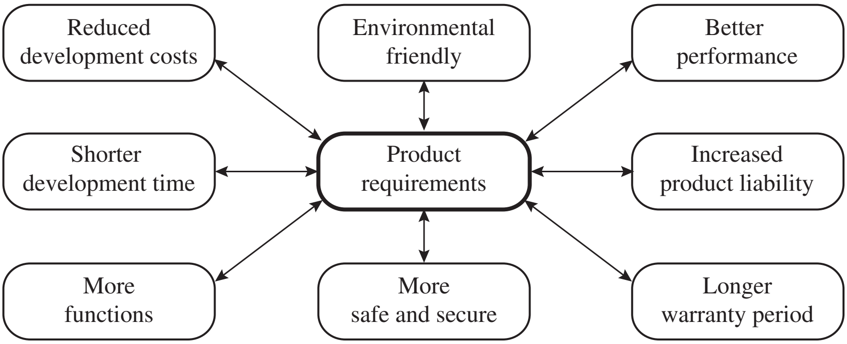


Figure 1.9 Factors that influence item requirements.

(6) For some items, high-speed operation reduces the tolerance of deviations and increases the consequences of failures, should they happen.

(7) There is an increasing focus on warranty. Companies have disappeared because of excessive warranty costs.

(8) An increasing number of items are now connected to a cybernetwork and are vulnerable to cyberattacks. Current challenges are related to the rapid developments of smart homes, smart cities, smart transport systems, the Internet of Things (IoT), cyber-physical systems, systems of systems, and Industry 4.0. Within few years, we expect to see many more new initiatives of similar nature. This will make reliability analyses even more challenging.

1.9 Standards and Guidelines

A range of standards and guidelines stating requirements to reliability and safety have been issued. Any reliability engineer needs to be familiar with the standards and guidelines that are applicable within her subject areas.

1.10 History of System Reliability

This section highlights some achievements in the history of system reliability starting from the 1930s. We realize that our presentation is biased because we put too much focus on activities in Europe and in the United States. In addition, we have included mainly events and books that have influenced our own learning and understanding of system reliability. The development of reliability theory has been strongly influenced by a series of accidents and catastrophic failures. Some of these are mentioned, but you may find that we have missed many important accidents.

Some of the achievements mentioned in this section may be difficult to comprehend fully at this stage, and it may therefore be wise to postpone the reading of this section until you have delved deeper into the subject.

1930s

At the beginning of the 1930s, *Walter Shewhart*, *Harold F. Dodge*, and *Harry G. Romig* laid down the theoretical basis for utilizing statistical methods in quality control of industrial products, but such methods were not used to any great extent until the beginning of World War II. Products that were composed of a large number of parts often failed, despite the fact that they were made of individual high-quality components.

An important achievement was made in the 1930s by the Swedish professor *Waloddi Weibull* (1887–1979) during his studies of the strength of materials. In Weibull (1939), he laid the basis for one of the most important probability distributions in reliability theory, the *Weibull distribution* (Weibull 1951).

1940s

It is often claimed that the first quantitative system reliability assessment can be attributed to *Robert Lusser* (1899–1969). He was a German engineer and aircraft designer who took part in several well-known Messerschmitt and Heinkel designs during World War II. During the war, a group in Germany was working under Wernher von Braun developing the V-1 missile, but the 10 first V-1 missiles were all fiascos. In spite of attempts to provide high-quality parts and careful attention to details, all the first missiles either exploded on the launching pad or landed "too soon" (in the English Channel). Robert Lusser was called in as a consultant. His task was to analyze the missile system, and he quickly derived the *product probability law of series components* saying that the reliability of series system is equal to the product of the reliabilities of the individual components that make up the system. If the system comprises a large number of components, the system reliability may therefore be low, even though the individual components have high reliabilities. A young mathematician, *Erich Pieruschka*, assisted Wernher von Braun and may have been as important as Lusser in developing Lusser's law. Some authors prefer to refer to Pieruschka's law instead of Lusser's law.

An important contribution to the subsequent reliability theory was made by the Russian mathematician *Boris V. Gnedenko* (1912–1995) in his 1943 paper "On the limiting distribution of the maximum term in a random series."[2] In this paper, Gnedenko provided rigorous proofs and formulated three classes of limit distributions, one of which was the Weibull distribution. Gnedenko was not the first to define the three limit distribution classes, but the first to provide proofs. The classes had earlier been defined by Fisher and Tippett (1928). The extreme value theorem proved by Gnedenko is often referred to as the Fisher–Tippett–Gnedenko theorem.

In the United States, attempts were made to compensate a low-system reliability by improving the quality of the individual components. Better raw materials and better designs for the products were demanded. A higher system reliability was obtained, but extensive systematic analysis of the problem was probably not carried out at that time.

After World War II, the development continued throughout the world as increasingly more complicated products were produced, composed of an ever-

[2] For a discussion of Gnedenko's contribution, see Smith (1992).

increasing number of components (e.g. television sets and electronic computers). With automation, the need for complicated control and safety systems also became steadily more pressing.

Several attempts to test and quantify the reliability of electronic components began in the 1940s during World War II. The war activities clearly revealed that electron (vacuum) tubes were the most failure-prone components in electronic systems (Denson 1998). Several groups tried to identify ways to improve the reliability of electronic systems, and it was suggested that the reliability of the components needed to be verified by testing before full-scale production.

In 1945, *Milton A. Miner* formulated the important Miner's rule for fatigue failures (Miner 1945). A similar rule was suggested by the Swedish engineer *Nils Arvid Palmgren* (1890–1971) already in 1924 while studying the life length of roller bearings. The rule is therefore also called the Palmgren–Miner's rule and the Miner–Palmgren's rule.

In 1949, the Institute of Electrical and Electronic Engineers (IEEE) formed a professional group on quality control as part of its Institute of Radio Engineers. The group got more and more focused on reliability issues and changed name several times. In 1979, the group got its current name, *IEEE Reliability Society*.

The first guideline on failure modes and effects analysis (FMEA) was issued in 1949 (MIL-P-1629 1949). This guideline was later developed into the military standard MIL-STD-1629A.

1950s

The Advisory Group on Reliability of Electronic Equipment (AGREE) was established in 1950 to survey the field and identify and promote actions that could provide more reliable electronic equipment. A big step forward was made by the report AGREE (1957).

The 1950s saw much pioneering work in the reliability discipline. The Weibull distribution was properly defined (Weibull 1951) and soon became popular and several US military handbooks were issued. The statistical branch of reliability theory was strongly enhanced by the paper "Life testing" (Epstein and Sobel 1953) and some years later by the Kaplan–Meier estimate (Kaplan and Meier 1958).

The UK Atomic Energy Authority (UKAEA) was formed in 1954. It soon got involved in performing safety and reliability assessments for outside bodies, due to its competence in such work in the nuclear field.

In the middle of the 1950s, Bell Telephone Laboratories started to develop the *fault tree* approach describing the possible causes of an undesired event, using Boolean algebra.

1960s

Reliability theory was significantly enhanced during the 1960s and several important books were published, among which are Bazovsky (1961), Lloyd and Lipow (1962), Barlow and Proschan (1965), and Shooman (1968).

In 1960, the first edition of the US military handbook MIL-HDBK-217F was released, outlining an approach for reliability prediction of electronic equipment.

In 1962, the Bell Telephone Laboratories published a report on the safety of the launch control system for the Minuteman intercontinental ballistic missile using *fault tree analysis*. This report is considered to be the birth of fault tree analysis. The same year, *David R. Cox* published his seminal book on renewal theory (Cox 1962).

In 1964, the "Reliability Engineering" handbook was published by Aeronautical Radio, Incorporated (ARINC). This book (ARINC 1964) was one of the first books describing engineering aspects of reliability theory. Another book on reliability engineering was Ireson (1966).

In 1968, the Air Transport Association (ATA) issued a document titled "Maintenance Evaluation and Program Development." This document gave rise to the approach "maintenance steering group" (MSG). The first version, called MSG-1, was used to ensure the safety of the new Boeing 747-100 aircraft. The MSG-1 process used failure modes, effects, and criticality analysis (FMECA) and a decision logic to develop scheduled maintenance. MSG-1 was later developed into MSG-2 and MSG-3, which is the current version.

The Reliability Analysis Center (RAC) was established in 1968 as a technical information center for the US Department of Defense, and soon played a very important role in the development of reliability theory and practice. The RAC journal was widely distributed, presenting updated information about new developments.

The military standard "Reliability program for systems and equipment" was published in 1969 (MIL-STD-785A 1969).

One of the most influential researchers on reliability theory in the 1960s was *Zygmunt Wilhelm Birnbaum* (1903–2000). He introduced a new importance metric of component reliability (Birnbaum 1969), made a probabilistic version of Miner's rule for fatigue life (Birnbaum and Saunders 1968), and made many other significant contributions.

1970s

A most important event for reliability in the 1970s was the release of the report from the *Reactor Safety Study* in 1975 (NUREG-75/014). The study was made by a group of experts lead by professor *Norman Rasmussen* of MIT. A high number of important methods were developed as part of – or inspired by – the Reactor Safety Study.

The US Nuclear Regulatory Commission (NRC) was established the same year (in 1975) and soon started to issue NRC Regulations, called NUREG.

The nuclear accident at Three Mile Island (TMI) near Harrisburg, PA occurred in 1979. In light of the recent Reactor Safety Study, it had a great impact of the development of system reliability theory.

In the early 1970s, several important results on network reliability were developed in Russia (e.g. see Lomonosov and Polesskii 1971). Many new books on system reliability were published. Among these are Green and Bourne (1972), Barlow and Proschan (1975), and Kapur and Lamberson (1977).

Analysis of reliability and lifetime data grew more important and the new book Mann et al. (1974) provided help on theory and methods. An even more important publication in this area was *David R. Cox*'s paper "Regression models and life tables (with discussions)" (Cox 1972).

Based on the ideas of the MSG-approach (see 1960s), a new maintenance planning approach called "reliability-centered maintenance" (RCM) was introduced in 1978 (Nowlan and Heap 1978). The RCM approach was initially developed for the defense industry, but is today used in many other applications and a high number of standards and guidelines have been issued.

In Norway, the first major accident in the offshore oil and gas industry occurred in 1977, the Bravo blowout in the Ekofisk field in the North Sea. This was a shock for the Norwegian industry and the government. As a consequence of this accident, a large research program, called "Safety Offshore" was launched by the Norwegian Research Council. A high number of safety and reliability projects were sponsored by the oil and gas industry. The first author of this book started lecturing a course in system reliability at the Norwegian University of Science and Technology (NTNU) in 1978.

The UKAEA Safety and Reliability Directorate (SRD), established in 1977, became a very active unit with a strong influence on the development of reliability theory, especially in Europe.

1980s

The 1980s started with a new journal *Reliability Engineering*, which had a great influence on the further development of reliability theory. The first editor of the journal was *Frank R. Farmer* (1914–2001), who made significant contributions in both risk and reliability theory. The title of the journal was later changed to *Reliability Engineering and System Safety*.

The Offshore Reliability Data (OREDA) project was initiated in 1981 and the first OREDA handbook was published in 1984. The same year another important reliability data handbook, IEEE Std. 500 (1984) also entered the market.

Reliability data analysis became more important, and several books on this topic were published in the early 1980s, the most influential may be Kalbfleisch and Prentice (1980), Lawless (1982), Nelson (1982), and Cox and Oakes (1984).

Fault tree analysis got more standardized through the *Fault Tree Handbook* that was published by the US NRC in 1981 (NUREG-0492). Bayesian probability entered into the field of reliability promoted by the book Martz and Waller (1982).

To strengthen the US semiconductor industry, the organization SEMATECH was established in 1987. SEMATECH prepared and made available a range of high-quality reliability guidelines that were studied far beyond the semiconductor industry.

Several universities established education programs in safety and reliability during the 1980s. Most notable were perhaps the programs provided by the Center of Risk and Reliability at the University of Maryland and the NTNU.

Several catastrophic accidents occurred in the 1980s and clearly showed the importance of risk and reliability. Among these were the capsizing of the Alexander Kielland offshore platform in 1980, the gas disaster in Bhopal, India in 1984, the fire and chemical spill at the Sandoz warehouse in Basel, Switzerland in 1986, the Challenger space shuttle accident in 1986, and the explosion on the offshore platform Piper Alpha in 1988. Several of these accidents prompted changes in legislation, new requirements to risk and reliability analyses, and initiated a range of research projects.

After 1990

The developments mentioned above continued and were strengthened in the years after 1990. The topic of system reliability got more and more popular and a range of new journals, new books, new education programs, new computer programs, new organizations, and a variety of reliability conferences emerged. The first edition of the current book was published in 1994, based on experience from reliability courses at NTNU.

The industry started to integrate reliability in their system development processes, often as part of a systems engineering framework. The topics of reliability qualification and technology readiness became more and more important and requirements were integrated in contracts of specialized products.

The first edition of the important standard IEC 61508 “Functional safety of electrical/electronic/programmable electronic safety-related systems” came in 1997 and required producers and users of safety-instrumented systems (SIS) to perform detailed reliability assessments.

During this period, more and more software has been introduced in almost all types of systems. Software quality and reliability are now an important part of most system reliability assessments. More recently, security aspects have also entered the scene.

The current survey has highlighted some few fragments of the history of system reliability. A more thorough treatment of the history is given by Coppola (1984), Denson (1998), and Knight (1991) and National Research Council (2015, Annex D). A lot of valuable information may also be found by searching the Internet.

1.11 Problems

1.1 Discuss the main similarities and differences between the concepts of quality and reliability.

1.2 List some of the services you make use of in your daily life. Which factors do you consider relevant in order to describe the reliability of each of these services?

1.3 Section 1.2 lists several application areas that are related to, and use terminology from reliability theory. Can you suggest some more application areas?

1.4 Discuss the main differences between hardware reliability and software reliability. Do you consider the term “software quality” to be more or less relevant than “software reliability”?

1.5 A *stakeholder* may be defined as a “person or organization that can affect, be affected by, or perceive themselves to be affected by a decision or activity.” Choose a specific item/system (e.g. a dangerous installation) and list the main stakeholders of a system reliability analysis of this item/system.

1.6 Evaluate the maintainability of a modern mobile phone. Can you suggest any design changes of the phone that will improve its maintainability?

1.7 List some technical items for which you consider it beneficial to use the physical (i.e. load-strength) approach to reliability analysis.

References

AGREE (1957). Reliability of Military Electronic Equipment. *Tech. Rep*. Washington, DC: Advisory Group on Reliability of Electronic Equipment, U.S. Department of Defense.

ARINC (1964). *Reliability Engineering*. Englewood Cliffs, NJ: Prentice-Hall.

Barlow, R.E. and Proschan, F. (1965). *Mathematical Theory of Reliability*. New York: Wiley.

Barlow, R.E. and Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing, Probability Models*. New York: Holt, Rinehart, and Winston.

Bazovsky, I. (1961). *Reliability Theory and Practice*. Englewood Cliffs, NJ: Prentice-Hall.

Birnbaum, Z.W. (1969). On the importance of different components in a multicomponent system. In: *Multivariate Analysis II* (ed. P.R. Krishnaiah), 581–592. New York: Academic Press.

Birnbaum, Z.W. and Saunders, S.C. (1968). A probabilistic interpretation of Miner's rule. *SIAM Journal of Applied Mathematics* 16: 637–652.

Birolini, A. (2014). *Reliability Engineering: Theory and Practice*, 7e. Heidelberg: Springer.

Coppola, A. (1984). Reliability engineering of electronic equipment: a historic perspective. *IEEE Transactions on Reliability* 33: 29–35.

Cox, D.R. (1962). *Renewal Theory*, In: *Methuen's Monographs on applied probability and statistics*. Methuen, London: Methuen & Co.

Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society* B 21: 411–421.

Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman and Hall.

Denson, W. (1998). The history of reliability prediction. *IEEE Transactions on Reliability* 47 (3): 321–328.

DNV-RP-A203 (2011). Qualification procedures for new technology, *Recommended practice*, DNV GL. Høvik, Norway.

Epstein, B. and Sobel, M. (1953). Life testing. *Journal of the American Statistical Association* 48 (263): 486–502.

Fisher, R.A. and Tippett, L.H.C. (1928). Limiting forms of the frequency distributions of the largest or smallest of a sample. *Proceedings of the Cambridge Philosophical Society* 24: 180–190.

Green, A.E. and Bourne, A.J. (1972). *Reliability Technology*. Chichester: Wiley.

IEC 62278 (2002). *Railway applications – specification and demonstration of reliability, availability, maintainability and safety (RAMS)*, *International standard*. Geneva: International Electrotechnical Commission.

IEEE Std. 500 (1984). *IEEE guide for the collection and presentation of electrical, electronic, sensing component, and mechanical equipment reliability data for nuclear power generating stations, Standard*. New York: Institute of Electrical and Electronics Engineers.

IEEE Std. 1366 (2012). *IEEE guide for electric power distribution reliability indices*, *Standard*. New York: Institute of Electrical and Electronics Engineers.

Ireson, W.G (ed.) (1966). *Reliability Handbook*. New York: McGraw-Hill.

ISO 20815 (2018). *Petroleum, petrochemical, and natural gas industries: production assurance and reliability management*, *International standard*. Geneva: International Organization for Standardization.

ISO 25010 (2011). *Systems and software engineering – systems and software quality requirements and evaluation (SQuaRE) – system and software quality models*, *International standard*. Geneva: International Organization for Standardization.

ISO 9000 (2015). *Quality management systems – fundamentals and vocabulary*, *Standard ISO9000*. Geneva: International Organization for Standardization.

Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. Hoboken, NJ: Wiley.

Kaplan, S. (1990). Bayes is for eagles. *IEEE Transactions on Reliability* 39: 130–131.

Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53 (282): 457–481.

Kapur, K.C. and Lamberson, L.R. (1977). *Reliability in Engineering Design*. Hoboken, NJ: Wiley.

Knight, C.R. (1991). Four decades of reliability progress. In: *Annual Reliability and Maintainability Symposium*. Orlando, FL: IEEE, 29-31 January 1991. DOI: 10.1109/ARMS.1991.154429, 156–160.

Laprie, J.C. (1992). *Dependability: Basic Concepts and Terminology*. Berlin: Springer.

Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*. Hoboken, NJ: Wiley.

Lloyd, D.K. and Lipow, M. (1962). *Reliability: Management, Methods, and Mathematics*. Englewood Cliffs, NJ: Prentice-Hall.

Lomonosov, M.V. and Polesskii, V.P. (1971). A lower bound for network reliability. *Problems of Information Transmission* 7 (4): 118–123.

Mann, N.R., Schafer, R.E., and Singpurwalla, N.D. (1974). *Methods for Statistical Analysis of Reliability and Lifetime Data*. Hoboken, NJ: Wiley.

Martz, H.F. and Waller, R.A. (1982). *Bayesian Reliability Analysis*. New York: Wiley.

Melchers, R.E. (1999). *Structural Reliability Analysis and Prediction*, 2e. Hoboken, NJ: Wiley.

MIL-HDBK-217F (1995). Reliability prediction of electronic equipment, *Military handbook*. Washington, DC: U.S. Department of Defense.

MIL-P-1629 (1949). Procedures for performing a failure modes, effects, and criticality analysis, *Military procedure*. Washington, DC: U.S. Department of Defense.

MIL-STD-785A (1969). Reliability program for systems and equipment development and production, *Military standard*. Washington, DC: U.S. Department of Defense.

MIL-STD-1629A (1980). Procedures for performing a failure mode, effects, and criticality analysis, *Military standard*. Washington, DC: U.S. Department of Defense.

Miner, M.A. (1945). Cumulative damage in fatigue. *Journal of Applied Mechanics* 12: A159–A164.

National Research Council (2015). *Reliability Growth: Enhancing Defense System Reliability*. Washington, DC: The National Academies Press.

Nelson, W. (1982). *Applied Life Data Analysis*. New York: Wiley.

Nowlan, F.S. and Heap, H.F. (1978). Reliability-Centered Maintenance. *Tech. Rep. A066-579*. San Francisco, CA: United Airlines.

NUREG-0492 (1981). Fault tree handbook, *Handbook NUREG-0492*. Washington, DC: U.S. Nuclear Regulatory Commission.

NUREG-75/014 (1975). Reactor Safety: An Assessment of Accident Risk in U.S. Commercial Nuclear Power Plants. *Report NUREG-75/014*. Washington, DC: U.S. Nuclear Regulatory Commission.

Shooman, M.L. (1968). *Probabilistic Reliability: An Engineering Approach*. New York: McGraw-Hill.

Smith, R.L. (1992). Introduction to Gnedenko (1943) on the limiting distribution of the maximum term in a random series. In: *Breakthroughs in Statistics* (ed. S. Kotz and N.L. Johnson). New York: Springer, 185–194.

Weibull, W. (1939). A Statistical Theory of the Strength of Materials. *Report 151*. Stockholm, Sweden: Royal Swedish Institute for Engineering Research.

Weibull, W. (1951). A statistical distribution function of wide applicability. *Journal of Applied Mechanics* 18: 293–297.

2

The Study Object and its Functions

2.1 Introduction

Our *study object* is usually a technical system, but can also be a single technical component. A component is an item that is not broken down into its constituent parts in a reliability analysis. This is contrary to a technical system, which is always broken down into its constituent parts, be it subsystems, modules, or components.

This chapter defines, delimits, and classifies the study object. The system boundary and its operating context are defined. The concepts of system functions and their performance criteria are defined and discussed, and some simple approaches to functional modeling and analysis are presented. This is followed by a brief introduction to the Newtonian–Cartesian paradigm and its implications for system analysis. Systems are classified as simple, complicated, or complex, and it is argued why complex systems are outside the scope of this book. The chapters end with an introduction to system structure modeling by *reliability block diagrams*.

2.2 System and System Elements

A (technical) system may be defined as follows:

Definition 2.1 (System)
A set of interrelated elements that are organized to achieve one or more stated purposes. □

The term "system" is derived from the Greek word *systema*, which means an organized relationship among functioning items. Aslaksen (2013) considers a system as the combination of three related sets: (i) a set of elements $\mathcal{E}$, (ii) a set $\mathcal{R}_I$ of

System Reliability Theory: Models, Statistical Methods, and Applications, Third Edition.
Marvin Rausand, Anne Barros, and Arnljot Høyland.

Companion Website: www.wiley.com/go/SystemReliabilityTheory3e

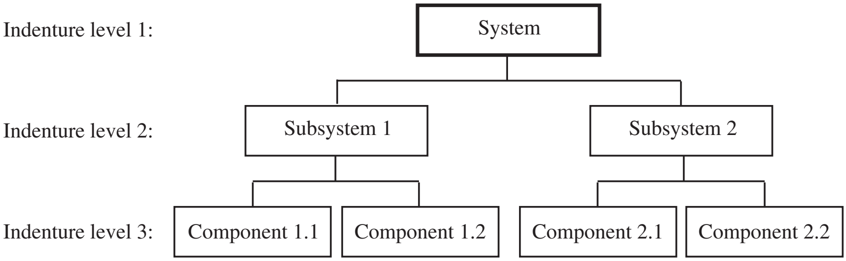


Figure 2.1 System breakdown structure (simplified).

internal interactions between elements, and (iii) a set $\mathcal{R}_E$ of external interactions between one or more elements and the external world (i.e. interactions that can be observed from outside the system).

For the purpose of a reliability study, the system elements are usually classified as subsystems, subsubsystems, and so on, down to the component level. The system elements may be organized by a *system breakdown structure* as shown (simplified) in Figure 2.1. The levels of the hierarchy are called *indenture levels*, where the first level is called indenture level 1, the next indenture level 2, and so on.[1] The number of levels required depends on the size of the system and the objectives of the reliability study. The various subsystems may have different numbers of levels.

The lowest level in the system breakdown structure – and in the reliability study – is called *component*. A component may itself be a system with many parts, but is considered a *black box* in the study. A black box is an element that is viewed in terms of its inputs and outputs, without concern about its internal structure and functions. When investigating the causes of a component failure, we sometimes need to study the states and conditions of the various *parts* of the component.

Subsystems are also referred to as *modules*. In system maintenance, terms such as *maintanable item* and *least replaceable unit* (LRU) are often used. A maintainable item is the lowest level in the system hierarchy that is specified for maintenance. A plethora of notions is used in the literature. Among these are the following: apparatus, component, element, equipment, instrument, item, module, part, product, system, and subsystem.

2.2.1 Item

To simplify the notation, the element we are currently studying is referred to as the *item*, whether it is a system, a subsystem, or a component. An *item* is defined as

[1] IEV defines indenture level as the "level of subdivision within a system hierarchy" (IEV 192-01-05).

Definition 2.2 (Item)
An entity that is able to perform at least one function of its own, under specified operational and environmental conditions and when the required energy and controls are available. □

We use the term *item*, unless when it is important to stress that we study a system consisting of subsystems, sub-subsystems, and so on.

2.2.2 Embedded Item

Embedded software is computer software that is written to control the technical item. An *embedded* item is a combination of hardware and software that together form a part of a larger item. An example of an embedded item is a microprocessor that controls a car engine. An embedded item is designed to run on its own without human intervention, and may be required to respond to events in real time. Today, we find embedded items in almost all our electric household units, such as refrigerators, washing machines, and ovens.

2.3 Boundary Conditions

A reliability study is always based on a range of assumptions and boundary conditions. The most notable is the *system boundary* that specifies which items are included in the study object and which are not. All systems are used in some sort of *environment* that may influence and be influenced by the system. To delimit the study object, a system boundary is drawn between the study object and its environment. The inputs to and outputs from the study object are drawn up, as shown in Figure 2.2. A slightly more detailed definition of the term system boundary is

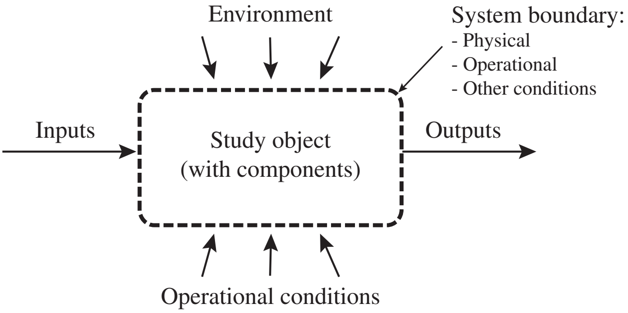


Figure 2.2 A study object (system) and its boundary.

Definition 2.3 (System boundary)
A system boundary is a boundary that separates the internal components and processes of a system from external entities. Internal to its boundary, the system has some degree of integrity, meaning the parts are working together, and this integrity gives the system a degree of autonomy. □

All assumptions and boundary conditions should be clearly stated in the documentation of the reliability study. Examples include answers to questions, such as

- What are the objectives of the study?
- What level of detail is required?
- What are the environmental conditions for the system?
- How is the system operated?
- Which operational phases are to be included in the study (e.g. start-up, steady state, maintenance, and disposal)?
- Which external stresses should be considered (e.g. earthquakes, lightning strikes, and sabotage)?

2.3.1 Closed and Open Systems

The study object may be a closed or an open system. A closed system may be defined as follows:

Definition 2.4 (Closed system)
A system where the interface to the environment is static and always according to the assumptions specified. □

In a closed system, the required inputs are always available, and random disturbances in the environment that may influence the study object are nonexisting. Most of the study objects considered in this book are closed systems. An open system is defined as follows:

Definition 2.5 (Open system)
A system where disturbances in the environment may influence the study object and where required system inputs and outputs may fluctuate or even be blocked. □

Open system are generally much more difficult to analyze than closed systems. Some open systems allow users to manipulate the system structure.

2.4 Operating Context

Items are generally designed and built for an *intended operating context* that should be clearly stated in the item specification and in the user documentation. The operating context specifies how the item is to be operated and maintained, limits to inputs, usage, and loads, and also which environmental conditions the item is supposed to work in and to tolerate. The user manual of a washing machine may, for example, specify intervals for the voltage and frequency of the power supply, the pressure and temperature of the water supply, the type and weight of laundry (e.g. clothes, carpets) put into the machine, the temperature in the room where the machine is located, and the surface on which the machine is placed. The operating context of the item is defined as follows:

Definition 2.6 (Operating context)
The environmental and operating conditions under which the item is (or is expected to be) operating. □

In some applications, the *concept of operations* (CONOPS) document describes the operating context of the item.

2.5 Functions and Performance Requirements

To be able to identify all potential item failures, the reliability engineer needs to have a thorough understanding of the various functions of the item and the performance criteria related to each function.

2.5.1 Functions

A *function* is a duty or an action the item has been designed to perform. A function requires one or more inputs to provide an output. The function is performed by technical and other resources and will usually also require some control (e.g. start signals). A function and its inputs and outputs are shown in Figure 2.3. The function and its elements are illustrated in Example 2.1.

Example 2.1 (Flashlight)
Consider a simple flashlight. The main function of the flashlight is to produce light. The required input is electric power coming from a battery. The resource is the flashlight with battery. The function is controlled by switching on/off the flashlight. □

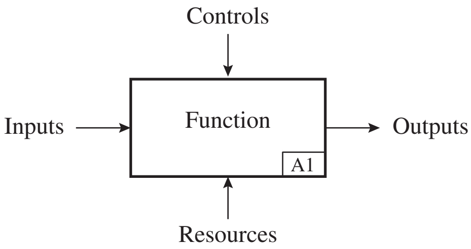


Figure 2.3 A function illustrated as a functional block.

A function may be defined as follows:

Definition 2.7 (Function)
An activity, process, or transformation stated by a verb and a noun that describes what must be accomplished. □

A function is an intended *effect* of an item and should be described such that each function has a single definite purpose. It is recommended to give the functions names that have a declarative structure, and say "what" is to be done rather than "how." The functions should preferably be expressed as a statement comprising a verb plus a noun; for example, provide light, close flow, contain fluid, pump fluid, and transmit signal. In practice, it is often difficult to specify a function with only two words, and additional words may need to be added.

2.5.2 Performance Requirements

New products and systems are developed to fulfill a set of *requirements*. These requirements are usually written into a *requirement document*. The requirements may be based on (i) identified customer needs, (ii) manufacturer's ideas to make the product more competitive, and (iii) requirements in standards, laws, and regulations. The IEV defines the term "requirement" as follows:

Definition 2.8 (Requirement)
Need or expectation that is stated, generally implied or obligatory (IEV 192-01-13). □

A *performance requirement* is a specification of the performance criteria related to a function. If, for example, the function is "pump water," a performance requirement may be that the output of water must be between 100 and 110 l/min. Some functions may have several performance requirements. Performance requirements are also referred to as *functional requirements* or *performance standards*.

2.5.3 Classification of Functions

A complicated item may have a high number of required functions. All functions are not equally important, and a classification may therefore be an aid for identification and analysis purposes. One way of classifying functions is as follows:

Essential functions. These are the functions required to fulfill the intended purpose of the item. The essential functions are simply the reasons for installing or using the item. The essential function is sometimes reflected in the name of the item. An essential function of a pump is, for example, to "pump fluid."

Auxiliary functions. These are the functions that are required to support the essential functions. The auxiliary functions are usually less obvious than the essential functions, but may in many cases be as important as the essential functions. Failure of an auxiliary function may in many cases be more safety-critical than a failure of an essential function. An auxiliary function of a pump is, for example, to "contain fluid."

Protective functions. These functions are intended to protect people, equipment, and the environment from damage and injury. The protective functions may be classified as follows:

(a) Safety functions (i.e. to prevent hazardous events and/or to reduce consequences to people, material assets, and the environment)
(b) Security functions (i.e. to prevent vulnerabilities, physical attacks, and cyberattacks)
(c) Environment functions (e.g. anti-pollution functions)
(d) Hygiene functions (e.g. for items used in food production or in hospitals).

Information functions. These functions cover condition monitoring, various gauges and alarms, communication monitoring, and so forth.

Interface functions. These functions apply to the interfaces between the item in question and other items. The interfaces may be active or passive. A passive interface is, for example, present when the item is a support or a base for another item.

Superfluous functions. These functions are never used and are often found in electronic equipment that have a wide range of "nice to have" functions that are not really necessary. Superfluous functions may further be found in systems that have been modified several times. Superfluous functions may also be present when the item has been designed for an operating context that is different from the actual operating context. In some cases, failure of a superfluous function may cause failure of other functions.

Some functions may belong to more than one class. For some applications, it may further be relevant to classify functions as follows:

(1) *Online functions*. These functions are operated either continuously or so often that the user has current knowledge about their status. The termination of an online function is called an *evident* or *detected* failure.
(2) *Off-line functions*. These functions are used intermittently or so infrequently that their availability is not known by the user without some special check or test. Some offline functions are not possible to test without damaging the item. An example of an offline function is the essential function of the airbag system of a car. Many protective functions are offline functions. The termination of the ability to perform an offline function is called a *hidden* or *undetected* failure.

2.5.4 Functional Modeling and Analysis

The objectives of a functional analysis are to

(1) Identify all the functions of the item.
(2) Identify the functions required in the various operating modes of the item.
(3) Provide a hierarchical decomposition of the item functions (see Section 2.5.5).
(4) Describe how each function is realized and provide the associated performance requirements.
(5) Identify the interrelationships between the functions.
(6) Identify interfaces with other systems and with the environment.

Functional analysis is an important step in *systems engineering* (Blanchard and Fabrycky 2011), and several analytical techniques have been developed. We briefly mention two of these techniques: Function trees and SADT / IDEF 0.

2.5.5 Function Trees

For complicated systems, it is sometimes beneficial to illustrate the various functions as a tree structure called a *function tree*. A function tree is a hierarchical functional breakdown structure starting with a system function or a system mission and illustrating the corresponding necessary functions on lower levels of indenture. The function tree is created by asking *how* an already established function is accomplished. This is repeated until functions on the lowest level are reached. The diagram may also be developed in the opposite direction by asking *why* a function is necessary. This is repeated until functions on the system level are reached. Function trees may be represented in many different ways. An example is shown in Figure 2.4.

A lower level function may be required by a number of main functions and may therefore appear several places in the function tree.

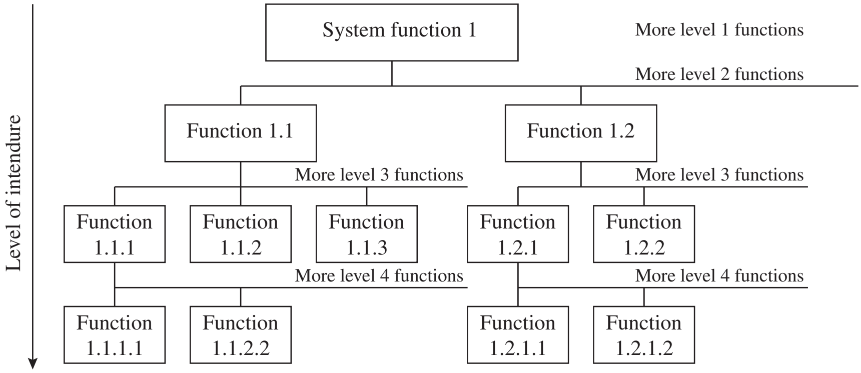


Figure 2.4 Function tree (generic).

2.5.6 SADT and IDEF 0

A widely used approach to functional modeling was introduced by Douglas T. Ross of Sof Tech Inc. in 1973, called the *structured analysis and design technique* (SADT). The SADT approach is described, for example, in Lambert et al. (1999) and Marca and McGowan (2006). In the SADT diagram each *functional block* is modeled according to a structure of five main elements, as shown in Figure 2.3

Function. Definition of the function to be performed.
Inputs. The energy, materials, and information necessary to perform the function.
Controls. The controls and other elements that constrain or govern how the function is carried out.
Resources. The people, systems, facilities, or equipment necessary to carry out the function.
Outputs. The result of the function. The outputs are sometimes split in two parts; the wanted outputs from the function, and unwanted outputs.

The output of a functional block may be the input to another functional block, or may act as a control of another functional block. This way the functional blocks can be linked to become a functional block diagram. An illustration of an SADT diagram for subsea oil and gas stimulation is shown in Figure 2.5. The diagram was developed as part of a student project at NTNU (Ødegaard 2002).

When constructing an SADT model, we use a top-down approach as shown in Figure 2.6. The top level represents a required system function. The functions necessary to fulfill the system function are established as an SADT diagram at the next level. Each function on this level is then broken down to lower level functions, and so on, until the desired level of decomposition has been reached. The hierarchy is maintained via a numbering system that organizes parent and child diagrams.

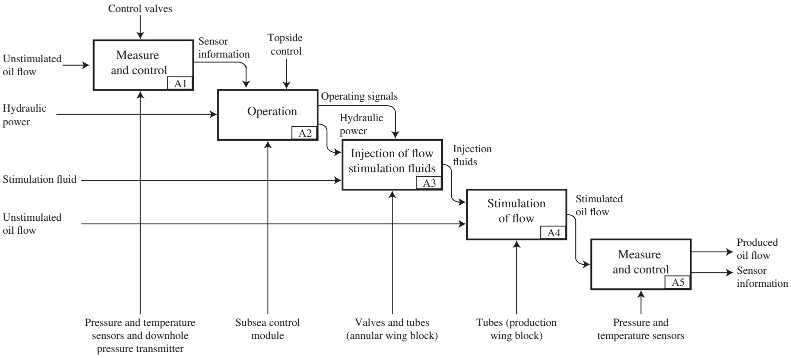
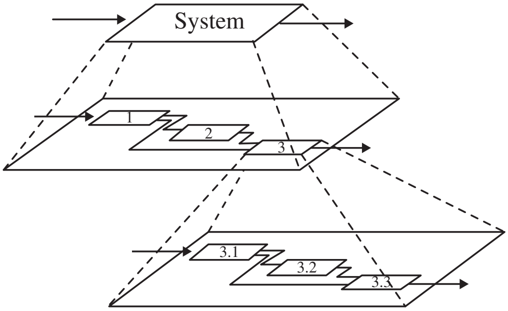


Figure 2.5 SADT diagram for subsea oil and gas stimulation.

Figure 2.6 Top-down approach to establish an SADT model.



The functional block in Figure 2.3 is also used in the *Integrated definition* language (IDEF), which is based on SADT and developed for the US Air Force. IDEF is divided into several modules. The module for modeling of system functions is called IDEF 0 (e.g. see U.S. Air Force 1981; U.S. DoD 2001; Marca and McGowan 2006).

For new systems, SADT and IDEF 0 may be used to define the requirements and specify the functions and as a basis for suggesting a solution that meets the requirements and performs the functions. For existing systems, SADT and IDEF 0 can be used to analyze the functions the system performs and to record the mechanisms (means) by which these functions are accomplished.

2.6 System Analysis

The term *analysis* means to break down – or decompose – a system or problem into its constituent components in order to get a better understanding of it. In a system analysis, all the constituent components are studied individually. The word "analysis" comes from an ancient Greek word that means "breaking up." To be able to analyze a system, the system must comply with the *Newtonian–Cartesian paradigm* (see box).

2.6.1 Synthesis

A *synthesis* is an opposite process of an analysis and is concerned with the combination of components and their properties to form a connected whole (i.e. a system).

In a system reliability study, we usually need to apply both analysis and synthesis to obtain a sufficient understanding of the system and its reliability.

The processes of system analysis and synthesis are illustrated in Figure 2.7.

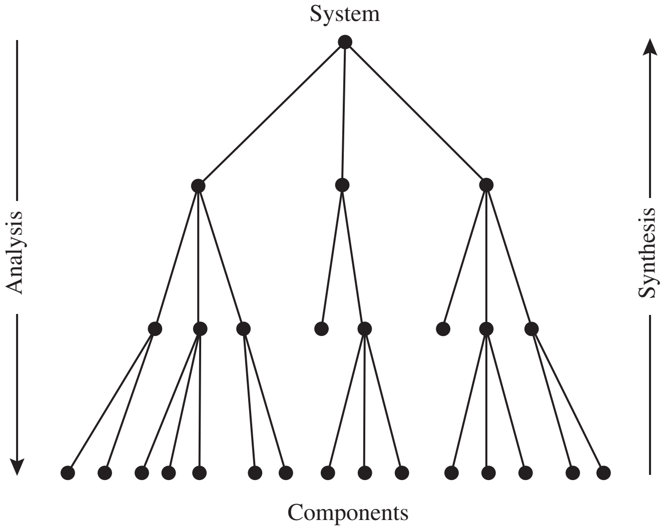


Figure 2.7 System analysis and synthesis.

2.7 Simple, Complicated, and Complex Systems

Most modern books on reliability theory and analysis seem to be concerned with "complex systems," but (almost) none of them define what they mean by the term *complex*. In our understanding, we may classify a system into one out of three categories:

Simple systems. A simple system is easy to understand and can be analyzed by following a defined procedure or algorithm. Most simple systems have a rather small number of components. Simple systems can generally be modeled by a series–parallel RBD (see Section 2.8).

The Newtonian–Cartesian Paradigm

A paradigm is a worldview underlying the theories and methodologies of a scientific subject. For system reliability, the Newtonian–Cartesian paradigm has been, and still is, the most essential. The basis for this paradigm was made by the French philosopher and scientist Réne Descartes (1596–1650) and the English mathematician and physicist Sir Isaac Newton (1642–1726).

The paradigm is based on Newton's three laws of forces and motion, his theories on universal gravitation, and the unifying theory that is called Newtonian mechanics. Another important basis for the paradigm is Descartes' theory

(Continued)

(Continued)

of reductionism and his division between mind and matter, between mental and physical processes. Reductionism implies that any system (or problem) can be adequately understood by reducing it, or decomposing it, to a set of its constituent components and by carefully and individually studying each component. When all the components on the lowest level have been carefully studied, a synthesis process can be started. By combining the knowledge about the components that feed into a module on the upper, next level, the paradigm implies that all important properties of this module can be deduced from the properties of its constituent components. This is then continued until the system level is reached (see Figure 2.7).

The Newtonian–Cartesian paradigm sees the world as a number of discrete, unchanging objects in an empty space. These objects interact in a linear, cause and effect manner. The time is linear and universal and not affected by speed or gravitation. The system behavior is deterministic, such that a particular cause leads to a unique effect. The paradigm supports the analysis of systems with a finite number of (mainly) independent parts that interact in a well-defined manner with relatively few interconnections.

The Newtonian–Cartesian paradigm is also called the *Newtonian paradigm* and the *mechanistic paradigm*.

The Newtonian–Cartesian paradigm has had an enormous success and most of our current knowledge about physical systems are based on this paradigm. Much more information about the Newtonian–Cartesian paradigm can be found by visiting a good library or searching the Internet.

Complicated systems. A complicated system has a high number of components with a fair degree of interrelationships and interdependencies between the components. By using current knowledge (e.g. by involving subject experts), we are able to understand the relevant system properties and to analyze it.

Complex systems. In a complex system, the behavior of at least some of the components or the interactions between them do not comply with the requirements of the Newtonian–Cartesian paradigm. A complex system cannot be adequately understood and analyzed by traditional approaches because the system is something more than a sum of its components.

An *emergent property* is a system property that cannot be deduced from the properties of the system components. In many cases, emergent properties lead to unexpected system behavior that may be dangerous. A system is usually not designed or built to be complex, but may develop into a complex system through changes, coupling, and emergence.

There is a considerable disagreement about how to delimit the concept of emergence. Some authors interpret emergence very widely and say that "properties" such as reliability, quality, and safety are emergent properties of a system.

Simple and complicated systems can be studied based on the Newtonian–Cartesian paradigm, whereas complex systems cannot be adequately studied within this paradigm. A new worldview called the *complexity paradigm* is therefore being developed.

All the examples in this book are related to simple systems, but the theory and methods presented may also be applied to complicated systems and many aspects of complex systems. Complex systems as such are not studied in this book.

Remark 2.1 (Classical methods ⇒ waste of time?)
Finally, you may wonder if the effort you make to learn the theory and methods described in this book is a waste of time when your study object is complex. According to Einstein and Infeld (1938), the development of new theory may be compared with climbing a mountain. When you have come to a certain height, you get a better overview, but you may realize that you need another strategy to reach the summit. To have reached the present height is an achievement that gives a good understanding of the further climbing efforts. □

2.8 System Structure Modeling

An early step of a system reliability study is to establish a model of the system structure. The model defines the system boundary and the elements of the system (i.e. inside the system boundary) and the interactions between these elements. We also make assumptions about how the system is operated and the environmental conditions and constraints that may affect the system elements and their behavior. A range of system modeling techniques are presented in later chapters. Here, we delimit the presentation to a rather simple approach – reliability block diagrams.

2.8.1 Reliability Block Diagram

This section describes how a system function (SF) can be modeled by a reliability block diagram (RBD). An RBD is a success-oriented graph with a single source (a) and a single terminal (b). The nodes of the RBD are called *blocks* or *functional blocks*. Each block represents a component function (or a combination of two or more functions). We assume that the blocks are numbered $1, 2, \ldots, n$, where n is known. This numbering is for convenience. In practical applications, a combination of letters and digits are often used to identify component functions. An RBD with n blocks is called an RBD of *order n*.

Each block is either *functioning* or *failed*, but the terms *up* and *down* are also used. Intermediate states are not allowed. To block i (for $i = 1, 2, \dots, n$) is connected a binary *state variable* x_i, defined as follows:

$$x_i = \begin{cases} 1 & \text{if block } i \text{ is functioning (up)} \\ 0 & \text{if block } i \text{ is failed (down)} \end{cases}. \tag{2.1}$$

Observe that $x_i = 1$ means that the specified function of block i is up. It does not mean that all the functions of the component associated with block i are up.

Blocks are drawn as squares or rectangles, as shown is Figure 2.8 for component function i. Connection between the end points (a) and (b) in Figure 2.8 means that block i is functioning (i.e. $x_i = 1$). It is possible to enter more information into the block and include a brief description of the required component function. An example is shown in Figure 2.9, where the component is a safety shutdown valve that is installed in a pipeline. A label is used to identify the block.

An RBD with three blocks representing a system function, SF, is shown in Figure 2.10. The system function, SF, is up if block 1 is functioning and either block 2, block 3, or both are functioning.

The blocks in Figure 2.10 are connected by *arcs*. Arcs are also called *edges*. The arcs are not directed, but directed arcs may sometimes be used to clarify the logic of the diagram. The system function, SF, is up if there exists a *path* from (a) to (b) through functioning blocks, otherwise, it is down. The RBD in Figure 2.10 is seen to have two paths $\{1, 2\}$ and $\{1, 3\}$.

System Structure

The RBD is *not* a physical layout diagram of the system, but a logic diagram that shows how and when the system function, SF, is up. The sequence of failures is

Figure 2.8 Component function *i* shown as a block.

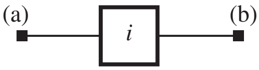


Figure 2.9 Alternative representation of the block in Figure 2.8

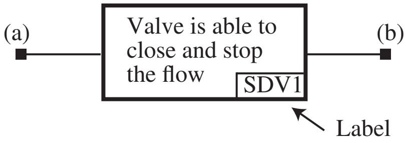
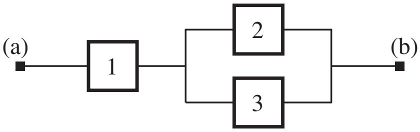


Figure 2.10 A simple reliability block diagram with three blocks.



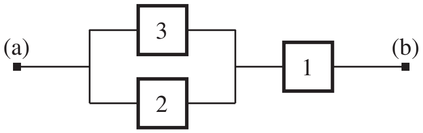


Figure 2.11 An alternative, and identical, version of the RBD in Figure 2.10.

not important, and the RBD in Figure 2.10 is therefore equivalent to the RBD in Figure 2.11. The RBD shows the *structure* of the system with respect to a specified system function, SF. When discussing RBD, we talk about the structure instead of the system. Separate RBDs have to be established for each system function.

Boolean Representation

Arranging several components along a path means connecting them by an AND- operation and arranging several components in parallel paths represents and OR-operation. In essence, a RBD is a graphical representation of a *Boolean expression*. Boolean expressions are discussed further in Section 4.6. The system function SF in Figure 2.10 is seen to be up if block 1 is up AND block 2 OR block 3 is up.

2.8.2 Series Structure

A series structure is functioning if and only if all the n blocks are functioning. This means that the structure fails as soon as one block fails. The RBD of a series structure of n blocks is shown in Figure 2.12. A path is seen to be available between the end points (a) and (b) – and the system is functioning – if and only if all the n blocks are functioning. The system function can be represented by the Boolean expression: The series structure is functioning if block 1 AND block 2 AND $\cdots$ AND block n are all functioning. As mentioned above, the sequence of the blocks in Figure 2.12 is not important, and we might have drawn the RBD with the n blocks in any sequence.

2.8.3 Parallel Structure

A parallel structure is functioning as long as at least one of its n blocks is able to function. The RBD of a parallel structure is shown in Figure 2.13. For this structure, there are n different paths between the end points (a) and (b). The structure is functioning if any one of these n paths is functioning. This means that the structure

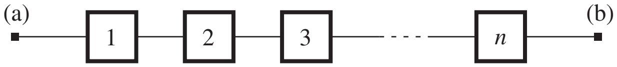


Figure 2.12 RBD for a series structure.

is functioning if at least one of the n blocks is functioning. The parallel structure can be represented by the Boolean expression: The parallel structure is functioning if block 1 OR block 2 $\cdots$ OR block n is functioning.

2.8.4 Redundancy

Redundancy is a means to improve the reliability of a structure. Redundancy may be defined as follows:

Definition 2.9 (Redundancy)
The provision of more than one means or parallel paths in a structure for performing a given function such that all means must fail before causing system failure□

The parallel structure in Figure 2.13 has redundancy because all the n blocks have to fail to cause the specified system failure, SF. Because n blocks have to fail, the system is said to have redundancy of *order n*.

Parallel or redundant paths can be installed for a single block, for a selection of blocks, or for the entire system function, SF.

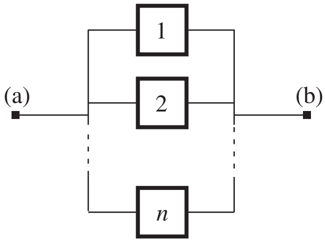
For hardware, redundancy may be achieved by installing one or more extra hardware items in parallel with the initial item. The redundant items may be identical or diverse. Adding redundancy increases the cost and makes the system more complicated, but if the cost of failure is high, redundancy is often an attractive option.

2.8.5 Voted Structure

A k-out-of-n ($k$oo$n$) voted structure is functioning as long as at least k of its n blocks are functioning ($k \leq n$). Observe that an $n$oo$n$ voted structure is a series structure and a 1oo$n$ structure is a parallel structure.

A 2oo3 voted structure is shown in Figure 2.14. Two different diagrams are shown. The diagram to the left is a physical diagram that shows the 2oo3 logic, whereas the RBD to the right is a series–parallel structure. In the RBD, we see

Figure 2.13 Parallel structure.



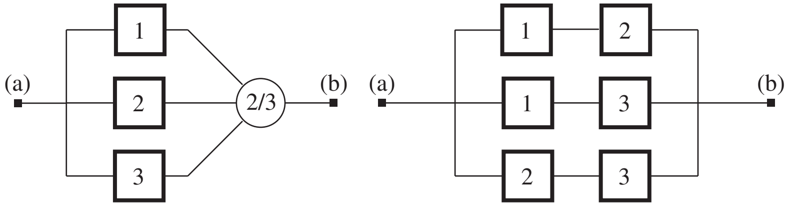


Figure 2.14 Voted structure 2oo3, (left) a physical diagram and (right) an RBD.

that the system is functioning when block 1 AND block 2 are functioning OR block 1 AND block 3 are functioning OR block 2 AND block 3 are functioning. Observe that each block appears in two different places in this RBD. This shows that an RBD is not a physical layout diagram, but a logical graph illustrating the specified function of the system.

2.8.6 Standby Structure

Redundancy may either be *active*, in which case the redundant items operate simultaneously in performing the same function (as for the parallel structure), or *standby*, such that the redundant items are only activated when the primary item fails. With standby redundancy, the standby items may be in cold standby or in partly loaded standby. With cold standby, the redundant item is considered to be as-good-as-new when activated. With partly loaded standby, the item may be failed or worn when activated.

A simple standby structure with two blocks is shown in Figure 2.15. Initially, block 1 is functioning. When block 1 fails, a signal is sent to the switch S to activate block 2 and a repair action of block 1 may be started. The switch S may be automatic, or a manual action to connect and start block 2. Depending on the operating rules, block 1 may be activated again as soon as the repair action is completed, or block 2 may run until it fails.

2.8.7 More Complicated Structures

Many of the structures, we study in this book, can be represented by a series–parallel RBD. A simple example of such a structure is shown in Figure 2.16.

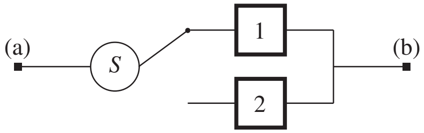


Figure 2.15 Standby structure.

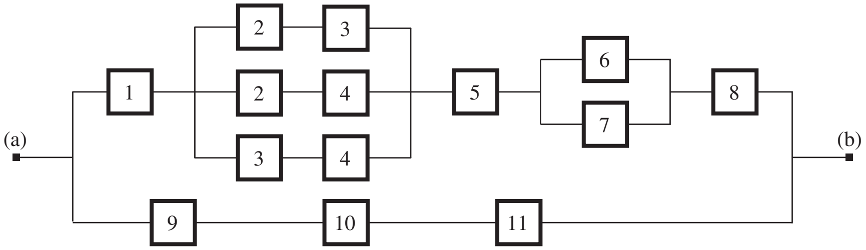


Figure 2.16 RBD for a series–parallel structure.

Remark 2.2 (Series–parallel structures)
The term series–parallel structure is not used in the same way by all authors. Some authors use the term to describe a series structure, where one or more of the blocks have added redundancy, that is, have parallel paths. The same authors use the term "parallel-series structure" to describe a parallel structure where two or more blocks appear in at least one of the parallel paths. In this book, we use the term series–parallel structure to describe a structure, where the blocks are arranged in any combination of series and parallel structures (as indicated in Figure 2.16). □

2.8.8 Two Different System Functions

The fact that different system functions give rise to different RBDs is illustrated in Example 2.2.

Example 2.2 (Pipeline with safety valves)
Consider a pipeline with two independent safety valves V_1 and V_2 that are physically installed in series, as shown in Figure 2.17a. The valves are supplied with a spring loaded fail-safe-close hydraulic actuator. The valves are opened and held open by hydraulic pressure and is closed automatically by spring force whenever the hydraulic pressure is removed or lost. In normal operation, both valves are held open. The essential function of the valve system is to act as a safety barrier, that is, to close and "stop flow" in the pipeline in case of an emergency.

The two blocks in Figure 2.17b represent the valve function "stop flow" for valve 1 and 2, respectively. This means that each valve is able to close and stop the flow in the pipeline. To achieve the system function "stop flow," it is sufficient that at least one of the individual valves can "stop flow." The associated RBD is therefore a parallel structure with respect to the system function "stop flow."

The valves may close spuriously, that is, without a control signal, and stop the flow in the pipeline. The two blocks in Figure 2.17c represent the valve function

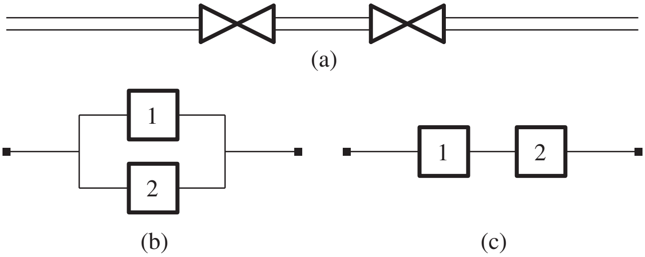


Figure 2.17 Two safety valves in a pipeline: (a) physical layout, (b) RBD for the safety barrier function, and (c) RBD for spurious closure.

"maintain flow" in the pipeline, for valves 1 and 2, respectively. Because the flow in the pipeline stops when one of the valves closes, the system function "maintain flow" is fulfilled only when both valves function with respect to the valve function "maintain flow". The associated RBD is therefore a series structure for the system function "maintain flow." □

Example 2.2 shows that two different functions of a single system give rise to two different RBD. Observe also that the blocks in the two RBDs represent different component functions in (b) and (c).

Remark 2.3 (Terminology problem)
Many authors use the term "component" instead of block. There is nothing wrong with this terminology–and we also use it later in this book–but we have to be very careful when, for example, saying that "component *i* is functioning." In cases, when it is not fully obvious, we should always add "with respect to the specified function." □

2.8.9 Practical Construction of RBDs

A specific system function, SF, usually requires a long range of subfunctions. For the essential function of a car, for example, we need the functions of the engine, the brakes, the steering, the ventilation, and many more. The RBD for the SF is then a long series structure of the required subsystem functions, as shown in Figure 2.18. Each of the required subfunctions may again need sub-subfunctions.

How many levels are required to depend on how complicated the system function is and the objectives of the analysis. RBDs are further discussed in Chapter 4. Chapter 6 deals with quantitative reliability analysis based on RBDs.

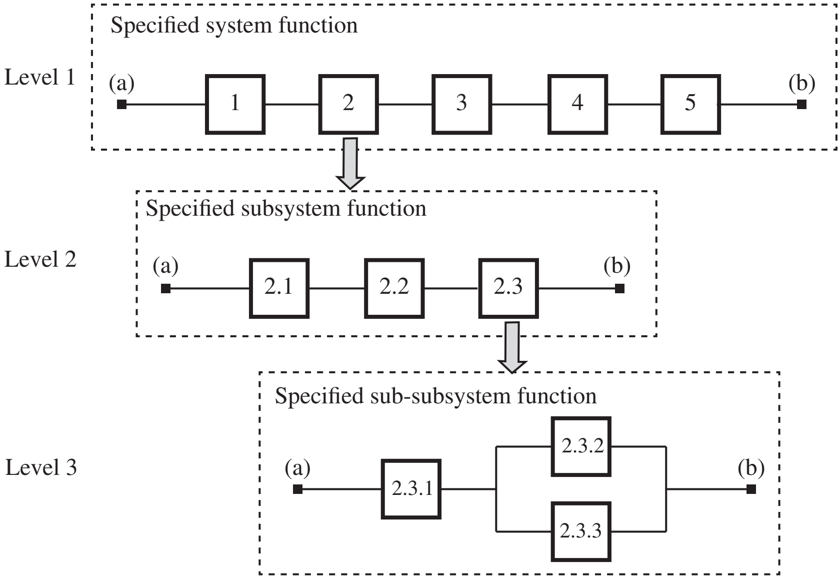

Specified system function
Level 1
(a)
1
2
3
4
5
(b)
Specified subsystem function
Level 2
(a)
2.1
2.2
2.3
(b)
Specified sub-subsystem function
Level 3
(a)
2.3.1
2.3.2
2.3.3
(b)

Figure 2.18 Construction of the RBD in levels.

2.9 Problems

2.1 Identify and describe briefly the main subsystems of a family car and establish a system breakdown structure for the car.

2.2 Establish a function tree for a (domestic) refrigerator.

2.3 List the environmental, operating, and maintenance factors that should be considered when defining the operating context of a family car.

2.4 List some information functions that are available in a modern car.

2.5 Identify the main functions of a family car and establish a function tree for the car.

2.6 List some safety functions of a modern car. Are the identified functions online or offline functions?

2.7 Identify and describe the functions of the front door of a house.

2.8 Describe the functions of a vacuum flask (thermos) and suggest relevant performance criteria.

2.9 Describe briefly a system you consider to be complex.

2.10 Refer to the SADT functional block (see Figure 2.3) and list all the inputs, controls, and resources you need to bake a pizza. The output from the function is the new-baked pizza. How would you set up the performance criteria for your pizza?

2.11 Based on an Internet search, explain what is meant by a CONOPS and list its main elements.

2.12 Based on an Internet search, list the main elements that are typically included in a system requirements document (or a system requirements specification).

2.13 Establish an RBD of the braking system of a family car.[2]

2.14 Consider a voted *k*oo*n* structure. The voting can be specified in two different ways:

- As the number *k* out of the *n* components that need to function for the system to function.
- As the number *k* of the *n* components that need to fail to cause system failure.

In the first case, we often write *k*oo*n*:G (for "good") and in the second case, we write *k*oo*n*:F (for failed).

(a) Determine the number *x* such that a 2oo4:G structure corresponds to a *x*oo4:F structure.

(b) Determine the number *x* such that a *k*oo*n*:G structure corresponds to a *x*oo*n*:F structure.

2.15 Are there any examples of standby redundancy in a family car? Justify your answer.

References

Aslaksen, E.W. (2013). *The System Concept and Its Application to Engineering*. Heidelberg: Springer-Verlag.

[2] You may need to search the Internet to find technical information on the braking system.

Blanchard, B.S. and Fabrycky, W.J. (2011). *Systems Engineering and Analysis*, 5e. Boston, MA: Pearson.

Einstein, A. and Infeld, L. (1938). *The Evolution of Physics*. Cambridge University Press.

Lambert, M., Riera, B., and Martel, G. (1999). Application of functional analysis techniques to supervisory systems. *Reliability Engineering & System Safety* 64 (2): 209–224.

Marca, D.A. and McGowan, C.L. (2006). *IDEF0 and SADT: A Modeler's Guide*. Auburndale, MA: OpenProcess.

Ødegaard, S. (2002). Reliability assessment of a subsea production tree. Project thesis. Trondheim, Norway: Norwegian University of Science and Technology.

U.S. Air Force (1981). Integrated Computer Aided Manufacturing (ICAM) Architecture. Part II. Volume IV, Functional Modeling Manual (IDEF 0). *Technical Report AFB AFWAL-TR-81-4023*. Wright Patterson Air Force Base, OH: Air Force Materials Laboratory.

U.S. DoD (2001). *Systems Engineering Fundamentals*. Fort Belvoir, VA: Defense Acquisition University Press.

3

Failures and Faults

3.1 Introduction

Failure is the most important concept in any reliability study, where typical questions addressed include:

- How long time will the item, on the average, be able to operate until the first *failure* occurs?
- What will the frequency of *failures* be? How many failures per year should we expect?
- What is the probability that the item will operate without *failure* during a specified time interval?
- If an item is demanded, what is the probability that it will *fail* to perform as required?

If we do not have a clear understanding of what a failure is, the reliability study may be of limited value. The term failure is used frequently in our daily language with many different interpretations and we also use a plethora of terms with similar meaning. Among these terms are blunder, breakdown, bug, collapse, defect, deficiency, error, fault, flaw, impairment, malfunction, mishap, mistake, and nonconformance.

How the term *failure* is interpreted varies between professional disciplines. Engineers working with quality, maintenance, warranty, safety, and reliability may have quite different opinions about whether or not a particular event constitutes a failure.

To perform a reliability study, it is important to understand thoroughly what is meant by the term failure in the context of reliability. Several definitions of failure have been proposed. IEV 192-03-01, for example, defines failure as "loss of the ability to perform as required."

System Reliability Theory: Models, Statistical Methods, and Applications, Third Edition.
Marvin Rausand, Anne Barros, and Arnljot Høyland.

Companion Website: www.wiley.com/go/SystemReliabilityTheory3e

This chapter is concerned with failures of single items only. Aspects related to interactions between several items in a *system* are treated in Chapter 4. Before continuing the discussion of failures, the concepts of states, transitions, and operational modes need to be introduced.

3.1.1 States and Transitions

At a given time, an item may be in one out of several *states*. The functions performed in one state may be different from the functions performed in other states. The item changes state by a *transition*. The transition may be automatic or manual and may occur at a random time or as a result of a command. Complicated items may have a high number of states and transitions.

Example 3.1 (Safety valve)
Consider a safety valve with a hydraulic fail-safe-close actuator. The valve is held open by hydraulic pressure during normal operation. When a specific critical situation occurs, a closing signal is sent to the safety valve and the valve closes by the force of the fail-safe actuator. The valve has two functioning states: open and closed. Transitions between these two states are facilitated by the actuator. The states and transitions are shown in Figure 3.1.

The essential function in state "open" is to provide a conduct for the medium/fluid through the valve, and the essential function in state "closed" is to stop the flow through the valve. An auxiliary function for both states is to contain the fluid and thereby to prevent leakage to the environment. □

Remark 3.1 (States and transition)
The difference between states and transitions is clear and intuitive for many items, but may be confusing for some items. The concepts of states and transition should therefore be used with care. □

3.1.2 Operational Modes

A complicated item may have many operational modes, and one or more functions for each operational mode. Operational modes may include normal operating

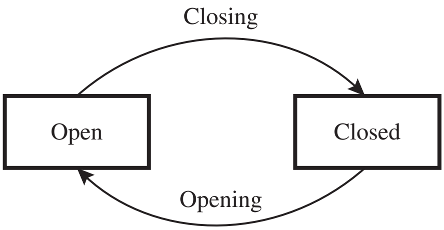


Figure 3.1 States and transitions for a safety valve.

modes, test modes, transition modes, and contingency modes induced by failures or operator errors. The establishment of the different operational modes is recommended for two reasons:

(1) It reveals functions that might be overlooked when focusing too much on the essential functions.
(2) It provides a structured basis for identifying failure modes that are connected to, and dependent on, the given operational mode.

Operational modes are therefore an aid in identifying both functions and failure modes. Failure modes are discussed in Section 3.4.

3.2 Failures

Even if we are able to identify all the required functions of an item, we may not be able to identify all the potential failures. This is because each function may fail in several different ways. No formal procedure seems to exist that help us to identify and classify all the potential failures.

In this section, we consider a specific item within its boundary in its intended operating context. Failure is, in many applications, a complicated and confusing concept. We try to shed some light on this concept and start by defining failure of an item as:

Definition 3.1 (Failure of an item)
The termination of the ability of an item to perform as required. □

The following comments to Definition 3.1 may be given:

(a) Definition 3.1 is mainly a rephrasing of IEV's definition of a failure: "loss of ability to perform as required" (IEV 192-03-01), but the expression "loss of " is replaced with the expression "the termination of" to make it even more clear that a failure is an *event* that takes place at a certain point in time (e.g. at time t_0).
(b) In the context of reliability, the expression "ability to perform as required" does not imply that all aspects of the item are perfect, but that the item must be able to perform the functions that are required for a given purpose.
(c) The item may deteriorate as a slow process. Failure occurs when a required function no longer fulfills its performance requirements, and it may not be any significant change in performance when the threshold is passed, as shown in Example 3.2.

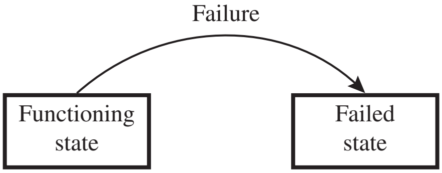


Figure 3.2 Failure as a transition from a functioning state to a failed state.

(d) One user may interpret "as required" different from another user. A failure that is important (and costly) in a warranty context may, for example, be irrelevant in a risk assessment context.
The performance requirements for an item are usually available in the item specification document and partly in the user's manuals, but users seldom read the specifications and the complete user's manual.

(e) We use the verb *fail* to express that a failure occurs. When a failure occurs at time t_0, the item fails at time t_0.

A failure may be interpreted as a *transition* from a *functioning state* to a *failed state*, as shown in Figure 3.2. Example 3.2 illustrates that we may not always be able to observe the failure event and the time t_0 of the failure.

Example 3.2 (Car tires)
When a car is used, the tires wear and the tire tread depth is continuously reduced and thereby the performance of the tires is degrading. When the depth becomes smaller than a certain legal limit d_0 (may be different in the different countries), the tires have to be replaced. A failure occurs when the tread depth passes d_0. In this case, it is not possible to determine exactly the time of failure, and there is no dramatic change of performance when the failure occurs, but the risk of water planning and of puncture is considered to be unacceptable with a smaller depth than d_0. □

3.2.1 Failures in a State

It is sometimes useful to distinguish between failures that occur in a state from failures that occur during a transition. The types of failures occurring in a state are illustrated in Examples 3.3, 3.4, and 3.5.

Example 3.3 (Water pump)
Consider an electric driven water pump. The essential function of the pump is to pump water at a certain rate. Assume that the target rate is 100 l/min, with performance criterion saying that the rate need to be between 95 and 105 l/min. In case of internal fouling, the pumping rate may decrease such that the performance

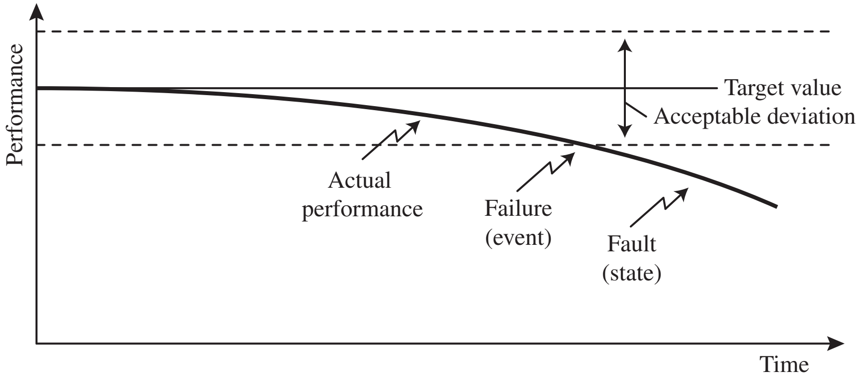


Figure 3.3 Illustration of the difference between failure and fault for a degrading item.

criterion is no longer met. When the rate passes the lower threshold rate, a pump failure occurs and the pump has to be stopped. The pump remains in this state until it has been cleaned/repaired. This process is illustrated in Figure 3.3. □

Example 3.4 (Light bulb–continuously "on")
Consider a light bulb that is always switched on. The function of the bulb is to provide light. When the light bulb fails, the failure occurs in an operating state. If someone is present and can observe the loss of light event, the precise time of the failure can be recorded. □

Example 3.5 (Light bulb–"on" only on demand)
Reconsider a light bulb, similar to the one in Example 3.4, but assume that the light bulb is very seldom switched on and that it each time is energized for a short time period. The bulb may also fail in passive state (e.g. due to vibrations). A failure in passive state is not observable and leaves a *hidden fault*. The hidden fault is not revealed until the light bulb is switched on next time. The time t_0 of the occurrence of the failure is unknown. When we try to switch on the light and observe that it has failed, we only know that the failure occurred in the time interval since the preceding use of the light bulb. (In this example, we assume that the switch is functioning without failure.) □

3.2.2 Failures During Transition

A failure during transition may either be caused by an existing hidden fault or an erroneously performed transition, as illustrated in Examples 3.6 and 3.7.

Example 3.6 (Lawn mower)
Consider a lawn mower with a petrol engine that is started by pulling a rope. To start the lawn mower involves a transition from a passive to an active state of the mower. A failure during this transition may be caused by an internal defect (e.g. corrosion, or contaminated petrol), but may also be due to incorrect starting procedure. □

Example 3.7 (Safety valve)
Reconsider the safety valve in Example 3.1 and assume that the valve is in fully open state when an emergency occurs on the downstream side of the valve. The valve receives a signal to close and the transition is initiated. Due to debris in the valve cavity, the movement is stopped before the valve reaches the closed state. □

3.3 Faults

The term *fault* is mentioned in Section 3.2, but without a proper definition. We define a fault as:

Definition 3.2 (Fault of an item)
A state of an item, where the item is not able to perform as required. □

The duration of the fault may range from negligible to permanent. There are two main types of faults.

Type 1 fault is a fault that occurs as a consequence of a failure. The failure causes a transition from a functioning state into a fault, which is also called a *failed state*. In Example 3.4, the failure of the light bulb left the bulb in a state where it cannot give light. In this example, the bulb has to be replaced to function again.

Type 2 fault is a fault that is introduced in the item due to human error or misjudgment in the specification, design, manufacture, transportation, installation, operation, or maintenance of the item. This type of fault enters the item without any preceding item failure and is a dormant fault that remains hidden until the item is activated or inspected. A type 2 fault is also called a *systematic fault*. A software bug is a typical example of such a fault. Another example is faults caused by design errors or installation errors.

3.4 Failure Modes

We define a failure mode of an item as:

Definition 3.3 (Failure mode)
The manner in which a failure occurs, independent of the cause of the failure. □

A failure mode is a description of how a failure occurs but does not say anything about why the failure occurred. Example 3.8 illustrates how the failure mode concept is usually interpreted.

Example 3.8 (Failure modes of a sink faucet)
Consider a sink faucet used in a bathroom. The main functions of the faucet are to open/close the water supply, to contain the water, and to regulate the water temperature and flow. We consider only the faucet (the item) and assume that cold and hot water are available.

The faucet may have a number of failure modes. Among these are:

(1) Fail to open (on demand) and supply water
(2) Fail to close (on demand) and stop the flow of water
(3) Leakage through the faucet (i.e. dripping)
(4) Leakage out (from faucet seals)
(5) Fail to regulate water flow
(6) Fail to regulate temperature

The faucet has two main states, closed and open. The first two failures (1 and 2) occur during intended transitions between these states. The next two failure modes (3 and 4) occur in a state. For these failure modes, the faucet is in a state where it is leaking and not able to perform as required. The two last failure modes (5 and 6) may be interpreted to be somewhere between the two other types. □

Example 3.9 shows that a failure mode sometimes describes the "manner by which a failure occurs" and sometimes the "manner by which a fault is present."

Example 3.9 (Electric doorbell)
A simple doorbell system is shown in Figure 3.4. The pushbutton activates a switch that closes a circuit from a battery to a solenoid that activates a clapper, which again makes sound by hammering on a bell. When your finger is lifted from the pushbutton, the switch should open, cut the circuit, and thereby stop the doorbell sound. The following failure modes may be defined:

(1) No sound when the pushbutton is activated (by a finger.)
(2) Doorbell sound does not stop when finger is lifted from pushbutton.
(3) Doorbell sounds without activating the pushbutton.

A similar doorbell system is analyzed in NASA (2002). □

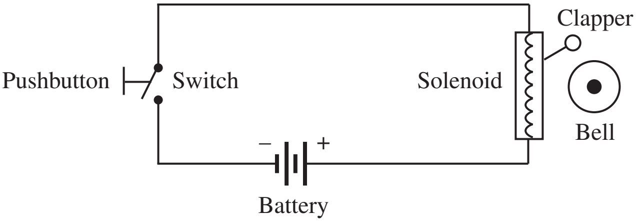


Figure 3.4 Doorbell and associated circuitry.

3.5 Failure Causes and Effects

A failure mode is generally caused by one or more failure causes and may result in a failure effect, as shown in Figure 3.5.

3.5.1 Failure Causes

All failures have at least one cause. We define failure cause as follows.

Definition 3.4 (Failure cause)
Set of circumstances that leads to failure. □

The failure cause may originate during specification, design, manufacture, installation, operation, or maintenance of an item (IEV 192-03-11). The failure cause may be an action, an event, a condition, a factor, a state, or a process that is – at least partly – responsible for the occurrence of a failure. To be responsible for a failure, the cause must be present before the failure occurs, and the presence of the cause should increase the likelihood of the failure.

When studying several similar failures, we should see a positive *correlation* between the presence of the cause and the occurrence of the failure(s), but positive correlation is not a sufficient condition for claiming that something is a cause of a failure. It is very easy to find correlated factors that are totally unrelated. The correlation may, for example, be that the two factors are both caused by the same third factor. Causality is a complicated philosophical subject. A lot more information may be found by searching the Internet. The authors especially recommend consulting (Pearl 2009).

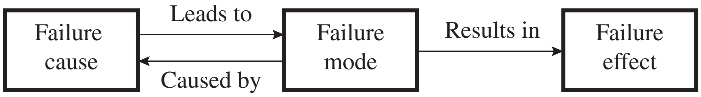


Figure 3.5 Relation between failure causes, failure modes, and failure effects.

Several failure analysis techniques have been developed to identify the causes of a failure that has occurred. Among these are *cause and effect analysis* and *root cause analysis* that are described in Section 3.7.

3.5.2 Proximate Causes and Root Causes

The term *root cause* is often used in analyses of failures that have occurred. The term is defined in several standards, and each standard seems to have its own particular definition. Before giving our preferred definition, we define the term *proximate cause*, which is an immediately and (often) readily seen cause of a failure.

Definition 3.5 (Proximate cause)
An event that occurred, or a condition that existed immediately before the failure occurred, and, if eliminated or modified, would have prevented the failure. □

A proximate cause is also known as a *direct cause*. A proximate cause is often not the real (or root) cause of a failure, as illustrated in Example 3.10.

Example 3.10 (Flashlight)
A flashlight is part of the safety equipment in a plant. During an emergency, the flashlight is switched on, but does not give any light. A proximate (or direct) cause is that the battery is dead. If we have access to the flashlight and the battery after the emergency is over, it is straightforward to verify whether or not this was the true proximate cause.

Any battery will sooner or later go dead and if the flashlight is an essential safety equipment, it is part of the maintenance duties to test and, if necessary, replace batteries at regular intervals. "The battery has not been tested/replaced at prescribed intervals" is therefore a cause of the proximate cause. By asking "why?" this happened several times, we may get to the root cause of the failure. □

For the purpose of this book, we define a root cause as:

Definition 3.6 (Root cause)
One of multiple factors (events, conditions, or organizational factors) that contributed to or created the proximate cause and subsequent failure and, if eliminated, or modified would have prevented the failure. □

For some failure modes, it may be possible to identify a single root cause, but most failure modes will have several contributing causes. All too often, failures are attributed to a proximate cause, such as human error or technical failure.

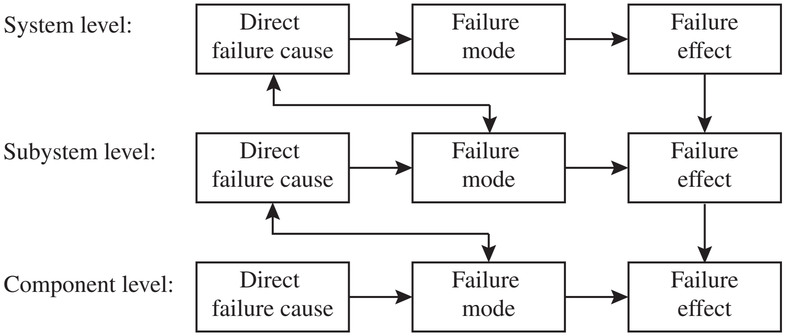


Figure 3.6 Relationship between failure cause, failure mode, and failure effect.

These are often merely *symptoms*, and not the root causes of the failure. Very often, the root causes turn out to be much more, such as (i) process or program deficiencies, (ii) system or organization deficiencies, (iii) inadequate or ambiguous work instructions, and/or (iv) inadequate training.

To identify root causes of failures and to rectify these is important for any system in the operational phase. It does not help only to correct the proximate causes (such as to replace the battery of the flashlight in Example 3.10) when a failure has occurred. This way, the same failure may recur many times. If, on the other hand, the root cause is rectified, the failure may never recur. Root cause analysis is briefly discussed in Section 3.7.

3.5.3 Hierarchy of Causes

The functions of a system may usually be split into subfunctions. Failure modes at one level in the hierarchy may be caused by failure modes on the next lower level. It is important to link failure modes on lower levels to the main top level responses, in order to provide traceability to the essential system responses as the functional structure is refined. This is shown in Figure 3.6 for a hardware structure breakdown. Figure 3.6 is further discussed in Section 3.6.5.

3.6 Classification of Failures and Failure Modes

It is important to realize that a failure mode is a manifestation of the failure as seen from the outside, that is, the nonfulfillment of one or more functions. “Internal leakage” is thus a failure mode of a shutdown valve because the valve loses its required function to “close flow,” whereas wear of the valve seal represents a cause of failure and is hence not a failure mode of the valve.

Failures and failure modes may be classified according to many different criteria. We briefly mention some of these classifications.

3.6.1 Classification According to Local Consequence

Blache and Shrivastava (1994) classify failures according to the completeness of the failure.

(1) *Intermittent failure*. Failure that results in the loss of a required function only for a very short period of time. The item reverts to its fully operational standard immediately after the failure.
(2) *Extended failure*. Failure that results in the loss of a required function that will continue until some part of the item is replaced or repaired. An extended failure may be further classified as:
    (a) *Complete failure*. Failure that causes complete loss of a required function.
    (b) *Partial failure*. Failure that leads to a deviation from accepted item performance but do not cause a complete loss of the required function.

    Both the complete failures and the partial failures may be further classified as:

    (a) *Sudden failure*. Failure that could not be forecast by prior testing or examination.
    (b) *Gradual failure*. Failure that could be forecast by testing or examination. A gradual failure represents a gradual "drifting out" of the specified range of performance values. The recognition of a gradual failure requires comparison of actual item performance with a performance requirement, and may in some cases be a difficult task.

    Extended failures may be split into four categories; two of these are given specific names:

    (a) *Catastrophic failures*. A failure that is both sudden and complete.
    (b) *Degraded failure*. A failure that is both partial and gradual (such as the wear of the tires on a car).

The failure classification described above is shown in Figure 3.7, which is adapted from Blache and Shrivastava (1994).

3.6.2 Classification According to Cause

Failures may be classified according to their causes as follows.

Primary Failures

A primary failure, also called a *random hardware failure* in IEC 61508, occurs when the item is used in its intended operating context. In most cases, the primary

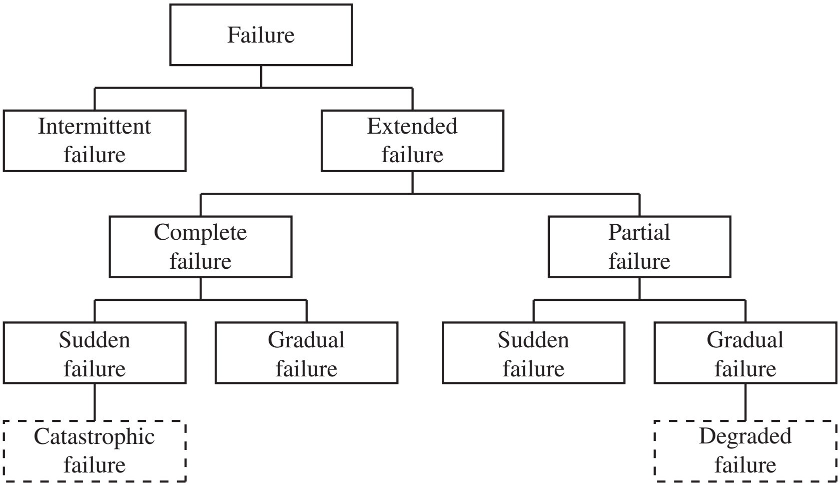


Figure 3.7 Failure classification. Source: Adapted from Blache and Shrivastava (1994).

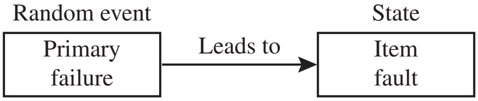


Figure 3.8 A primary failure leading to an item fault.

failure results in an item fault and a repair action is usually necessary to return the item to a functioning state. Primary failures are generally random failures, where the cause of failure can be attributed to aging and the properties of the item itself. A primary failure is illustrated in Figure 3.8. Primary failures are the only category of failures that we justifiably can claim compensation for under warranty. Primary failures are not relevant for software.

Secondary Failures

A secondary failure, also called *overstress* or *overload failure*, is a failure caused by excessive stresses outside the intended operating context of the item. Typical stresses include shocks from thermal, mechanical, electrical, chemical, magnetic, or radioactive energy sources, or erroneous operating procedures. The stresses may be caused by neighboring items, the environment, or by users/system operators/plant personnel. Environmental stresses, such as lightning, earthquake, and falling object, are sometimes called *threats* to the item. We may, for example, say that lightning is a threat to a computer system and that heavy snowfall and storm are threats to an electric power grid. The overstress event leads to a secondary failure with some probability p that depends on the stress level and on the *vulnerability* of the item. Overloads of software systems may also be classified as secondary failures.

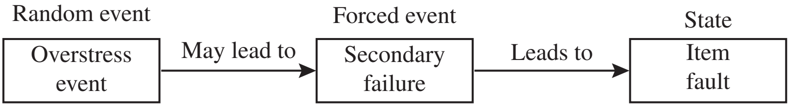


Figure 3.9 A secondary failure, caused by an overstress event, leading to an item fault.

A secondary failure usually leads to an item fault, and a repair action is usually necessary to return the item to a functioning state. The structure of a secondary failure is shown in Figure 3.9. Secondary failures are generally random events, but it is the overstress event that is the main contributor to the randomness.

Systematic Failures

A systematic failure is a failure due to a *systematic cause* that may be attributed to a human error or misjudgment in the specification, design, manufacture, installation, operation, or maintenance of the item. A software bug is a typical example of a systematic fault. After the error is made, the systematic cause remains dormant and hidden in the item. Examples of systematic causes are given in Example 3.12.

A systematic failure occurs when a certain *trigger* or *activation condition* occurs. The trigger can be a transient event that activates the systematic cause, but can also be a long-lasting state such as environmental conditions, as illustrated in Example 3.14. The trigger event is often a random event, but may also be deterministic.

A systematic failure can be reproduced by deliberately applying the same trigger. The term *systematic* means that the same failure will occur whenever the identified trigger or activation condition is present and for all identical copies of the item. A systematic cause can only be eliminated by a modification of the design or of the manufacturing process, operational procedures, or other relevant factors (IEC 61508 2010). A systematic fault leading to a systematic failure by the "help" of a trigger is shown in Figure 3.10. Systematic failures are often, but not always, random events, but it is the trigger that is random, whereas the item failure is a consequence of the trigger event.

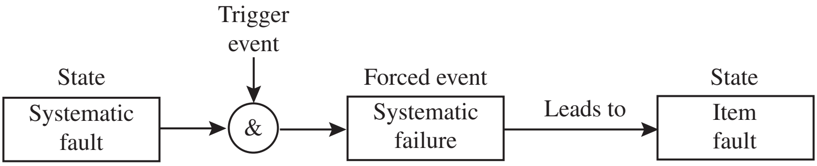


Figure 3.10 A systematic fault leading to a systematic failure.

Example 3.11 (Airbag system in a car)
A new car model was launched and a person driving such a car crashed into another car. The airbags did not operate as intended and the driver was critically injured. After the accident, it was found that the airbag system was not correctly installed. Later, it was found that the same error was made for all cars of the same type. The airbag failure was due to a systematic cause and all the cars of the same type had the same systematic fault. All these cars had to be recalled for repair and modification. There was nothing wrong with the airbag system as such and the airbag system manufacturer could not be blamed for the accident (unless the installation instructions were misleading or ambiguous). The car manufacturer had to cover the consequences of the failure. For drivers and passengers, the cause of the failure does not matter. A systematic failure has the same consequences as a primary (random hardware) failure. □

Example 3.12 (Failure causes of a gas detection system)
A heavy (i.e. heavier than air) and dangerous gas is used in a chemical process. If a gas leakage occurs, it is important to raise an alarm and shut down the process as fast as possible. For this purpose, a safety-instrumented system (SIS) is installed, with one or more gas detectors. The SIS has three main parts (i) gas detectors, (ii) a logic solver that receives, interprets, and transmits signals, and (iii) a set of actuating items (e.g. alarms, shutdown valves, door closing mechanisms). The purpose of the SIS is to give an automatic and rapid response to a gas leakage. Many more details about SIS may be found in Chapter 13.

Assume that a gas leak has occurred without any response from the SIS. Possible causes of the failure may include the following:

- A primary (i.e. random hardware) failure of the SIS.
- The installed gas detectors are not sensitive to this particular type of gas, or have been mis-calibrated.
- The gas detectors have been installed high up on walls or in the ceiling (remember, the gas is heavier than air.)
- The gas detectors have been installed close to a fan (no gas will reach them.)
- The gas detectors have been inhibited during maintenance (and the inhibits have not been removed.)
- The gas detector does not raise alarm due to a software bug. (Most modern gas detectors have software-based self-testing features.)
- The gas detector is damaged by, for example, sand-blasting. (Has happened several times in the offshore oil and gas industry.) □

Security Failures

A security failure is a failure caused by a deliberate human action. Many systems are exposed to a number of *threats*. The threats may be related to physical actions or cyberattacks. Physical threats include arson, sabotage, theft, and many more. A cyberattack is only relevant for systems that are connected to a cyber network (e.g. Internet, or mobile phone network). A threat may be used by a *threat actor* to attack the system. The system may have a number of *vulnerabilities* (i.e. weaknesses) that may be exploited by the threat actor to make a "successful" attack.

With the development of new technologies, such as cyber-physical systems, the Internet of Things (IoT), smart-grids, smart cities, remote operation and maintenance, and many more, cyberattacks come more frequently and we can now hardly open a newspaper without articles about cyberattacks. Many of these attacks are directed toward critical infrastructure and industrial control and safety systems.

The structure of a security failure is illustrated in Figure 3.11. A threat, a threat actor, and a vulnerability are required "inputs" for a security failure. The threat actor uses a threat to attack the system, and the threat inspires the threat actor. The attack can only be successful if the system has one or more vulnerabilities.

A security failure is not a random event, but the consequence of a deliberate action made by the threat actor. To reduce the likelihood of security failures, vulnerabilities should be identified and removed during system design.

Additional Types of Failures

When an item fails, the failure is often claimed to be caused by the control of the item, the input/output to/from the item, or misuse of the item. These causes are usually outside the boundary of the item and not something the manufacturer of the item can be responsible for.

Control failures. A control failure is an item failure caused by an improper control signal or noise, that is, due to factors outside the boundary of the item. A repair action may or may not be required to return the item to a functioning state. Failures caused by inadequate, or not followed operating procedures may also be classified as control failures.

Input/output failures. An input/output failure is a failure caused by inadequate or lacking item inputs or outputs, that is, due to factors outside the boundary of the item. For a washing machine, the washing service is stopped due to inadequate

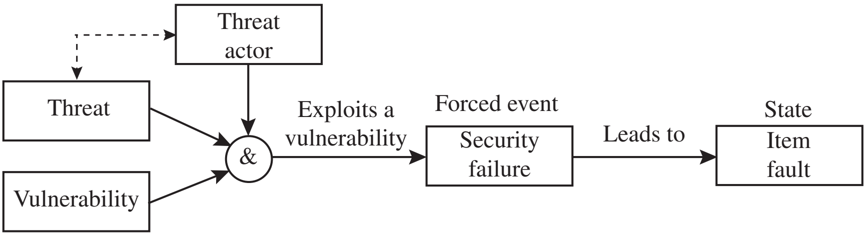


Figure 3.11 The structure of a security failure.

or lacking supply of electricity, water, or detergent, or due to inadequacies of the drainage system. Input/output failures will stop the service provided by the item but will usually not leave the item in a failed state. The item may not need any repair after an input/output failure. Input/output failures tell very little about the reliability of the item as such.

Misuse/mishandling failure. A misuse/mishandling failure is a failure that occurs because the item is used for a purpose that it was not designed for, or is mishandled. The mishandling may be due to a human error or a deliberate action such as sabotage. Some laws and standards (e.g. EU-2006/42/EC) require that *foreseeable misuse* shall be considered and compensated for in the design and development of the item, and be covered in the operating context of the item.

The categories of failures listed above are not fully mutually exclusive. Some control failures may, for example, also be due to systematic causes.

Remark 3.2 (Functionally unavailable)

The US Nuclear Regulatory Commission (NRC) introduces the term *functionally unavailable* for an item that is capable of operation, but where the function normally provided by the item is unavailable due to lack of proper input, lack of support function from a source outside the component (i.e. motive power, actuation signal), maintenance, testing, the improper interference of a person, and so on.

The NRC-term is seen to cover failures/faults of several of the categories above, most notably input/output and control failures. □

Failures Named According to the Cause of Failure

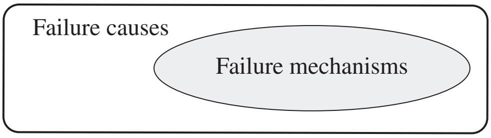
Failures are sometimes named according to (i) the main cause of the failure, such as corrosion failure, fatigue failure, aging failure, calibration failure, systematic failure, and so forth, (ii) the type of technology that fails, such as mechanical failure, electrical failure, interface failure, and software bug, and (iii) the life cycle phase in which the failure cause originates, such as design failure, manufacturing failure, and maintenance failure.

When using this type of labeling, we should remember that the failure description does not tell how the failure is manifested, that is, which failure mode that occurs. The same failure mode may occur due to many different failure causes.

3.6.3 Failure Mechanisms

A failure mechanism is a physical, chemical, logical, or other process or mechanism that may lead to failure. Examples of failure mechanisms include wear, corrosion, fatigue, hardening, swelling, pitting, and oxidation. Failure mechanisms are hence specific failure causes as shown in Figure 3.12.

Figure 3.12 Failure causes and mechanisms. A failure mechanism is a specific type of failure cause.



Each mechanism can have its root in different stages of the item's life cycle. Wear can, for instance, be a result of wrong material specification (design failure), usage outside specification limits (misuse failure), poor maintenance, inadequate lubrication (mishandling failure), and so on.

A failure mechanism may be seen as a process that leads to a failure cause.

3.6.4 Software Faults

An increasing number of item functions are being replaced by software-based functions and a fair proportion of item failures are caused by *software bugs*. IEV defines a software fault/bug as:

Definition 3.7 (Software fault/bug)
State of a software item that prevents it from performing as required (IEV 192-04-02). □

Combined with a particular demand or trigger, the software bug may lead to item failure. Such a failure is a systematic failure and is sometimes called a software failure (see Figure 3.10). If the trigger is a random event, the software failure is random. Software bugs are difficult to reveal and software development projects therefore include a detailed process for finding and correcting bugs. This process is called *debugging*.

Software does not deteriorate and software bugs do not occur at random in the operational phase. They have been programmed into the software and remain until the software is modified. New software bugs are often introduced when new patches or new versions of the software are installed to remove known bugs. The same software failure occurs each time the same activation condition or trigger occurs. If relevant activating conditions or triggers do not occur, the software bug remains undetected. Installations of the same software may show very different frequencies of software failures because the failure frequency is proportional to the frequency of the occurrence of activating conditions or triggers.

3.6.5 Failure Effects

Failure effect is an undesired consequence of a failure mode. Failure effects may be categorized as follows:

(1) Injuries or damage to personnel or to the public.

(2) Damage to the environment.
(3) Damage to the system where the failure occurred.
(4) Material or financial loss.
(5) Interruptions of the system operation (e.g. loss of production, cancelled or delayed transport means, interruptions of electric or water supply, interruption of computer/telephone network service.)

A failure mode may lead to many different failure effects, on the item where the failure occurred, and on other items. Failure effects are classified as local effects, next higher effects, and end effects. These effects are illustrated in Example 3.13.

Example 3.13 (Failure effects of brake pad failure)
Consider a (total) wear-out failure of a brake pad on the left front wheel of a car. The local effect is that the braking effect on the left front wheel is strongly reduced and that the brake disc may be damaged. The next higher effect is that the braking effect of the car is uneven and not adequate. The end effect is that the car cannot provide a safe drive and must be stopped. □

A general picture of the relationship between cause and effect is that each failure mode can be caused by several different failure causes, leading to several different failure effects. To get a broader understanding of the relationship between these terms, the level of indenture being analyzed should be brought into account. This is shown in Figure 3.6.

Figure 3.6 shows that a failure mode on the lowest level of indenture is one of the failure causes on the next higher level of indenture, and the failure effect on the lowest level equals the failure mode on the next higher level. The failure mode "leakage from sealing" for the seal component is, for example, one of the possible failure causes for the failure mode "internal leakage" for the pump, and the failure effect (on the next higher level) "internal leakage" resulting from "leakage from sealing" is the same as the failure mode "internal leakage" of the pump.

Failure effects are often classified according to their *criticality* as discussed in Chapter 4.

3.7 Failure/Fault Analysis

A failure or fault analysis is a systematic investigation of a failure or a fault that has occurred, in order to identify the root causes of the failure/fault and to propose corrective actions needed to prevent future failures/faults of the same, or similar, types.

This section gives an introduction to two commonly used failure/fault analysis techniques (i) cause and effect analysis and (ii) root cause analysis. Both techniques are primarily used to analyze real failures/faults that have occurred, but may also be used to analyze potential failures or faults.

3.7.1 Cause and Effect Analysis

Cause and effect analyses are frequently used in quality engineering to identify and illustrate possible causes of quality problems. The same approach may also be used in reliability engineering to find the potential causes for system failures or faults. The cause and effect analysis is documented in a *cause and effect diagram*.

The cause and effect diagram, also called Ishikawa diagram (Ishikawa 1986), was developed in 1943 by the Japanese professor Kaoru Ishikawa (1915–1989). The diagram is used to identify and describe all the potential causes (or events) that may result in a specified failure. Causes are arranged in a tree structure that resembles the skeleton of a fish with the main causal categories drawn as *bones* attached to the spine of the fish. The cause and effect diagram is therefore also known as a *fishbone* diagram.

To construct a cause and effect diagram, we start with an item failure. The item failure is briefly described, enclosed in a box and placed at the right end of the diagram, as the "head of the fish." The analysis is carried out by a team, using an idea-generating technique, such as *brainstorming*. Failure causes are suggested by the team and organized under headings such as

(1) Manpower
(2) Methods
(3) Materials
(4) Machinery
(5) Milieu (environment)

This is a common classification for failure/fault analysis and is referred to as the 5M approach, but other categories may also be used. The main structure of a 5M cause and effect diagram is shown in Figure 3.13.

When the team members agree that an adequate amount of detail has been provided under each major category, they analyze the diagram, and group the causes. An important part of this analysis is to eliminate irrelevant causes from the diagram and tidy it up. One should especially look for causes that appear in more than one category. For those items identified as the "most likely causes," the team should reach consensus on listing those causes in priority order with the first cause being the "most likely cause."

Some cause and effect analyses also include an evaluation of how easy it is to verify each of the identified causes in the diagram. Three classes are sometimes

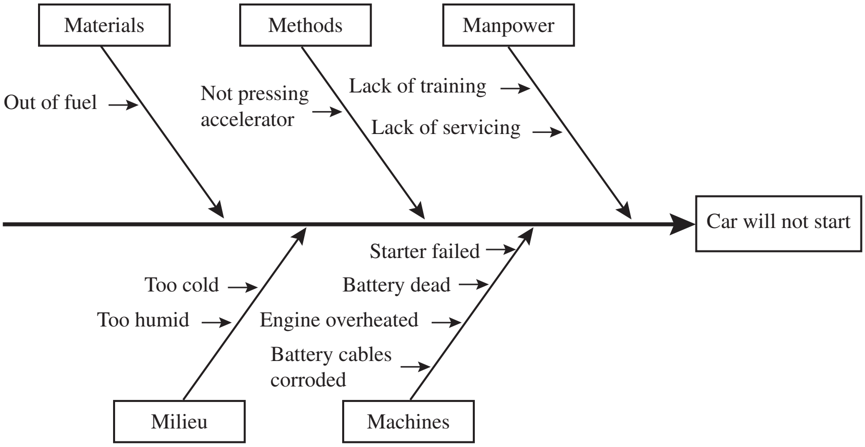


Figure 3.13 Cause and effect diagram for the event "car will not start."

used: (i) very easy, (ii) somewhat easy, and (iii) not easy. A final step to propose actions to rectify the identified causes, may or may not be included in the analysis.

The cause and effects diagram cannot be used for quantitative analyses, but is generally considered to be an excellent aid for problem solving, and to illustrate the potential causes of an item failure/fault. Cause and effect analysis is also a recommended step in a more comprehensive root cause analysis (see Section 3.7.2).

Example 3.14 (Car will not start)
Consider a car that will not start after having been idle for a period. The causes suggested by the team are shown in the cause and effect diagram in Figure 3.14. A number of similar cause and event diagrams may be found on the Internet. □

3.7.2 Root Cause Analysis

A root cause analysis may be defined as:

Definition 3.8 (Root cause analysis)
A systematic investigation of a failure or a fault to identify its likely root causes, such that they can be removed by design, process, or procedure changes. □

The root cause analysis is reactive, starting with (i) a failure that has happened, or (ii) a potential failure that has been identified. The root cause analysis should continue until organizational factors have been identified, or until data are exhausted. Root cause analysis may be used to investigate a wide range of undesired events, not only failures and faults but also our description is delimited to failure/fault analysis.

The main steps of a root cause (failure) analysis are:

(1) Clearly define the failure or fault. Explain clearly what went wrong.
(2) Gather data/evidence. The evidence should provide answers to the following questions:
- When did the failure occur?
- Where did it occur?
- What conditions were present prior to its occurrence?
- What controls or barriers could have prevented its occurrence but did not?
- What are the potential causes? (Make a preliminary list of likely causes).
- Which actions can prevent recurrence?

(3) Ask why and identify the true root cause associated with the defined failure/fault.
(4) Check the logic and eliminate items that are not causes.
(5) Identify corrective action(s) that will prevent recurrence of the failure/fault – and that address both proximate and root causes.
(6) Implement the corrective action(s).
(7) Observe the corrective actions to ensure effectiveness.
(8) If necessary, reexamine the root cause analysis.

The root cause analysis is done by a team using idea generation techniques, such as brainstorming, and is often started by a cause and effect analysis link: (see Section 3.7.1). To identify root causes, it is usually recommended to ask "why?" at least five times for each main cause identified. The five whys are illustrated in Figure 3.14.

The root causes must be thoroughly understood before corrective actions are proposed. By correcting root causes, it is hoped that the likelihood of failure recurrence is minimized.

Example 3.15 (Car will not start)

Reconsider the car that will not start in Example 3.14. The following sequence of five questions and answers may illustrate the analysis process.

(1) Why will not the car start?
Cause: The engine will not turn over.

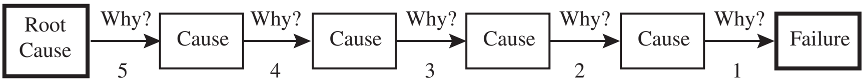


Figure 3.14 Repeatedly asking why?

(2) Why will the engine not turn over?
Cause: The battery is dead.

(3) Why is the battery dead?
Cause: The alternator is not functioning.

(4) Why is the alternator not functioning?
Cause: The belt is broken.

(5) Why is the alternator belt broken?
Cause: The belt was not replaced according to the manufacturer's maintenance schedule.

This example is strongly influenced by the presentation "Corrective action and root cause analysis" by David S. Korcal (found on the Internet). □

Careful studies of failures that occur should add to our "lessons learned," and we therefore end this chapter optimistically by quoting Henry Ford (1863–1947):

> Failure is the opportunity to begin again more intelligently.

3.8 Problems

3.1 Consider the exterior door of a family house. The door is locked/unlocked by using a standard key.
(a) List all relevant functions of the door (including lock).
(b) List all relevant failure modes of the door.
(c) Classify the failure modes by using the classification system outlined in this chapter.
(d) Do you consider it relevant to include *misuse failures*? If "yes," provide examples.

3.2 Consider a filter coffee maker/brewer that you are familiar with.
(a) List all potential failure modes of the coffee brewer.
(b) Identify potential causes of each failure mode.
(c) Identify potential effects of each failure mode.

3.3 Identify and describe possible failure modes of a (domestic) refrigerator.

3.4 Assume that your mobile phone is "dead." Illustrate the possible causes of this fault by a cause and effect diagram.

3.5 Consider a smoke detector used in a private home and list possible causes of systematic faults of this detector.

3.6 Explain the differences between the terms *failure* and *fault*. Illustrate you explanation by practical examples.

3.7 Consider a domestic washing machine.
(a) Identify as many causes of potential failures as possible.
(b) Define categories of failure causes.
(c) Use these categories to classify the identified failure causes.

3.8 Suggest a technical system that can be divided into several levels of indenture. If you cannot propose anything better, you may use a family car. Assume that a specific component failure mode occurs in the system and exemplify the relationships that are illustrated in Figure 3.6.

3.9 Reconsider the coffee maker in Problem 3.2. When you press the on/off switch, no coffee is supplied.
(a) Analyze the "failure" by using a cause and effect diagram.
(b) Analyze the same "failure" by a root cause analysis.

References

Blache, K.M. and Shrivastava, A.B. (1994). Defining failure of manufacturing machinery and equipment. *Proceedings Annual Reliability and Maintainability Symposium*, pp. 69–75.

EU-2006/42/EC (2006). Council Directive 2006/42/EC of 17 May 2006 on machinery. *Official Journal of the European Union*, L 157/24 (2006). Brussels.

IEC 61508 (2010). *Functional safety of electrical/electronic/programmable electronic safety-related systems*. Parts 1-7, *International standard*. Geneva: International Electrotechnical Commission.

Ishikawa, K. (1986). *Guide to Quality Control*. White Plains, NY: Asian Productivity Organization – Quality Resources.

NASA (2002). Fault Tree Handbook with Aerospace Applications, *Handbook*. Washington, DC: U.S. National Aeronautics and Space Administration.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*, 2e. Cambridge: Cambridge University Press.

4

Qualitative System Reliability Analysis

4.1 Introduction

This chapter presents five different approaches/methods for qualitative system reliability analysis.

(1) *Failure modes, effects, and criticality analysis (FMECA).* This is a common approach to identify the potential failure modes of system components and subsystems, to identify the causes of each failure mode, and to study the effects these failure modes might have on the system. FMECA was developed as a tool for designers, but it is frequently used as a basis for more detailed reliability analyses and for maintenance planning.

(2) *Fault tree analysis (FTA).* A fault tree illustrates all possible combinations of potential failures and events that may cause a specified system failure. Fault tree construction is a deductive approach where we start with the specified system failure and ask "what are the causes for this failure?" Failures and events are combined through logic gates in a binary approach. The fault tree may be evaluated quantitatively if we have access to probability estimates for the basic events. Quantitative FTA is discussed in Chapter 6.

(3) *Event tree analysis (ETA).* ETA is an inductive method that starts with a system deviation and identifies how this deviation may develop. The possible events following the deviation will usually depend on the various barriers and safety functions that are designed into the system. Quantitative ETA is discussed briefly in Chapter 6.

(4) *Reliability block diagrams (RBDs).* RBDs were introduced in Section 2.8. In this chapter, the structure of the RBD is described mathematically by structure functions. Structure functions are used in the following chapters to calculate system reliability indices. Further quantitative RBD analysis is discussed in Chapter 6.

System Reliability Theory: Models, Statistical Methods, and Applications, Third Edition.
Marvin Rausand, Anne Barros, and Arnljot Høyland.

Companion Website: www.wiley.com/go/SystemReliabilityTheory3e

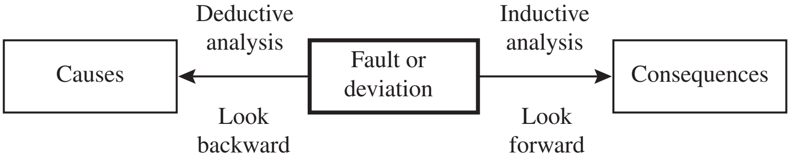


Figure 4.1 Deductive versus inductive analysis of a fault or deviation in the study object.

Table 4.1 Deductive versus inductive methods.

Model/method	Deductive	Inductive
FMECA	Δ	Δ
Fault tree analysis	X	–
Event tree analysis	–	X
Reliability block diagrams	X	–
Bayesian networks	X	X

(5) *Bayesian networks (BNs).* A BN is a directed acyclic graph (DAG) that can replace and extend traditional fault trees and event trees and accommodate causal dependencies between items. Quantitative BN analysis is discussed in Chapter 6.

4.1.1 Deductive Versus Inductive Analysis

The methods in this chapter start with a defined fault or deviation in the study object. With this starting point, we may look backwards and try to identify the causes of the fault or deviation. This is done by a *deductive analysis* that backwardly deduces the causes of the fault or deviation. Alternatively, we may start with the same fault or deviation and look forward and try to figure out the potential consequences of the fault or deviation. This is done by an *inductive analysis* that forwardly induces the consequences. The two approaches are illustrated schematically in Figure 4.1.

Some of the five methods listed above are deductive, others are inductive, and some have elements (Δ) of both, as indicated in Table 4.1.

4.2 FMEA/FMECA

The first *failure mode and effects analysis* (FMEA) guideline was published as early as 1949 (see Section 1.10), and FMEA is still the most commonly used method for potential failure analysis. FMEA reviews components, assemblies, and subsystems to identify potential failure modes, their causes, and effects. For each component, the failure modes and their resulting effects on the rest of the system are recorded in a specific FMEA worksheet. There are numerous variants of such worksheets. A typical example is shown in Figure 4.4.

An FMEA becomes a *failure mode, effects, and criticality analysis* (FMECA) if criticality or priority is assigned to each failure mode effect. In the following, we do not distinguish between FMEA and FMECA, and use the term "FMECA" for both. More detailed information on how to conduct FMECA may be found in several standards, such as IEC 60812 (2018), MIL-STD-1629A (1980), SAE ARP 5580 (2012), and SAE J1739 (2009).

4.2.1 Types of FMECA

FMECAs come in many flavors, depending on the study object and in which phase of its life cycle the analysis is performed. The following four types are, for example, used in the automotive industry (SAE J1739 2009):

(1) *Concept FMECA* analyzes new product concepts in the concept and early design phases.
(2) *Design FMECA* analyzes products before they are released to production.
(3) *Machinery FMECA* analyzes special machinery (equipment and tools) that allows for customized selection of component parts, machine structure, tooling, bearings, coolants, and so on.
(4) *Process FMECA* analyzes manufacturing and assembly processes.

Additional Variants of FMECA

Several new variants of FMECA have been developed for specific purposes:

- *Interface FMECA* analyzes potential problems related to the interfaces between components or subsystems.
- *Software FMECA* identifies and prevents potential bugs in software (e.g. see Haapanen and Helminen 2002).
- *FMEDA* (failure modes, effects and diagnostic analysis) analyzes systems that have built-in diagnostic testing and is especially applied to safety-instrumented systems (e.g. see Goble and Brombacher 1999).
- *FMVEA* (failure modes, vulnerabilities, and effects analysis) identifies and prevents system vulnerabilities that may be exploited by threat actors (e.g. see Schmittner et al. 2014).
- *CyberFMECA* has a similar purpose as FMVEA.

A timeline of the development of FMECA and the additional variants listed above is shown in Figure 4.2.

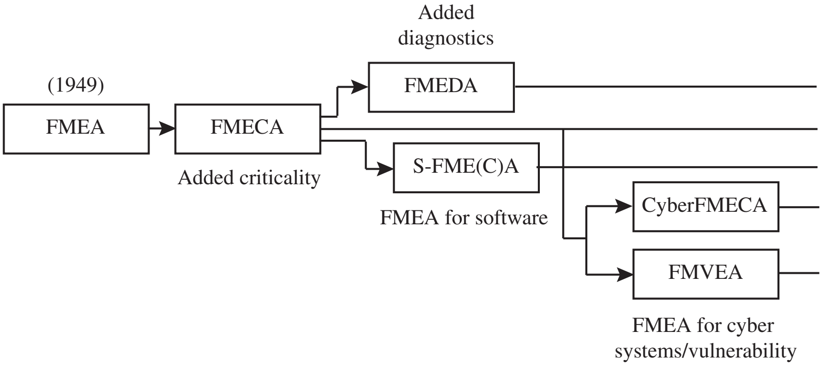


Figure 4.2 Timeline of the development of FMECA variants (not in scale).

Hardware Versus Functional Approach

Two main approaches may be chosen for FMECA of technical items. These are

Hardware FMECA is used to analyze existing systems and system concepts. The individual components on the lowest level in the system hierarchy are analyzed to identify potential failure modes, their causes and effects. When the components on the lowest level are analyzed, we move to the next upper level in the hierarchy, and so on. Hardware FMECA is said to be carried out as a *bottom-up* approach.

Functional FMECA is mainly used in the early design phases of a system. The analysis starts with a top-level system function, and we ask How can this function conceivably fail, what could the causes be, and what could the consequences be? The same procedure is followed for each functional failure. Functional FMECA is said to be carried out as a *top-down* approach.

The rest of this section is delimited to presenting hardware FMECA used for design analysis. Other applications are similar, and the reader should be able to make the appropriate adjustments.

4.2.2 Objectives of FMECA

The objectives of a hardware FMECA in the design phase are the following: (IEEE Std. 352):

(1) Assist in selecting design alternatives with high reliability and high safety potential during the early design phase.
(2) Ensure that all conceivable failure modes and their effects on operational success of the system have been considered.
(3) List potential failures and identify the magnitude of their effects.

(4) Develop early criteria for test planning and the design of the test and checkout systems.
(5) Provide a basis for quantitative reliability and availability analyses.
(6) Provide historical documentation for future reference to aid in analysis of field failures and consideration of design changes.
(7) Provide input data for tradeoff studies.
(8) Provide basis for establishing corrective action priorities.
(9) Assist in the objective evaluation of design requirements related to redundancy, failure detection systems, fail-safe characteristics, and automatic and manual override.

FMECA is mainly a qualitative analysis and should be carried out by the designers during the design phase of a system. The purpose is to identify design areas where improvements are needed to meet reliability requirements. An updated FMECA is an important basis for design reviews and inspections, and also for maintenance planning.

4.2.3 FMECA Procedure

FMECA does not require any advanced analytical skills, but the analysts need to be familiar with and understand the purpose of the study object and the constraints under which it has to operate. FMECA is carried out as a sequence of seven main steps, as shown in Figure 4.3. The number and content of the steps depend on the application and the delimitations of the analysis. Further details for FMECA in the automobile industry may be found in Ford (2004).

The various entries in the FMECA worksheet are best illustrated by going through a specific worksheet column by column. We use the FMECA worksheet in Figure 4.4 as an example.

(1) *Reference*. The name/tag of the item or reference to a drawing is given in the first column.
(2) *Function*. The function(s) of the item is (are) described in this column.
(3) *Operational mode*. The item may have various operational modes, for example, running or standby. Operational modes for an airplane include, for example, taxi, take-off, climb, cruise, descent, approach, flare-out, and

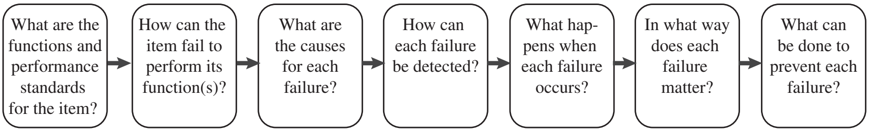
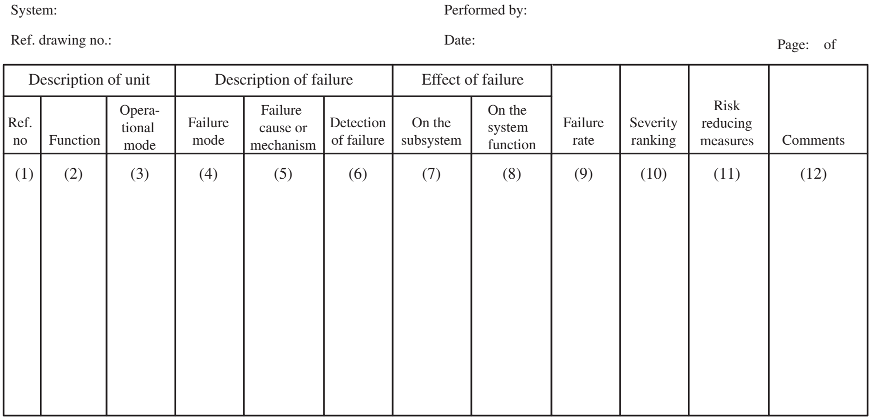

What are the functions and performance standards for the item?
How can the item fail to perform its function(s)?
What are the causes for each failure?
How can each failure be detected?
What happens when each failure occurs?
In what way does each failure matter?
What can be done to prevent each failure?

Figure 4.3 The mains steps of FMECA.



System:

Performed by:

Ref. drawing no.:

Date:

Page: of

Description of unit			Description of failure			Effect of failure					
Ref. no	Function	Opera-tional mode	Failure mode	Failure cause or mechanism	Detection of failure	On the subsystem	On the system function	Failure rate	Severity ranking	Risk reducing measures	Comments
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)

Figure 4.4 Example of an FMECA worksheet.

roll. In applications where it is not relevant to distinguish between operational modes, this column may be omitted.

(4) *Failure mode*. For each component's function and operational mode, all the failure modes are identified and recorded. Observe that the failure modes should be defined as nonfulfillment of the functional requirements of the functions specified in column 2.

(5) *Failure causes and mechanisms*. The possible failure mechanisms (corrosion, erosion, fatigue, etc.) that may produce the identified failure modes are recorded in this column. Other failure causes should also be recorded. To identify all potential failure causes, it may be useful to remember the interfaces shown in Figure 2.2.

(6) *Detection of failure*. The way to detect each failure mode is then recorded. Options include alarms, testing, human perception, and so forth. The ability to detect a failure mode is sometimes rated.

(7) *Effects on other components in the same subsystem*. All the main effects of the identified failure modes on other components in the subsystem are recorded.

(8) *Effects on the function of the system*. All the main effects of the identified failure mode on the function of the system are then recorded. The resulting operational status of the system after the failure may also be recorded, that is, whether the system is functioning or not, or is switched over to another operational mode.

Remark 4.1 (Safety and availability)
In some applications, it may be relevant to replace columns 7 and 8 by, for example, *Effect on safety* and *Effect on availability*. □

Table 4.2 Occurrence rating (example).

Frequent	Once per 1 mo or more often
Probable	Once per 1 yr
Occasional	Once per 10 yr
Remote	Once per 100 yr
Very unlikely	Once per 1000 yr or more seldom

(9) *Failure rate*. Failure rates for each failure mode are then recorded. In many cases, it is more suitable to classify the failure rate in classes, such as shown in Table 4.2.

Observe that the failure rate with respect to a failure mode might be different for the various operational modes. The failure mode "Leakage to the environment" for a valve may, as an example, be more likely when the valve is closed and pressurized, than when the valve is open.

(10) *Severity*. The severity of a failure mode is the potential consequence of the failure, determined by the degree of injury, property damage, or system damage that could ultimately occur. The ranking categories in Table 4.3 are sometimes used.

(11) *Risk reduction measures*. Possible actions to correct the failure and restore the function or prevent serious consequences are recorded. Actions that are likely to reduce the frequency of the failure modes may also be recorded.

(12) *Comments*. This column may be used to record pertinent information not included in the other columns.

By combining the failure rate (column 9) and the severity (column 10), the criticality of the different failure modes may be ranked. This ranking is shown in Figure 4.5 as a *risk matrix*. In this example, the failure rate is classified into five

Table 4.3 Severity rating (example).

Catastrophic	Any failure that could result in deaths or injuries or prevent performance of the intended mission.
Critical	Any failure that will degrade the system beyond acceptable limits and create a safety hazard (cause death or injury if corrective action is not immediately taken).
Major	Any failure that will degrade the system beyond acceptable limits but can be adequately counteracted or controlled by alternate means.
Minor	Any failure that does not degrade the overall performance beyond acceptable limits – one of the nuisance variety.

	Severity group			
Failure rate	Minor	Major	Critical	Catastrophic
Frequent				
Probable				
Occasional	(x)			
Remote		(x)		
Very unlikely	(x)		(x)	

Figure 4.5 Risk matrix of the different failure modes.

classes, and the severity is classified into four classes. The most critical failure modes are represented by (x) in the upper right corner of the risk matrix, whereas the least critical failure modes get (x) in the lower left corner of the risk matrix. In practical analyses, (x) is replaced by an abbreviated indicator for the actual failure mode.

Risk Priority Number

In some application areas, for example in the automobile industry, it is common to present the "risk" related to a failure mode as a *risk priority number* (RPN). The RPN is calculated on the as the product of the severity (S), occurrence (O), and detection (D) ratings.

$$\text{RPN} = \text{S} \times \text{O} \times \text{D}. \tag{4.1}$$

The ratings are given as follows:

Severity (S). The severity rating is a numerical value, subjectively chosen as an integer between 1 and 10 that assesses how severe the customer perceives the effect of the failure.

Occurrence rate (O). The occurrence rating is a numerical value, subjectively chosen as an integer between 1 and 10 that estimates the probability that the failure mode will occur during the lifetime of the item.

Detection (D). The detection rating is a numerical value, subjectively chosen as an integer between 1 and 10 that assesses the effectiveness of the controls to prevent or detect the failure before the failure reaches the customer.

The RPN, as such, does not have any specific meaning, but the RPN (between 1 and 1000) may be used to rank the concerns in the design. In many applications, however, the severity should have higher priority than the RPN. RPNs are used only to prioritize potential design weaknesses for consideration of possible design actions to reduce criticality and/or to make the design less sensitive to manufacturing variation.

4.2.4 Applications

Many industries require FMECA to be integrated in the design process of technical systems and that FMECA worksheets be part of the system documentation. This is, for example a common practice for suppliers to the defense, the aerospace, and the automobile industry. The same requirements are becoming more and more usual within the offshore oil and gas industry.

FMECA gives the highest value when carried out during the design phase of a system. The main objective of the analysis is to reveal weaknesses and potential failures at an early stage, to enable the designer to incorporate corrections and barriers in the design. The results from FMECA may also be useful during modifications of the system and for maintenance planning. Designers are trained to think in terms of functions; how to design the system to meet specified functional requirements. Through FMECA, designers are also "forced" to consider potential failures. By early awareness of potential failures, many failures may be designed-out of the system.

Many industries are introducing a reliability-centered maintenance (RCM) program for maintenance planning. FMECA is one of the basic tools of RCM and is further discussed in Chapter 9.

Because all failure modes, failure mechanisms, and symptoms are documented in FMECA, this provides valuable information as a basis for fault diagnostic procedures and for a repairman's checklists. FMECA is very effective when applied to a system where system failures most likely are the results of single component failures. During the analysis, each failure is considered individually as an independent occurrence with no relation to other failures in the system. FMECA is not suitable for analysis of systems with a fair degree of redundancy. For such systems, FTA is a much better alternative. An introduction to FTA is given in Section 4.3. In addition, FMECA is not well suited for analyzing systems where common cause failures are considered to be a significant problem. Common cause failures are discussed in Chapter 8.

A limitation of FMECA is further the inadequate attention generally given to human errors. This is mainly due to the concentration on hardware failures.

Perhaps the worst drawback is that all component failures are examined and documented, including those that do not have any significant consequences. For large systems, especially systems with a high degree of redundancy, the amount of unnecessary documentation work is a major disadvantage.

4.3 Fault Tree Analysis

FTA was introduced in 1962 at Bell Telephone Laboratories (see Section 1.10). Today, FTA is one of the most commonly used techniques for risk and reliability studies. In particular, FTA has been used with success to analyze safety systems in nuclear power stations, such as in the Reactor Safety Study (NUREG-75/014, 1975).

A fault tree is a logic diagram that displays the relationships between a *potential system fault* and the causes of this fault. In risk analysis, the system fault is often a potential accident. The causes may be environmental conditions, human errors, normal events (events that are expected to occur during the life span of the system), and specific component failures. Observe that the potential system fault may, or may not, occur sometime in the future.

FTA may be qualitative, quantitative, or both, depending on the objectives of the analysis. Possible results from the analysis may, for example be

- A listing of the possible combinations of environmental factors, human errors, normal events, and component faults that may result in the system fault.
- The probability that the system fault will occur at a specified time or during a specified time interval.

Only qualitative FTA is covered in this chapter. Quantitative FTA is discussed in Chapter 6. FTA is thoroughly described in standards and guidelines (e.g. see IEC 61025 2006; NUREG-0492 1981; NASA 2002).

4.3.1 Fault Tree Symbols and Elements

FTA is a *deductive* method, based on a top-down approach starting with a specified system fault. The system fault is called the TOP *event* of the fault tree. The analysis is started by assuming that the potential system fault has occurred (i.e. exists). The immediate causal events $A_1, A_2, \dots$ that, either alone or in combination, lead to the TOP event are identified and connected to the TOP event through a *logic gate*. Next, we identify all potential causal events $A_{i,1}, A_{i,2}, \dots$ that may lead to event A_i for $i = 1, 2, \dots$. These events are connected to event A_i through a logic gate. This procedure is continued deductively (i.e. backwards in the causal chain) until we reach a suitable level of detail. The events on the lowest level are called the *basic events* of the fault tree. Basic events may include component faults, human errors, environmental conditions, and normal events. A simple fault tree is shown in Figure 4.6. The main symbols used in the fault tree are shown and explained in Table 4.4.

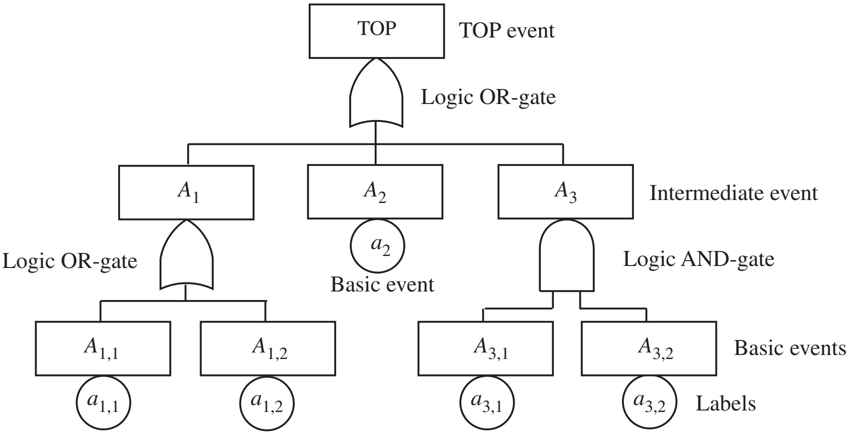


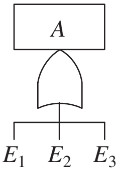
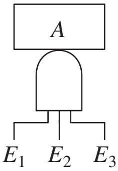
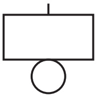
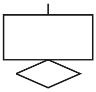
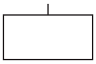
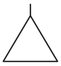
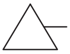
Figure 4.6 A simple fault tree.

The fault tree in Figure 4.6 shows that the TOP event occurs when one of the events A_1, A_2, or A_3 occurs. These three events are connected to the TOP event by a logic OR-gate. We may also read this as "TOP event occurs if event A_1 occurs, OR event A_2 occurs, OR event A_3 occurs." Event A_1 and event A_3 are called intermediate events because they are developed further by logic gates. Event A_2 is a basic event. The symbol in the circle is a label that uniquely identifies the basic event in the fault tree. Event A_1 is connected to its causal events $A_{1,1}$ and $A_{1,2}$ by an OR-gate and we say that "event A_1 occurs if event $A_{1,1}$ OR event $A_{1,2}$ occurs." Event A_3 is connected to its causes, event $A_{3,1}$ and event $A_{3,2}$ by an AND-gate and we say that "event A_3 occurs if event $A_{3,1}$ AND event $A_{3,2}$ occur at the same time."

Remark 4.2 (Terminology)
Observe that the method is called *fault* tree analysis and not *failure* tree analysis and recall from Chapter 3 that fault is a state, whereas failure is an event. Also observe that the fault tree construction is started by a potential (i.e. future) system failure that we imagine has occurred. This means that we start with a system fault and we ask "what could the causes be for this state to exist?" The fault state exists, so the causes are also states, even though the term "event" is used to describe them (TOP event, intermediate event, and basic event) □

FTA is a *binary* analysis. All events are assumed either to occur, or not to occur; there are no intermediate states. In the basic version, the fault tree is static and cannot accommodate any dynamic effects.

Table 4.4 Fault tree symbols.

	Symbol	Description
Logic gates	OR-gate 	The OR-gate indicates that the output event A occurs if any of the input events E_i occur
	AND-gate 	The AND-gate indicates that the output event A occurs only when all the input events E_i occur at the same time
Input events	Basic event 	The Basic event represents a basic equipment failure that requires no further development of failure causes
	Undeveloped event 	The undeveloped event represents an event that is not examined further because information is unavailable or because its consequence is insignificant
Description	Comment rectangle 	The Comment rectangle is for supplementary information
Transfer symbols	Transfer-out 	The Transfer-out symbol indicates that the fault tree is developed further at the occurrence of the corresponding Transfer-in symbol
	Transfer-in 	

The graphical layouts of the fault tree symbols depend on what standard we follow. Table 4.4 shows the most commonly used fault tree symbols together with a brief description. A number of more advanced fault tree symbols are available, but they are not covered in this book. A thorough description may be found in, e.g. see NUREG-0492 (1981) and NASA (2002).

Observe that the fault tree symbols used in IEC 61025 (2006) are different from the symbols in Table 4.4, but the meaning of the corresponding symbols are the same.

An FTA is normally carried out in five steps[1]:

(1) Definition of the problem and the boundary conditions
(2) Construction of the fault tree
(3) Identification of minimal cut and/or path sets
(4) Qualitative analysis of the fault tree
(5) Quantitative analysis of the fault tree

Steps 1–4 are covered in this section and step 5 is discussed in Chapter 6.

4.3.2 Definition of the Problem and the Boundary Conditions

The first activity of FTA has two substeps:

- Definition of the TOP event to be analyzed.
- Definition of the boundary conditions for the analysis (see also Chapter 2).

It is important that the TOP event is given a clear and unambiguous definition. If not, the analysis is often of limited value. As an example, the event description "system breakdown" is far too general and vague. The description of the TOP event should always give answer to the questions *what*, *where*, and *when*:

What:	Describes the potential system failure that is to be studied, together with a clear system failure mode description.
Where:	Describes where the system failure mode may occur.
When:	Describes when the system failure occurs (e.g. during normal operation).

To get a consistent analysis, it is important that the boundary conditions for the analysis are carefully defined. General boundary conditions were discussed in Chapter 2. Specific boundary conditions for the fault tree construction include the following:

The initial conditions. What is the operational state of the system when the TOP event is occurring? Is the system running on full/reduced capacity? Which valves are open/closed, which pumps are running, and so on?

Boundary conditions with respect to external stresses. What type of external stresses should be included in the analysis? By external stresses, we mean stresses from war, sabotage, earthquake, lightning, and so on.

[1] The procedure described below is influenced by CCPS (2008).

The level of resolution. How far down in detail should we go to identify potential causes for a failed state? Should we, for example, be satisfied when we have identified the reason to be "valve fail to close," or should we break it further down to failures in the valve housing, valve stem, actuator, and so forth. When determining the preferred level of resolution, we should remember that the detailedness in the fault tree should be comparable to the detailedness of the information available.

4.3.3 Constructing the Fault Tree

The fault tree construction always starts with the TOP event. Thereafter, all fault events that are the immediate, necessary, and sufficient causes that result in the TOP event are identified. These causes are connected to the TOP event via a logic gate. It is important that the first level of causes under the TOP event is put up in a structured way. This first level is often referred to as the TOP *structure* of the fault tree. The TOP structure causes are often taken to be failures of the prime modules of the system, or of the prime functions of the system. We then proceed, level by level, until all fault events have been developed to the prescribed level of resolution. The analysis is in other words deductive and is carried out by repeatedly asking "What are the causes of this event?"

Rules for Fault Tree Construction

Let *fault event* denote any event in the fault tree, whether it is a basic event or an event higher up in the tree.

Describe the fault events. Each basic event should be carefully described (what, where, when) in a "comment rectangle."

Evaluate the fault events. The fault events may be of different types, such as technical failures, human errors, or environmental stresses. Each event should be carefully evaluated. As explained in Section 3.6.3, technical failures may be divided into groups, such as primary failures and secondary failures. Primary failures of components are usually classified as basic events, whereas secondary failures are classified as intermediate events that require a further investigation to identify the prime reasons.

When evaluating a fault event, we ask the question, "Can this fault be a primary failure?" If the answer is yes, we classify the fault event as a "normal" basic event. If the answer is no, we classify the fault event as either an intermediate event that has to be further developed, or as a "secondary" basic event. The "secondary" basic event is often called an *undeveloped* event and represents a fault event that is not examined further because information is unavailable or because its consequence is insignificant.

Complete the gates. All inputs to a specific gate should be completely defined and described before proceeding to the next gate. The fault tree should be completed in levels, and each level should be completed before beginning the next level.

Example 4.1 (Fire detector system)

Consider a simplified version of a fire detector system located in a production room. (Observe that this system is not a fully realistic fire detector system.)

The fire detector system is divided into two parts, heat detection and smoke detection. In addition, there is an alarm button that can be operated manually. The fire detector system can be described schematically, as shown in Figures 4.7 and 4.8.

Heat detection. In the production room, there is a closed, pneumatic pipe circuit with four identical fuse plugs, FP1, FP2, FP3, and FP4. These plugs let air out of the circuit if they are exposed to temperatures higher than 72 °C. The pneumatic system has a pressure of three bars and is connected to a pressure switch (pressostat) PS. If one or more of the plugs are activated, the switch will be activated and give an electrical signal to the start relay for the alarm and shutdown system. In order to have an electrical signal, the DC-source, DC, must be intact.

Smoke detection. The smoke detection system consists of three optical smoke detectors, SD1, SD2, and SD3; all are independent and have their own batteries. These detectors are very sensitive and can give warning of fire at an early stage. In order to avoid false alarms, the three smoke detectors are connected via a logical 2oo3:G voting unit (VU). This means that at least two detectors must give fire signal before the fire alarm is activated. If at least two of the three detectors are activated, the 2oo3:G voting unit will give an electric signal to the start relay (SR), for the alarm and shutdown system. Again, the DC voltage source, DC, must be intact to obtain an electrical signal.

Manual activation. Together with the pneumatic pipe circuit with the four fuse plugs, there is a manual switch MS that can be turned to relieve the pressure in the pipe circuit. If the operator, OP, who should be continually present, notices a fire, she can activate this switch. When the switch is activated, the pressure in

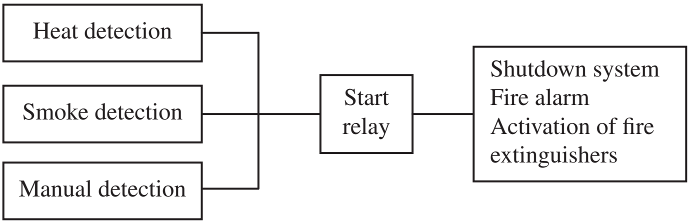


Figure 4.7 System overview of fire detector system.

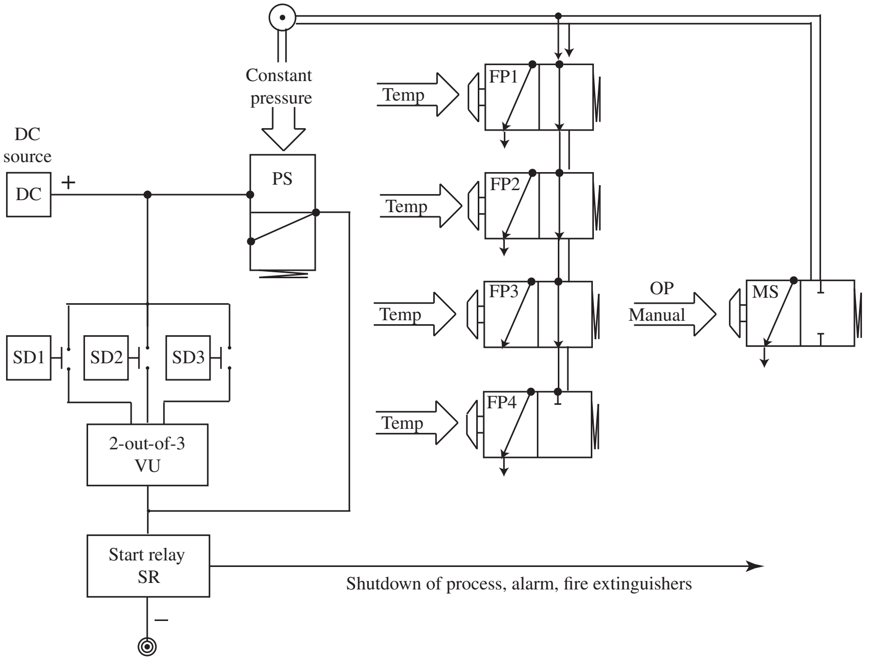


Figure 4.8 Schematic layout of the fire detector system.

the pipe circuit is relieved, and the pressure switch (PS), is activated and gives an electric signal to the start relay, SR. Again, the DC source must be intact.

The start relay. When the start relay SR receives an electrical signal from the detection systems, it is activated and gives a signal to shut down the process and to activate the alarm and the fire extinguishers.

Assume now that a fire starts. The fire detector system should detect and give warning about the fire. Let the TOP event be *“No signals from the start relay SR when a fire condition is present.”* A possible fault tree for this TOP event is shown in Figure 4.9. □

Remark 4.3 (The fault tree is not unique)
Observe that a fault tree does not show the causes of *all* system faults of the system. It only illustrates the causes of a specified fault, the TOP event. The fault tree is usually dependent on the analyst. Two different analysts will, in most cases, construct slightly different fault trees. □

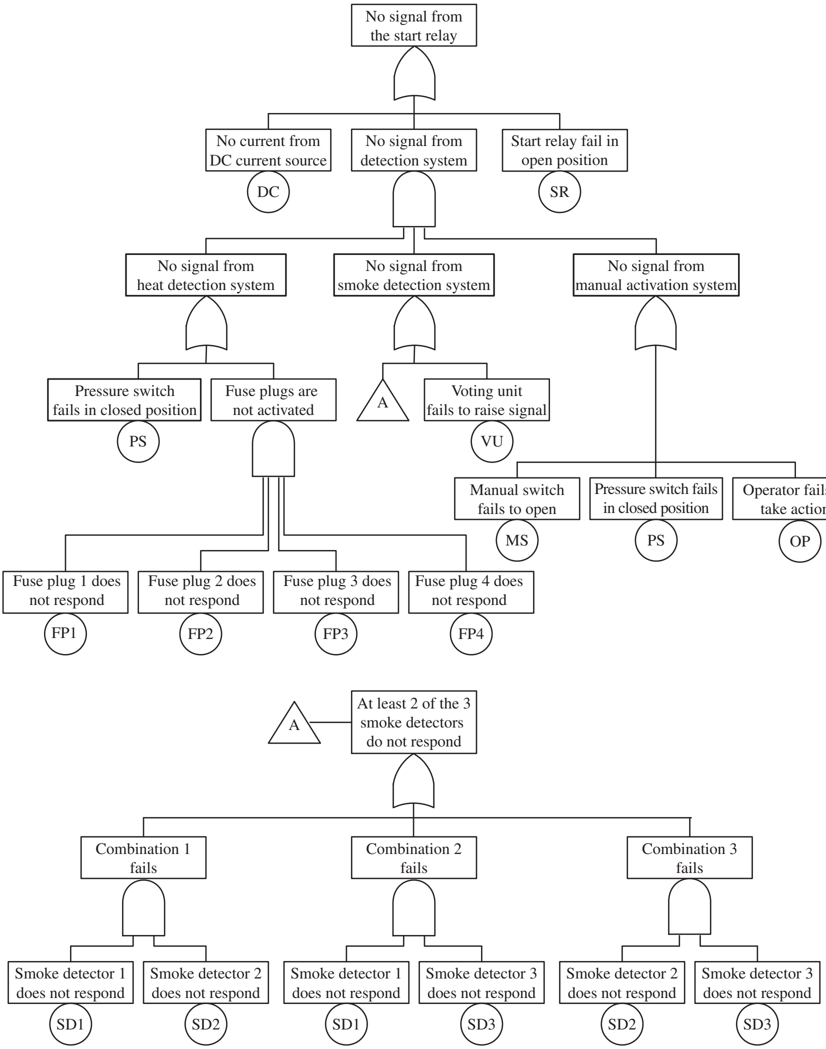

No signal from the start relay
No current from DC current source
DC
No signal from detection system
Start relay fail in open position
SR
No signal from heat detection system
No signal from smoke detection system
No signal from manual activation system
Pressure switch fails in closed position
PS
Fuse plugs are not activated
A
Voting unit fails to raise signal
VU
Manual switch fails to open
MS
Pressure switch fails in closed position
PS
Operator fails to take action
OP
Fuse plug 1 does not respond
FP1
Fuse plug 2 does not respond
FP2
Fuse plug 3 does not respond
FP3
Fuse plug 4 does not respond
FP4
A
At least 2 of the 3 smoke detectors do not respond
Combination 1 fails
Combination 2 fails
Combination 3 fails
Smoke detector 1 does not respond
SD1
Smoke detector 2 does not respond
SD2
Smoke detector 1 does not respond
SD1
Smoke detector 3 does not respond
SD3
Smoke detector 2 does not respond
SD2
Smoke detector 3 does not respond
SD3

Figure 4.9 Fault tree for the fire detector system in Example 4.1.

4.3.4 Identification of Minimal Cut and Path Sets

A fault tree provides valuable information about possible combinations of fault events that will result in the TOP event. Such a combination of fault events is called a cut set. In the fault tree terminology, a cut set is defined as follows:

Definition 4.1 (Minimal cut set in fault tree)
A cut set in a fault tree is a set of basic events whose occurrence (at the same time) ensures that the TOP event occurs. A cut set is said to be *minimal* if the set cannot be reduced without losing its status as a cut set. □

The number of different basic events in a minimal cut set is called the *order* of the cut set. For small and simple fault trees, it is feasible to identify the minimal sets by inspection without any formal procedure/algorithm. For large or complicated fault trees, we need an efficient algorithm.

4.3.5 MOCUS

MOCUS (method for obtaining cut sets) is an algorithm that can be used to find the minimal cut sets of a fault tree. The algorithm is best explained by an example. Consider the fault tree in Figure 4.10, where the gates are numbered from $G0$ to $G6$. The example fault tree is adapted from Barlow and Lambert (1975).

The algorithm starts at the $G0$ gate representing the TOP event. If this is an OR-gate, each input to the gate is written in separate rows. (The inputs may be new gates). Similarly, if the $G0$ gate is an AND-gate, the inputs to the gate are written in separate columns.

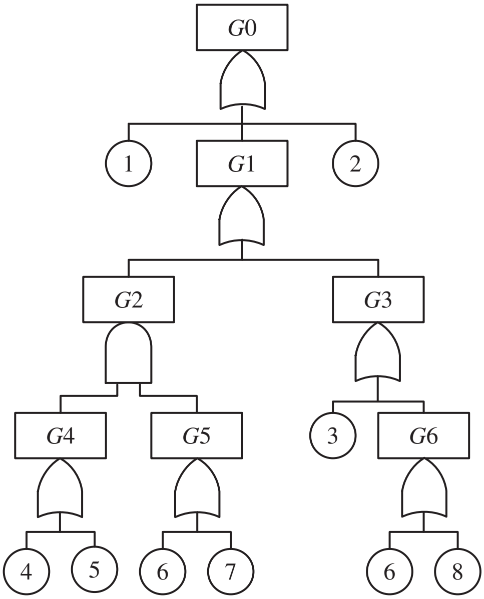


Figure 4.10 Example of a fault tree.

In our example, *G*0 is an OR-gate, hence, we start with

1
*G*1
2

Because each of the three inputs, 1, *G*1, and 2 will cause the TOP event to occur, each of them will constitute a cut set.

The idea is to successively replace each gate with its inputs (basic events and new gates) until one has gone through the whole fault tree and is left with just the basic events. When this procedure is completed, the rows in the established matrix represent the cut sets in the fault tree.

Because *G*1 is an OR-gate:

1
*G*2
*G*3
2

Because *G*2 is an AND-gate:

1
*G*4,*G*5
*G*3
2

Because *G*3 is an OR-gate:

1
*G*4,*G*5
3
*G*6
2

Because *G*4 is an OR-gate:

1
4,*G*5
5,*G*5
3
*G*6
2

Because *G*5 is an OR-gate:

1
4,6
4,7
5,6
5,7
3
*G*6
2

Because *G*6 is an OR-gate:

1
4,6
4,7
5,6
5,7
3
6
8
2

We are then left with the following nine cut sets:

{1}
{2}
{3}
{6}
{8}
{4,6}
{4,7}
{5,6}
{5,7}

Because {6} is a cut set, {4,6} and {5,6} are not minimal. If we leave these out, we are left with the following list of minimal cut sets:

{1}, {2}, {3}, {6}, {8}, {4,7}, {5,7}

In other words, five minimal cut sets of order 1 and two minimal cut sets of order 2. The reason that the algorithm in this case leads to nonminimal cut sets is that basic event 6 occurs several places in the fault tree.

In some situations, it may be of interest to identify the possible combinations of components which by functioning secure that the system is functioning. Such a combination of components (basic events) is called a *path set*. In the fault tree terminology, a path set is defined as follows:

Definition 4.2 (Minimal path set in fault tree)
A path set in a fault tree is a set of basic events whose nonoccurrence (at the same time) ensures that the TOP event does not occur. A path set is said to be *minimal* if the set cannot be reduced without losing its status as a path set. □

The number of different basic events in a minimal path set is called the *order* of the path set. To find the minimal path sets in the fault tree, we may start with the so-called dual fault tree. This can be obtained by replacing all the AND-gates in the original fault tree with OR-gates, and vice versa. In addition, we let the events in the dual fault tree be complements of the corresponding events in the original fault tree. The same procedure, as described above applied to the dual fault tree, will now yield the minimal path sets.

For relatively "simple" fault trees, one can apply the MOCUS algorithm by hand. More complicated fault trees require the use of a computer. A number of computer programs for minimal cut (path) set identification are available. Some of these are based on MOCUS, but faster algorithms have been developed.

4.3.6 Qualitative Evaluation of the Fault Tree

A fault tree may be evaluated qualitatively[2] based on the minimal cut sets. The criticality of a cut set obviously depends on the number of basic events in the cut

[2] This section is influenced by CCPS (2008).

Table 4.5 Criticality ranking of minimal cut sets of order 2.

Rank	Basic event 1 (type)	Basic event 2 (type)
1	Human error	Human error
2	Human error	Active equipment failure
3	Human error	Passive equipment failure
4	Active equipment failure	Active equipment failure
5	Active equipment failure	Passive equipment failure
6	Passive equipment failure	Passive equipment failure

set (i.e. the order of the cut set). A cut set of order one is usually more critical than a cut set of order two, or more. When we have a cut set of order one, the TOP event will occur as soon as the corresponding basic event occurs. When a cut set has two basic events, both of these have to occur at the same time to cause the TOP event to occur.

Another important factor is the type of basic events of a minimal cut set. We may rank the criticality of the various cut sets according to the following ranking of basic events:

(1) Human error
(2) Active equipment failure
(3) Passive equipment failure

This ranking is based on the assumption that human errors occur more frequently than active equipment failures and that active equipment is more prone to failure than passive equipment (e.g. an active or running pump is more exposed to failures than a passive standby pump). Based on this ranking, we get the ranking in Table 4.5 of the criticality of minimal cut sets of order 2. (Rank 1 is the most critical one.)

Example 4.2 (Offshore separator)

Consider a part of the processing section on an offshore oil and gas production installation. A mixture of oil, gas, and water coming from the various wells is collected in a wellhead manifold and led into two identical process trains. The gas, oil, and water are separated in several separators. The gas from the process trains is then collected in a compressor manifold and led to the gas export pipeline via compressors. The oil is loaded onto tankers, and the water is cleaned and re-injected into the reservoir. Figure 4.11 shows a simplified sketch of a section of one of the process trains. The mixture of oil, gas, and water from the wellhead manifold is led into the separator, where the gas is (partly) separated from the fluids. The process is controlled by a *process control system* that is not shown in the figure. If the process control system fails, a separate *process safety system* should prevent a major accident. The rest of this example is limited to the process safety system. The process safety system has three protection layers:

(1) On the inlet pipeline, there are installed two process shutdown (PSD) valves, PSD_1, and PSD_2 in series. The valves are fail-safe close and are held open by hydraulic (or pneumatic) pressure. When the hydraulic (or pneumatic) pressure is bled off, the valves will close by the force of a precharged actuator. The system supplying hydraulic (or pneumatic) pressure to the valve actuators is not shown in Figure 4.11.

Two pressure switches, PS_1 and PS_2 are installed in the separator. If the pressure in the separator increases above a set value, the pressure switches should send a signal to a logic unit (LU). If the LU receives at least one signal from the pressure switches, it will send a signal to the PSD valves to close.

(2) Two pressure safety valves (PSV) are installed to relieve the pressure in the separator in case the pressure increases beyond a specified high pressure. The PSV valves, PSV_1, and PSV_2 are equipped with a spring-loaded actuator that may be adjusted to a preset pressure.

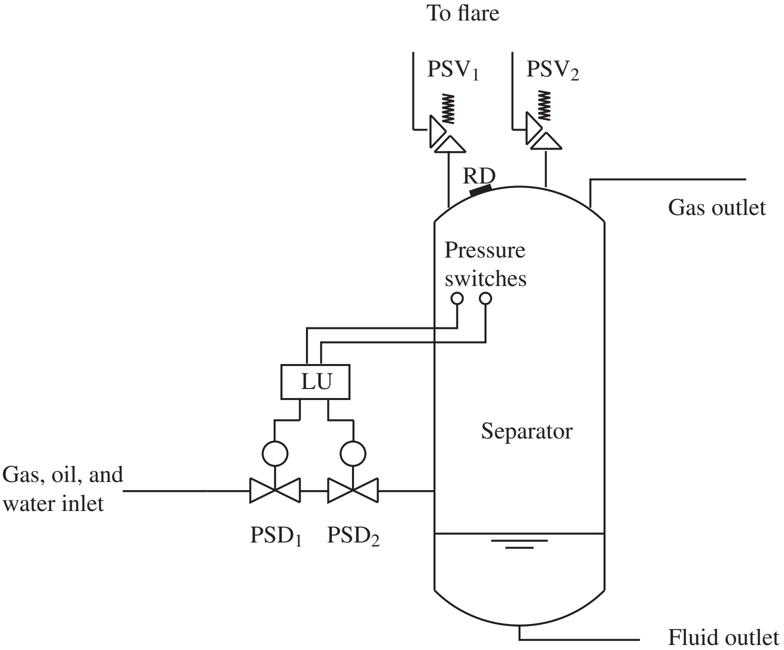


Figure 4.11 Sketch of a first stage gas separator.

(3) A rupture disc (RD) is installed on top of the separator as a last safety barrier. If the other safety systems fail, the rupture disc will open and prevent the separator from rupturing or exploding. If the rupture disc opens, the gas will blow out from the top of the separator and maybe into a blowdown system.

The reliability of the process safety system may be analyzed by different approaches. We will here illustrate how a fault tree can be performed.

Fault Tree Analysis. The most critical situation will arise if the gas outlet line A is suddenly blocked. The pressure in the separator will then rapidly increase and will very soon reach a critical overpressure, if the process safety system does not function properly. A relevant TOP event is therefore "Critical overpressure in the first stage separator." We assume that the critical situation occurs during normal production and that the fluid level in the separator is normal when the event occurs. We may therefore disregard the fluid outlet line from the FTA. A possible fault tree for this TOP event is shown in Figure 4.12. Chapter 6 deals with how to enter failure rates and other reliability parameters into the fault tree, and how to calculate the probability $Q_0(t)$ of the TOP event when gas outlet is suddenly blocked.

Before constructing the fault tree in Figure 4.12, we have made a number of assumptions. The assumptions should be recorded in a separate file and integrated in the report from the analysis. The lowest level of resolution in the fault tree in Figure 4.12 is a failure mode of a technical item. Some of these items are rather complicated, and it might be of interest to break them down into subitems and attribute failures to these. The valves may, for example be broken down into valve body and actuator. These subitems may again be broken down to sub-subitems, and so on. The failure of the pressure switches to give signal may be split into two parts, individual failures and common cause failures that cause both pressure switches to fail at the same time. A pressure switch may fail due to an inherent component failure, or due to miscalibration by the maintenance crew. How far we should proceed depends on the objective of the analysis. Anyway, the assumptions made should be recorded. □

4.3.7 Dynamic Fault Trees

A *dynamic fault tree* (DFT) extends the traditional fault tree by taking certain dynamic effects into account. A typical dynamic effect occurs when the output event of a gate depends not only on the logical combination of its input events but also on the order in which all the input events occur. To cater for relevant dynamic effects, several new gates have to be introduced in addition to the AND and OR gates. An example of such an effect occurs when a specific event (called a trigger) occurs and causes otherwise independent events to occur at (almost) the same time. The trigger event may, for example be a control system failure or a power failure.

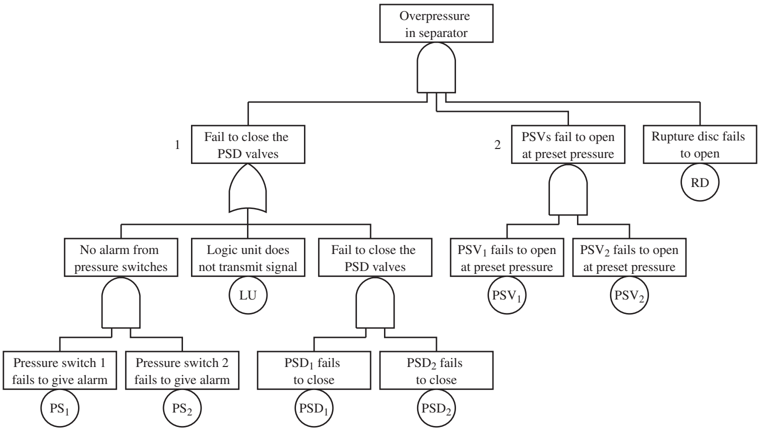

Overpressure in separator
1
Fail to close the PSD valves
2
PSVs fail to open at preset pressure
Rupture disc fails to open
RD
No alarm from pressure switches
Logic unit does not transmit signal
LU
Fail to close the PSD valves
PSV_1 fails to open at preset pressure
PSV_2 fails to open at preset pressure
PSV_1
PSV_2
Pressure switch 1 fails to give alarm
Pressure switch 2 fails to give alarm
PSD_1 fails to close
PSD_2 fails to close
PS_1
PS_2
PSD_1
PSD_2

Figure 4.12 Fault tree for the first stage separator in Example 4.2.

Analysis of DFTs is rather complicated and goes beyond the scope of this book. Readers who want to pursue this topic may start by reading Chapter 8 of NASA (2002). Quantitative DFT analysis is usually accomplished either by converting the DFT to a Markov model (see Chapter 11) or by Monte Carlo simulation (see Section 6.10.1). For further information, see Dugan (2000) and Xu et al. (2006). DFT is not discussed further in this book.

4.4 Event Tree Analysis

In many accident scenarios, the initiating event, such as a ruptured pipeline, may have a wide spectrum of possible outcomes, ranging from no consequences to a disaster. In most well-designed systems, a number of safety functions, or barriers, are provided to stop, or mitigate, the consequences of potential initiating events. The safety functions may comprise technical equipment, human interventions, emergency procedures, and combinations of these. Examples of technical safety functions are fire and gas detection systems, emergency shutdown (ESD) systems, automatic train stop systems, fire-fighting systems, fire walls, and evacuation systems. The consequences of the initiating event are determined by how the accident progression is affected by subsequent failure or operation of these safety functions, by human errors made in responding to the initiating event, and by various factors such as weather conditions and time of the day.

The accident progression is best analyzed by an inductive method. The most common method is *event tree analysis* (ETA). An event tree is a logic tree diagram that starts from an initiating event and provides a systematic coverage of the time sequence of event propagation to its potential outcomes or consequences. In the development of the event tree, we follow each of the possible sequences of events that result from assuming failure or success of the safety functions affected as the accident propagates. Each event in the tree is conditional on the occurrence of the previous events in the event chain. The outcomes of each event are most often assumed to be binary (*true* or *false* – *yes* or *no*), but may also include multiple outcomes (e.g. *yes*, *partly*, and *no*).

ETA is a natural part of most risk analyses but may also be used as a design tool to demonstrate the effectiveness of protective systems in a plant. Event tree analyses are also used for human reliability assessment, for example as part of the technique for human error-rate prediction (THERP) technique (NUREG/CR-1278).

The ETA may be qualitative, quantitative, or both, depending on the objectives of the analysis. In quantitative risk assessment application, event trees may be developed independently or follow on from FTA.

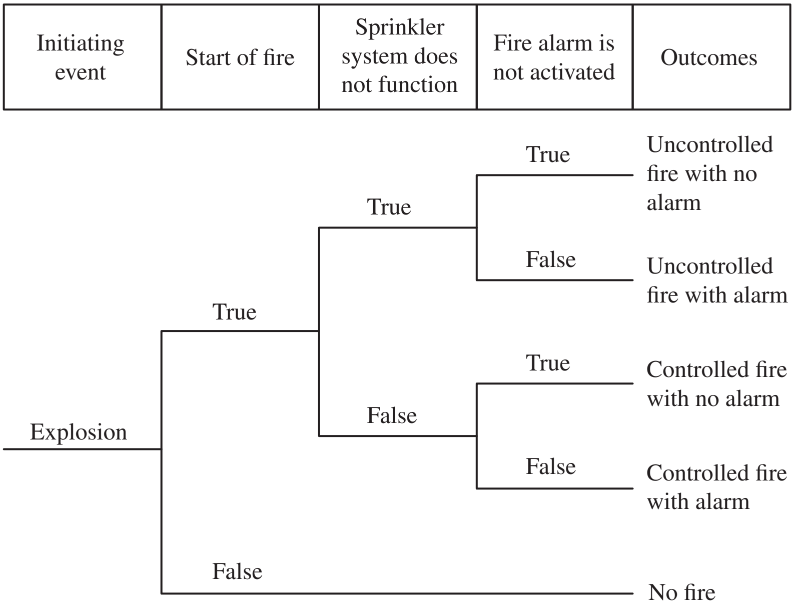

Initiating event
Start of fire
Sprinkler system does not function
Fire alarm is not activated
Outcomes
Explosion
True
False
True
False
True
False
True
False
Uncontrolled fire with no alarm
Uncontrolled fire with alarm
Controlled fire with no alarm
Controlled fire with alarm
No fire

Figure 4.13 A simple event tree for a dust explosion.

An ETA is usually carried out in six steps (CCPS 2008):

(1) Identification of a relevant initiating (hazardous) event that may give rise to unwanted consequences.
(2) Identification of the safety functions that are designed to deal with the initiating event.
(3) Construction of the event tree.
(4) Description of the resulting accident event sequences.
(5) Calculation of probabilities/frequencies for the identified consequences.
(6) Compilation and presentation of the results from the analysis.

A simple event tree for a (dust) explosion is shown in Figure 4.13. Following the initiating event explosion in Figure 4.13, fire may, or may not, break out. A sprinkler system and an alarm system have been installed. These may, or may not, function. Quantitative analysis of event trees is discussed in Section 6.8.

4.4.1 Initiating Event

Selection of a relevant initiating event is very important for the analysis. The initiating event is usually defined as the first significant deviation from the normal situation that may lead to a system failure or an accident. The initiating event may be a technical failure or some human error and may have been identified by techniques such as FMECA. To be of interest for further analysis, the initiating event must give rise to a number of consequence sequences. If the initiating event gives rise to only one consequence sequence, FTA is a more suitable technique to analyze the problem.

The initiating event is often identified and anticipated as a possible critical event already in the design phase. In such cases, barriers and safety functions have usually been introduced to deal with the event.

Various analysts may define slightly different initiating events. For a safety analysis of, for example an oxidation reactor, one analyst may choose "Loss of cooling water to the reactor" as a relevant initiating event. Another analyst may, for example choose "Rupture of cooling water pipeline" as initiating event. Both of these are equally correct.

4.4.2 Safety Functions

Safety functions (e.g. barriers, safety systems, procedures, and operator actions) that respond to the initiating event may be thought of as the system's defense against the occurrence of the initiating event. Safety functions may be classified in the following groups (CCPS 2008):

- Safety systems that automatically respond to the initiating event (e.g. automatic shutdown systems)
- Alarms that alert the operator(s) when the initiating event occurs (e.g. fire alarm systems)
- Operator procedures following an alarm
- Barriers or containment methods that are intended to limit the effects of the initiating event

The analyst must identify all barriers and safety functions that have impact on the consequences of an initiating event, in the sequence they are assumed to be activated.

The possible event chains, and sometimes also the safety functions, may be affected by various hazard contributing factors (events or states), such as

- Ignition or no ignition of a gas release
- Explosion or no explosion
- Time of the day
- Wind direction toward community or not
- Meteorological conditions
- Liquid/gas release contained or not

4.4.3 Event Tree Construction

The event tree displays the chronological development of event chains, starting with the initiating event and proceeding through successes and/or failures of the safety functions that respond to the initiating event. The consequences are clearly defined events that result from the initiating event.

The diagram is usually drawn from left to right, starting from the initiating event. Each safety function or hazard contributing factor is called a *node* in the event tree and is formulated either as an event description, or as a question, usually with two possible outcomes (*true* or *false* – *yes* or *no*). At each node, the tree splits into two branches: the upper branch signifying that the event description in the box above that node is *true*, and a lower branch, signifying that it is *false*. If we formulate the description of each node such that the worst outcome will always be on the upper branch, the consequences will usually be ranked in a descending order, with the worst consequence highest up in the list.

The outputs from one event lead to other events. The development is continued to the resulting consequences. If the diagram is too big to be drawn on a single page, it is possible to isolate branches and draw them on different pages. The different pages may be linked together by transfer symbols. Observe that for a sequence of n events, there will be 2^n branches of the tree. The number may in many cases be reduced by eliminating impossible branches.

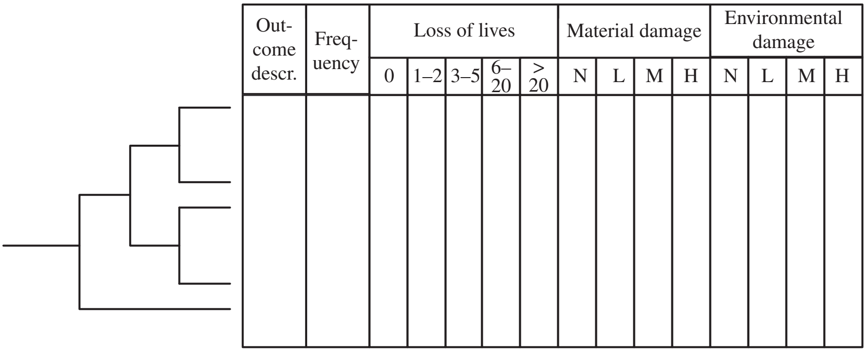
4.4.4 Description of Resulting Event Sequences

The last step in the qualitative part of the analysis is to describe the different event sequences arising from the initiating event. One or more of the sequences may represent a safe recovery and a return to normal operation or an orderly shutdown. The sequences of importance, from a safety point of view, are those that result in accidents.

The analyst must strive to describe the resulting consequences in a clear and unambiguous way. When the consequences are described, the analyst may rank them according to their criticality. The structure of the diagram, clearly showing the progression of the accident, helps the analyst in specifying where additional procedures or safety systems will be most effective in protecting against these accidents.

Sometimes, we may find it beneficial to split the end consequences (outcomes) of the ETA into various consequence categories as shown in Figure 4.14. In this example, the following categories are used:

- Loss of lives
- Material damage
- Environmental damage



Outcome descr.	Frequency	Loss of lives					Material damage				Environmental damage			
		0	1–2	3–5	6–20	> 20	N	L	M	H	N	L	M	H

Figure 4.14 Presentation of results from ETA.

Within each category, the consequences may be ranked. For the category "loss of lives," the subcategories 0, 1–2, 3–5, 6–20, and ≥21 are proposed. For the categories "material damage" and "environmental damage" the subcategories are negligible (N), low (L), medium (M), and high (H). What is meant by these categories has to be defined in each particular case. If we are unable to put the consequences into a single group, we may give a probability distribution over the subcategories. The outcome of an event chain may, for example be that nobody will be killed with probability 50%, 1–2 persons will be killed with probability 40%, and 3–5 persons will be killed with probability 10%. If we in addition are able to estimate the frequency of the outcome it is straightforward to estimate the fatal accident rate[3] (FAR) associated with the specified initiating event.

Example 4.3 (Offshore separator–event tree)

Reconsider the offshore separator in Example 4.2. The activation pressures for the three protection layers of the process safety system are shown in Figure 4.15. We get different consequences depending on whether or not the three protection systems are functioning, and the system is therefore suitable for ETA. The initiating event is "blockage of the gas outlet line." A possible event tree for this initiating event is presented in Figure 4.16. The four outcomes are seen to give very different consequences. The most critical outcome is "rupture or explosion of separator" and may lead to total loss of the installation if the gas is ignited. The probability of this outcome is, however, very low because the rupture disc is a very simple and reliable item. The second most critical outcome is "gas flowing out of rupture disc." The criticality of this outcome depends on the design of the system, but may for some installations be very critical, if the gas is ignited. The next outcome

[3] FAR is a commonly used metric for personnel risk and is defined as the expected number of fatalities per 10^8 hours of exposure.

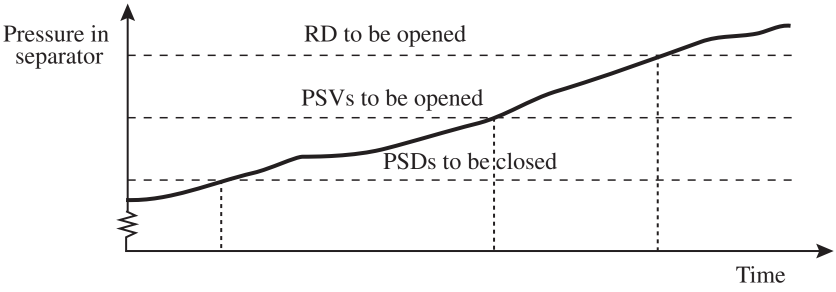


Figure 4.15 Activation pressures for the three protection layers of the process safety system.

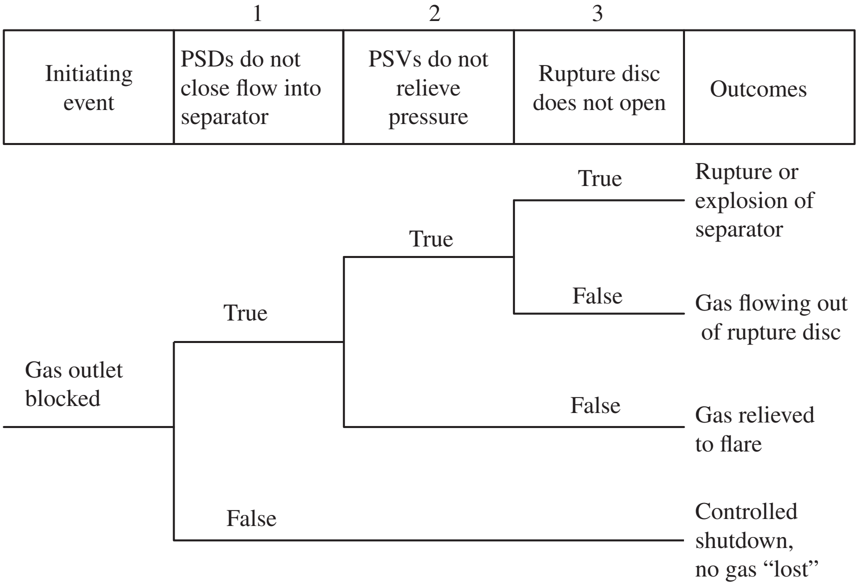

1
2
3
Initiating event
PSDs do not close flow into separator
PSVs do not relieve pressure
Rupture disc does not open
Outcomes
True
Rupture or explosion of separator
True
False
Gas flowing out of rupture disc
True
Gas outlet blocked
False
Gas relieved to flare
False
Controlled shutdown, no gas “lost”

Figure 4.16 An event tree for the initiating event “blockage of the gas outlet line.”

“gas relieved to flare” is usually a noncritical event, but this will lead to an economic loss (CO_2 tax) and production downtime. The last outcome is a controlled shutdown that will only lead to production downtime.

In this case, ETA is seen to provide more detailed results than FTA. The two analyses may be combined. The causes of failure of barrier 1 (PSDs do not close flow into separator) are found in branch 1 of the fault tree in Figure 4.12. The causes of failure of barrier 2 (PSVs do not relieve pressure) are found in branch 2 of the fault tree in Figure 4.12. □

4.5 Fault Trees versus Reliability Block Diagrams

RBDs were introduced in Chapter 2 and provide roughly the same information as fault trees. In some practical applications, we may choose whether to model the system structure by a fault tree or by an RBD. When the fault tree is limited to only OR-gates and AND-gates, both methods give the same result, and we may convert the fault tree to an RBD, and vice versa.

Remark 4.4 (Terminology)
When block i in an RBD is functioning, this means that a specific function i of the associated component is in order, for $i = 1, 2, \dots, n$. Instead of saying "block i is functioning" we will from now on say that "component i is functioning." Even if this change is not fully correct, it simplifies the presentation and it also brings our presentation in line with most other textbooks on system reliability. □

In an RBD, connection through a block means that the associated component is functioning. This again means that one, or a specified set, of failure modes of the component is not occurring. In a fault tree, we may let a basic event be the occurrence of the same failure mode – or the same specified set of failure modes – for the component. When the TOP event in the fault tree represents "system failure" and the basic events are defined as above, it is easy to see, for instance, that a series structure is equivalent to a fault tree where all the basic events are connected through an OR-gate. The TOP event occurs and the series system fails, if either component 1, OR component 2, OR component 3, OR $\cdots$ OR component n fails.

In the same way, a parallel structure may be represented as a fault tree where all the basic events are connected through an AND-gate. The TOP event occurs (i.e. the parallel structure fails), if component 1, AND component 2, AND component 3, AND $\cdots$ AND component n fail. The relationships between some simple RBDs and fault trees are shown in Figure 4.17.

Example 4.4 (Example 4.1 (cont.))
It is usually an easy task to convert a fault tree to an RBD. The RBD corresponding to the fault tree for the fire detector system in Figure 4.8 is shown in Figure 4.18. In this conversion, we start from the TOP event and replace the gates successively. OR-gates are replaced by series structures of the "components" directly beneath the gate, and AND-gates are replaced by a parallel structure of the "components" directly beneath the gate. □

From Figure 4.18, we observe that some of the components are represented in two different locations in the diagram. It should be emphasized that an RBD is not a physical layout diagram for the system. It is a logic diagram, illustrating the function of the system.

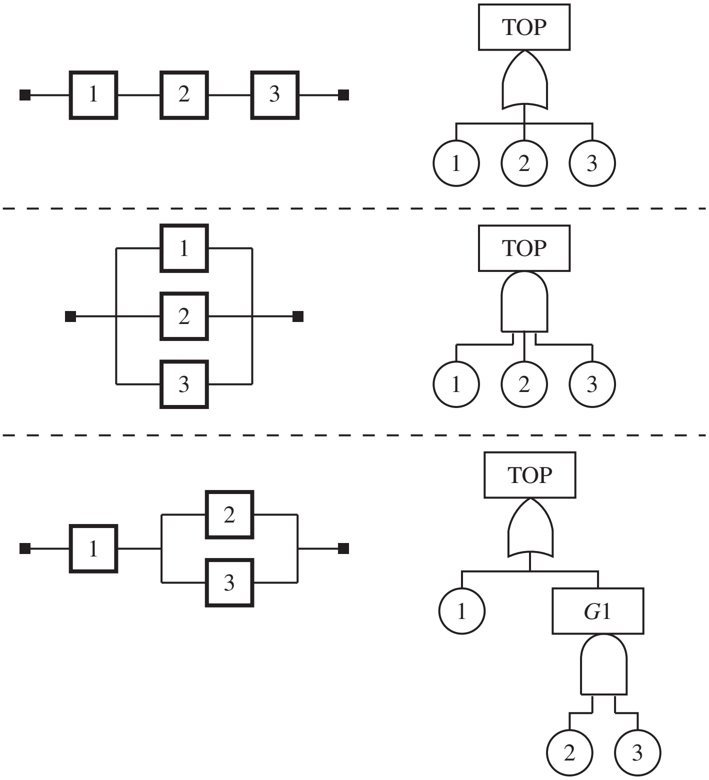


Figure 4.17 Relationship between some simple RBDs and fault trees.

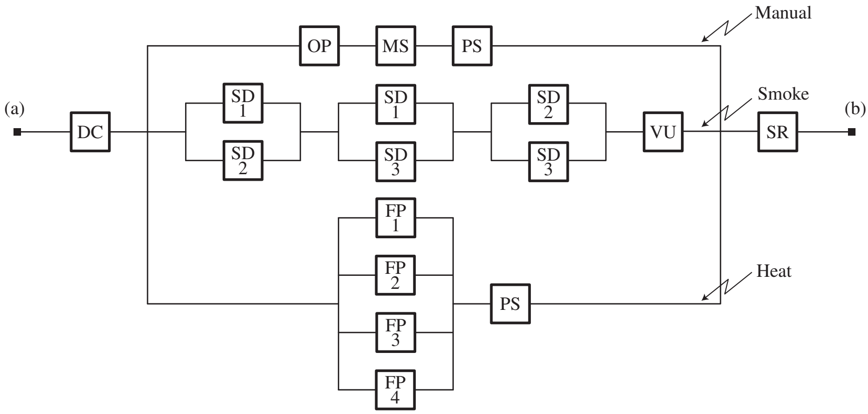


Figure 4.18 RBD for the fire detector system.

4.5.1 Recommendation

For most practical applications, we recommend to start by constructing a fault tree instead of an RBD. When constructing a fault tree, we search for potential causes of the TOP event and all intermediate events. We think in terms of *faults* and will often reveal more potential fault causes than if we think in terms of *functions*. The construction of a fault tree will give the analyst a better understanding of the potential causes of fault. If the analysis is carried out in the design phase, the analyst may rethink the design and operation of the system and take actions to eliminate potential hazards.

When we establish an RBD, we think in terms of functions and will often overlook auxiliary functions and equipment that is, or should be, installed to protect the equipment, people, and/or the environment.

For further evaluation, it is often more natural to base these on an RBD. A fault tree will therefore sometimes be converted to an RBD for qualitative and quantitative analyses.

4.6 Structure Function

Consider a structure of n components.[4] The structure is said to be of *order n*, and the components are assumed to be numbered consecutively from 1 to n[5].

Each component is assumed to have only two states, a functioning state and a failed state. The same applies to the structure. The state of component i, $i = 1, 2, \dots, n$ can then be described by a binary[6] state variable x_i, where

$$x_i = \begin{cases} 1 & \text{if component } i \text{ is functioning} \\ 0 & \text{if component } i \text{ is in a failed state} \end{cases}, \tag{4.2}$$

$\boldsymbol{x} = (x_1, x_2, \dots, x_n)$ is called the *state vector* of the structure. Furthermore, we assume that by knowing the states of all the n components, we also know whether the structure is functioning or not.

Similarly, the state of the structure can be described by a binary function

$$\phi(\boldsymbol{x}) = \phi(x_1, x_2, \dots, x_n),$$

where

$$\phi(\boldsymbol{x}) = \begin{cases} 1 & \text{if the structure is functioning} \\ 0 & \text{if the structure is in a failed state} \end{cases}, \tag{4.3}$$

[4] Remember that the term "component" is used here to denote a specified function of the component.

[5] Sections 4.6 and 4.7 are influenced by Barlow and Proschan (1975).

[6] In this context, a binary variable (function) is a variable (function) that can take only the two values, 0 or 1.

$\phi(\boldsymbol{x})$ is called the *structure function* or just the *structure*. Examples of simple structures are given in Sections 4.6.1–4.6.3.

4.6.1 Series Structure

A series structure is functioning if and only if *all* of its n components are functioning. The structure function is

$$\phi(\boldsymbol{x}) = x_1 \cdot x_2 \ \cdots \ x_n = \prod_{i=1}^{n} x_i, \tag{4.4}$$

where $\prod$ is the multiplication sign. A series structure of order n is illustrated by the RBD in Figure 2.12. Observe that the structure function of a series structure may be written as

$$\phi(\boldsymbol{x}) = \prod_{i=1}^{n} x_i = \min_{1\le i\le n} x_i.$$

4.6.2 Parallel Structure

A parallel structure is functioning if at least one of its n components is functioning. The structure function is

$$\phi(\boldsymbol{x}) = 1 - (1 - x_1)(1 - x_2) \cdots (1 - x_n) = 1 - \prod_{i=1}^{n} (1 - x_i). \tag{4.5}$$

A parallel structure of order n is illustrated by the RBD in Figure 2.13.

The expression on the right-hand side of (4.5) is often written as $\coprod_{i=1}^{n} x_i$, where $\amalg$ is read “ip.”

Hence, a parallel structure of order 2 has structure function

$$\phi(x_1, x_2) = 1 - (1 - x_1)(1 - x_2) = \coprod_{i=1}^{2} x_i.$$

The right-hand side may also be written as $x_1 \amalg x_2$, where $\amalg$ is the sign for logical OR. Observe that

$$\phi(x_1, x_2) = x_1 + x_2 - x_1 x_2. \tag{4.6}$$

Because x_1 and x_2 are binary variables, $x_1 \amalg x_2$ will be equal to the maximum of the x_i’s. Similarly,

$$\phi(\boldsymbol{x}) = \coprod_{i=1}^{n} x_i = \max_{1\le i\le n} x_i.$$

Boolean Algebra

Boolean algebra is a branch of mathematics dealing with variables that are *true* or *false*, usually denoted 1 and 0, respectively. Boolean algebra was introduced by the English mathematician George Boole (1815–1864). The basic operations are AND and OR. In the reliability literature, the AND-symbol is simply written as a product and the OR-symbol is written as $\amalg$.

$$\begin{aligned} \text{AND} &: \; x_1 \cdot x_2 = \min\{x, y\} = x_1 x_2 \\ \text{OR} &: \; x_1 \amalg x_2 = \max\{x, y\} = x_1 + x_2 - x_1 x_2. \end{aligned}$$

We also use Boolean algebra for *events* (sets in a sample space $\mathcal{S}$), where the AND-symbol is written ∩ and the OR-symbol is written ∪. The following rules apply:

$$\begin{array}{ll} A \cup B = B \cup A & A \cup A = A \\ A \cap B = B \cap A & A \cap A = A \\ A \cap (B \cup C) = (A \cap B) \cup (A \cap C) & \emptyset \cap A = \emptyset \\ A \cup (A \cap B) = A & \emptyset \cup A = A \\ \overline{A} \cap A = \emptyset & A \cup (\overline{A} \cap B = A \cup B \\ \overline{A} \cup A = \mathcal{S} & A \cap (\overline{A} \cup B) = A \cap B \\ \overline{A \cup B} = \overline{A} \cap \overline{B} & \overline{A \cap B} = \overline{A} \cup \overline{B}, \end{array}$$

where $\overline{A}$ is the negation of event A (i.e. $\overline{A} = \mathcal{S} - A$). Boolean algebra is commonly used when defining electronic systems.

4.6.3 *k*oo*n*:G Structure

A $k$oo$n$:G structure is functioning if and only if at least k of the n components are functioning (i.e. are "good"). A series structure is therefore an $n$oo$n$:G structure, and a parallel structure is a 1oo$n$:G structure.

In the rest of this chapter, all the $k$oo$n$ structures considered are $k$oo$n$:G structures. To simplify the notation, we omit the explicit reference to functioning (i.e. "good") components and simply write $k$oo$n$.

The structure function of a $k$oo$n$ structure can be written as

$$\phi(\boldsymbol{x}) = \begin{cases} 1 & \text{if} \quad \sum_{i=1}^{n} x_i \geq k \\ 0 & \text{if} \quad \sum_{i=1}^{n} x_i < k \end{cases}, \tag{4.7}$$

where $\sum$ is the summation sign. As an example, consider a 2oo3 structure, as shown in Figure 2.14. In this case, the failure of one component is tolerated, whereas two or more component failures lead to system failure.

A three-engined airplane that can stay in the air if and only if at least two of its three engines are functioning, is an example of a 2oo3 structure.

The structure function of the 2oo3 structure in Figure 2.14 may also be written as

$$\begin{aligned}
\phi(\boldsymbol{x}) &= x_1x_2 \amalg x_1x_3 \amalg x_2x_3 \\
&= 1 - (1 - x_1x_2)(1 - x_1x_3)(1 - x_2x_3) \\
&= x_1x_2 + x_1x_3 + x_2x_3 - x_1^2x_2x_3 - x_1x_2^2x_3 - x_1x_2x_3^2 + x_1^2x_2^2x_3^2 \\
&= x_1x_2 + x_1x_3 + x_2x_3 - 2x_1x_2x_3. \qquad (4.8)
\end{aligned}$$

(Observe that because x_i is a binary variable, $x_i^k = x_i$ for all i and k.)

Voted Structures in Safety Systems

The 2oo3 structures are typically used for safety systems such as gas detectors, in which case at least two of the gas detectors must signal the presence of gas to raise an alarm and/or to shut down a process. False alarm is often a problem with such systems and too many false alarms may weaken the confidence in the system. For a 2oo3 structure, at least two gas detectors have to raise simultaneous false alarms to give a system alarm and to shut down the process. Because false alarms are seldom events with a 2oo3 structure and because the 2oo3 structure has an adequate reliability, the 2oo3 structure is usually a preferred configuration of gas detectors. The 2/3 unit in Figure 2.14 is a LU (e.g. an electronic controller) that counts the number of incoming signals and sends a signal out only when there are at least two incoming signals. This is said to be a 2oo3 voting and the gas detector system is often said to be a 2oo3 *voted structure*. Voted structures are discussed in detail in Chapter 13.

4.6.4 Truth Tables

A *truth table* is a table listing all the possible values of the state variables $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$ together with the resulting value of the Boolean function $\phi(\boldsymbol{x})$. A truth table for a 2oo3 structure is shown in Table 4.6.

Observe that the 2oo3 structure is functioning (state 1) for the last four combination of states in Table 4.6 and failed (state 0) for the first four combinations.

4.7 System Structure Analysis

We now present some general properties of system structures.

Table 4.6 Truth table for a 2oo3 structure.

x_1	x_2	x_3	$\phi(x)$
0	0	0	0
0	0	1	0
0	1	0	0
1	0	0	0
0	1	1	1
1	0	1	1
1	1	0	1
1	1	1	1

4.7.1 Single Points of Failure

One of the first questions to ask for any system is: “Are there any single points of failure in the system?” A single point of failure is defined as follows:

Definition 4.3 (Single point of failure)
A component that by failing will cause the system to fail. □

When an RBD is established, it is easy to spot the single points of failure for that particular RBD (that has been established for a particular system function). For a complicated system with many system functions, it may be cumbersome to identify all single points of failure.

4.7.2 Coherent Structures

When establishing the structure of a system, it seems reasonable first to leave out all components that do not have any effect on the functioning of the system. The components we are left with are called *relevant*. The components that are not relevant are called *irrelevant*.

If component i is irrelevant, then:

$$\phi(1_i, \boldsymbol{x}) = \phi(0_i, \boldsymbol{x}) \qquad \text{for all}(\cdot_i, \boldsymbol{x}), \tag{4.9}$$

where $(1_i, \boldsymbol{x})$ represents a state vector, where the state of the ith component $= 1$, $(0_i, \boldsymbol{x})$ represents a state vector, where the state of the ith component $= 0$, and $(\cdot_i, \boldsymbol{x})$ represents a state vector where the state of the ith component $= 0$ or 1. In more

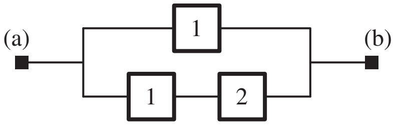


Figure 4.19 Component 2 is irrelevant.

detail,

$$(1_i, \boldsymbol{x}) = (x_1, \dots, x_{i-1}, 1, x_{i+1}, \dots, x_n)$$
$$(0_i, \boldsymbol{x}) = (x_1, \dots, x_{i-1}, 0, x_{i+1}, \dots, x_n)$$
$$(\cdot_i, \boldsymbol{x}) = (x_1, \dots, x_{i-1}, \cdot, x_{i+1}, \dots, x_n),$$

Figure 4.19 shows a system of order 2, where component 2 is irrelevant.

Remark 4.5 (Relevant and irrelevant components)
The notion "relevant/irrelevant" is sometimes misleading, as it is easy to find examples of components of great importance for a system without being relevant in the above sense. The RBD and the structure function are established for a specific system function, for example "separate gas from oil and water" in Example 4.2. To fulfill this system function, a number of components are required to function, and therefore relevant in the above sense. The shutdown function of the protection systems will be irrelevant with respect to this system function, as the production will not be influenced by the protections system's inability to shut down the process in an emergency.

When we say that a component is irrelevant, this is always with respect to a specific system function. The same component may be highly relevant with respect to another system function.

Also remember that x_i represents the state of a specific function (or, a specific subset of functions) of a component. When we say that component i is irrelevant, we in fact say that the specific function i of the physical component is irrelevant. In Example 4.2, "spurious shutdowns" of the protection system will be relevant for the system function "separate gas from oil and water," whereas the shutdown function of the same protection system will be irrelevant. □

Assume now that the system will not run worse than before if we replace a component in a failed state with one that is functioning. This is obviously the same as requiring that the structure function shall be nondecreasing in each of its arguments. Let us now define what is meant by a *coherent structure*.

Definition 4.4 (Coherent structure)
A structure where all components are relevant and the structure function is nondecreasing in each argument. □

All the systems that we have considered so far (except the one in Figure 4.19) are coherent. One might get the impression that all systems of interest must be coherent, but this is not the case. It is, for example easy to find systems where the failure of one component prevents another component from failing. This complication is discussed later.

4.7.3 General Properties of Coherent Structures

Coherent structures all have the three properties presented in this section.

Property 4.1

The structure function $\phi(\boldsymbol{x})$ of a coherent structure is a binary function that can only take the values 0 and 1. If $\phi(\mathbf{0}) = 1$, we must have that $\phi(\mathbf{0}) = \phi(\mathbf{1}) = 1$, because a coherent structure is nondecreasing in each argument. This implies that all the components in the structure are irrelevant, which contradicts the assumption that the structure is coherent. Hence,

$$\phi(\mathbf{0}) = 0. \tag{4.10}$$

Similarly $\phi(\mathbf{1}) = 0$ implies that $\phi(\mathbf{0}) = 0$, that is, that all the components are irrelevant. This contradicts the assumption of coherence. Hence,

$$\phi(\mathbf{1}) = 1. \tag{4.11}$$

The two results (4.10) and (4.11) simply say that

- If all the components in a coherent structure are functioning, then the structure is functioning.
- If all the components in a coherent structure are in a failed state, then the structure is in a failed state.

Property 4.2

Observe that $\prod_{i=1}^{n} x_i$ and $\coprod_{i=1}^{n} x_i$ are both binary and assume that $\prod_{i=1}^{n} x_i = 0$. Because we already know that $\phi(\boldsymbol{x}) \geq 0$, we know that $\prod_{i=1}^{n} x_i \leq \phi(\boldsymbol{x})$. Further assume that $\coprod_{i=1}^{n} x_i = 0$, that is, $\boldsymbol{x} = \mathbf{0}$. Then according to (4.10), $\phi(\boldsymbol{x}) = 0$, and $\phi(\boldsymbol{x}) \leq \coprod_{i=1}^{n} x_i$. Finally, assume that $\coprod_{i=1}^{n} x_i = 1$. Because we already know that $\phi(\boldsymbol{x}) \leq 1$, we conclude that

$$\prod_{i=1}^{n} x_i \leq \phi(\boldsymbol{x}) \leq \coprod_{i=1}^{n} x_i. \tag{4.12}$$

Property 4.2 says that any coherent structure is functioning at least as well as a corresponding structure where all the n components are connected in series and at most as well as a structure where all the n components are connected in parallel.

Property 4.3
Observe that $x_i \amalg y_i \geq x_i$ for all i. Because ϕ is coherent, ϕ is nondecreasing in each argument and therefore $\phi(\boldsymbol{x} \amalg \boldsymbol{y}) \geq \phi(\boldsymbol{x})$. In the same way, $\phi(\boldsymbol{x} \amalg \boldsymbol{y}) \geq \phi(\boldsymbol{y})$. Because $\phi(\boldsymbol{x})$ and $\phi(\boldsymbol{y})$ are both binary, we have that

$$\phi(\boldsymbol{x} \amalg \boldsymbol{y}) \geq \phi(\boldsymbol{x}) \amalg \phi(\boldsymbol{y}). \tag{4.13}$$

Similarly, we know that $x_i y_i \leq x_i$ for all i. Because ϕ is coherent, then $\phi(\boldsymbol{x} \cdot \boldsymbol{y}) \leq \phi(\boldsymbol{x})$ and $\phi(\boldsymbol{x} \cdot \boldsymbol{y}) \leq \phi(\boldsymbol{y})$. Because $\phi(\boldsymbol{x})$ and $\phi(\boldsymbol{y})$ are binary, then

$$\phi(\boldsymbol{x} \cdot \boldsymbol{y}) \leq \phi(\boldsymbol{x})\phi(\boldsymbol{y}). \tag{4.14}$$

Let us interpret Property 4.3 in common language. Consider the structure in Figure 4.20 with structure function $\phi(\boldsymbol{x})$. Assume that we also have an identical structure $\phi(\boldsymbol{y})$ with state vector $\boldsymbol{y}$. Figure 4.21 shows a structure with *redundancy at system level*. The structure function for this structure is $\phi(\boldsymbol{x}) \amalg \phi(\boldsymbol{y})$.

Next, consider the structure we get from Figure 4.20 when each pair x_i, y_i are connected in parallel, see Figure 4.22. This figure shows a structure with *redundancy at component level*. The structure function is $\phi(\boldsymbol{x} \amalg \boldsymbol{y})$.

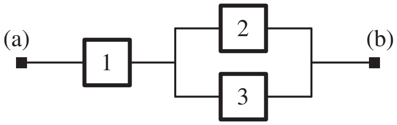


Figure 4.20 Example structure.

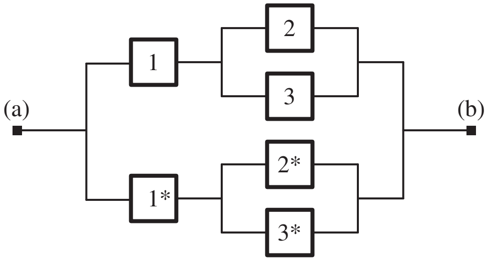


Figure 4.21 Redundancy at system level.

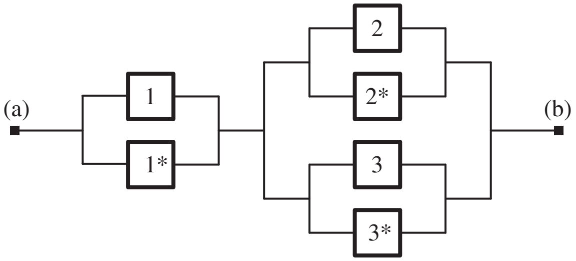


Figure 4.22 Redundancy at component level.

According to (4.13), $\phi(\boldsymbol{x} \amalg \boldsymbol{y}) \geq \phi(\boldsymbol{x}) \amalg \phi(\boldsymbol{y})$. This means that

> We obtain a "better" structure by introducing redundancy at the component level than by introducing redundancy at the system level.

This principle is well known to designers and was discussed already by Shooman (1968, pp. 281–289). The principle is not obvious when the system has two or more failure modes, for example "fail to function" and "false alarm" of a fire detection system.

4.7.4 Structures Represented by Paths and Cuts

A structure of order n consists of n components numbered from 1 to n. The set of components is denoted by

$$C = \{1, 2, \dots, n\}.$$

The concepts minimal path set and minimal cut set may be defined as follows:

Definition 4.5 (Minimal path set)
A path set P is a set of components in C that by functioning ensures that the structure is functioning. A path set is said to be *minimal* if it cannot be reduced without losing its status as a path set. □

Definition 4.6 (Minimal cut set)
A cut set K is a set of components in C that by failing causes the structure to fail. A cut set is said to be *minimal* if it cannot be reduced without losing its status as a cut set. □

We illustrate the concepts minimal path set and minimal cut set by some simple examples:

Example 4.5 Consider the RBD in Figure 4.20. The component set is $C = \{1, 2, 3\}$. The structure has the following path and cut sets.

Path sets:	Cut sets:
{1,2}∗	{1}∗
{1,3}∗	{2,3}∗
{1,2,3}	{1,2}
	{1,3}
	{1,2,3}

The minimal path sets and cut sets are marked with an ∗.

In this case, the minimal path sets are

$$P_1 = \{1, 2\} \quad \text{and} \quad P_2 = \{1, 3\},$$

whereas the minimal cut sets are

$$K_1 = \{1\} \quad \text{and} \quad K_2 = \{2, 3\}.$$

□

Example 4.6 (Bridge structure)
Consider a bridge structure such as that given by the physical network in Figure 4.23. The minimal path sets are

$$P_1 = \{1, 4\}, \quad P_2 = \{2, 5\}, \quad P_3 = \{1, 3, 5\}, \quad \text{and} \quad P_4 = \{2, 3, 4\}.$$

The minimal cut sets are

$$K_1 = \{1, 2\}, \quad K_2 = \{4, 5\}, \quad K_3 = \{1, 3, 5\}, \quad \text{and} \quad K_4 = \{2, 3, 4\}.$$

□

Example 4.7 (2oo3 structure)
Consider the 2oo3 structure in Figure 2.14. The minimal path sets are

$$P_1 = \{1, 2\}, \quad P_2 = \{1, 3\}, \quad \text{and} \quad P_3 = \{2, 3\}.$$

The minimal cut sets are

$$K_1 = \{1, 2\}, \quad K_2 = \{1, 3\}, \quad \text{and} \quad K_3 = \{2, 3\}.$$

The 2oo3 structure may therefore be represented as a series structure of its minimal cut parallel structures as shown in Figure 4.24. □

In these examples, the number of minimal cut sets coincides with the number of minimal path sets. This is usually not the case.

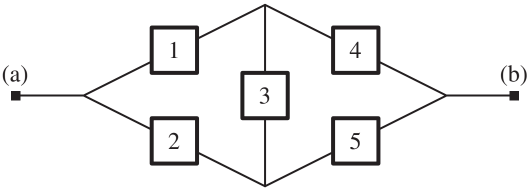


Figure 4.23 Bridge structure.

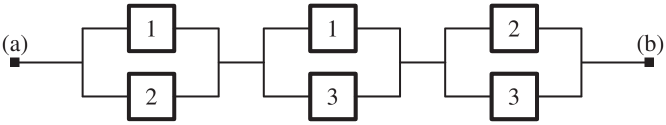


Figure 4.24 2oo3 structure represented as a series structure of the minimal cut parallel structures.

Consider the following two views:

The lazy designer's point of view. Consider a designer who wants to ensure that a structure is functioning with the least possible design effort. What the designer needs is a list of the minimal path sets from which one will be chosen for the design.

The lazy saboteur's point of view. Next, consider a saboteur who wants to bring the structure into a failed state, again with the least possible effort on his or her part. What the saboteur needs is a list of the minimal cut sets from which to choose one for the sabotage plan.

Consider an arbitrary structure with minimal path sets $P_1, P_2, \dots, P_p$ and minimal cut sets $K_1, K_2, \dots, K_k$. To the minimal path set P_j, we associate the binary function

$$\rho_j(\boldsymbol{x}) = \prod_{i \in P_j} x_i \qquad \text{for } j = 1, 2, \dots, s. \tag{4.15}$$

Observe that $\rho_j(\boldsymbol{x})$ represents the structure function of a series structure composed of the components in P_j. $\rho_j(\boldsymbol{x})$ is therefore called the jth minimal *path series structure.*

Because we know that the structure is functioning if and only if at least one of the minimal path series structures is functioning,

$$\phi(x) = \coprod_{j=1}^{p} \rho_j(\boldsymbol{x}) = 1 - \prod_{j=1}^{p} [1 - \rho_j(\boldsymbol{x})]. \tag{4.16}$$

This structure may be interpreted as a parallel structure of the minimal path series structures.

From (4.15) and (4.16), we get

$$\phi(\boldsymbol{x}) = \coprod_{j=1}^{p} \prod_{i \in P_j} x_i. \tag{4.17}$$

Example 4.8 (Example 4.7 (cont.))

In the bridge structure in Figure 4.23, the minimal path sets were $P_1 = \{1, 4\}$, $P_2 = \{2, 5\}$, $P_3 = \{1, 3, 5\}$, and $P_4 = \{2, 3, 4\}$. The corresponding minimal path series structures are

$$\rho_1(\boldsymbol{x}) = x_1 x_4$$

$$\rho_2(\boldsymbol{x}) = x_2 x_5$$

$$\rho_3(\boldsymbol{x}) = x_1 x_3 x_5$$

$$\rho_4(\boldsymbol{x}) = x_2 x_3 x_4$$

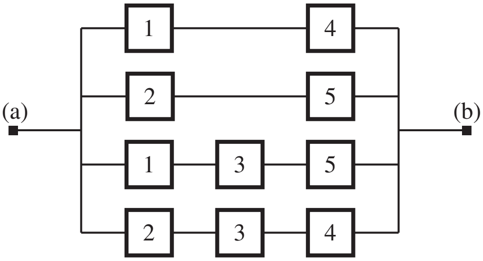


Figure 4.25 The bridge structure represented as a parallel structure of the minimal path series structures.

Accordingly, the structure function may be written as

$$\begin{aligned}\phi(\boldsymbol{x}) &= \coprod_{j=1}^{4} \rho_j(\boldsymbol{x}) = 1 - \prod_{j=1}^{4}(1 - \rho_j(\boldsymbol{x}))\\ &= 1 - (1 - \rho_1(\boldsymbol{x}))(1 - \rho_2(\boldsymbol{x}))(1 - \rho_3(\boldsymbol{x}))(1 - \rho_4(\boldsymbol{x}))\\ &= 1 - (1 - x_1x_4)(1 - x_2x_5)(1 - x_1x_3x_5)(1 - x_2x_3x_4)\\ &= x_1x_4 + x_2x_5 + x_1x_3x_5 + x_2x_3x_4 - x_1x_3x_4x_5 - x_1x_2x_3x_5\\ &\quad - x_1x_2x_3x_4 - x_2x_3x_4x_5 - x_1x_2x_4x_5 + 2x_1x_2x_3x_4x_5.\end{aligned}$$

(Remember that because x_i is a binary variable, $x_i^k = x_i$ for all i and k.)

Hence, the bridge structure can be represented by the RBD in Figure 4.25. □

Similarly, we can associate the following binary function to the minimal cut set K_j

$$\kappa_j(\boldsymbol{x}) = \coprod_{i \in K_j} x_i = 1 - \prod_{i \in K_j}(1 - x_i) \qquad \text{for } j = 1, 2, \dots, k. \tag{4.18}$$

We see that $\kappa_j(\boldsymbol{x})$ represents the structure function of a parallel structure composed of the components in K_j. $\kappa_j(x)$ is therefore called the jth minimal *cut parallel structure*.

Because we know that the structure is failed if and only if at least one of the minimal cut parallel structures is failed, then

$$\phi(\boldsymbol{x}) = \prod_{j=1}^{k} \kappa_j(\boldsymbol{x}). \tag{4.19}$$

Hence, we can regard this structure as a series structure of the minimal cut parallel structures. By combining (4.18) and (4.19) we get

$$\phi(\boldsymbol{x}) = \prod_{j=1}^{k} \coprod_{i \in K_j} x_i. \tag{4.20}$$

Example 4.9 (Example 4.8 (cont.))

In the bridge structure, the minimal cut sets were $K_1 = \{1, 2\}$, $K_2 = \{4, 5\}$, $K_3 = \{1, 3, 5\}$, and $K_4 = \{2, 3, 4\}$. The corresponding minimal cut parallel structures

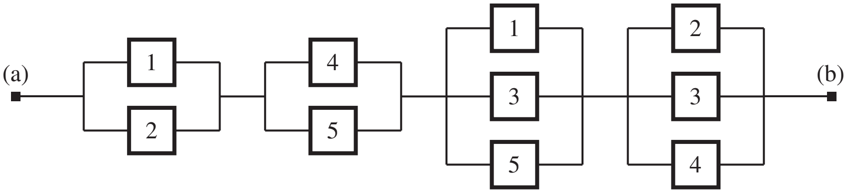


Figure 4.26 The bridge structure represented as a series structure of the minimal cut parallel structures.

become

$$\begin{aligned}\kappa_1(\boldsymbol{x}) &= x_1 \amalg x_2 = 1-(1-x_1)(1-x_2)\\ \kappa_2(\boldsymbol{x}) &= x_4 \amalg x_5 = 1-(1-x_4)(1-x_5)\\ \kappa_3(\boldsymbol{x}) &= x_1 \amalg x_3 \amalg x_5 = 1-(1-x_1)(1-x_3)(1-x_5)\\ \kappa_4(\boldsymbol{x}) &= x_2 \amalg x_3 \amalg x_4 = 1-(1-x_2)(1-x_3)(1-x_4),\end{aligned}$$

and we may find the structure function of the bridge structure by inserting these expressions into (4.19). The bridge structure may therefore be represented by the RBD in Figure 4.26. □

4.7.5 Pivotal Decomposition

The following identity holds for every structure function $\phi(\boldsymbol{x})$:

$$\phi(\boldsymbol{x}) \equiv x_i\phi(1_i,\boldsymbol{x}) + (1-x_i)\phi(0_i,\boldsymbol{x}) \text{ for all } \boldsymbol{x}. \tag{4.21}$$

We can easily see that this identity is correct from the fact that

$$x_i = 1 \Rightarrow \phi(\boldsymbol{x}) = \phi(1_i,\boldsymbol{x}) \quad \text{and} \quad x_i = 0 \Rightarrow \phi(\boldsymbol{x}) = \phi(0_i,\boldsymbol{x}).$$

Pivotal decomposition is also known as *Shannon expansion*.[7] In Chapter 6, we introduce probabilities of the various states. The probabilistic version of the pivotal decomposition Eq. (4.21) will then become nothing but the well-known *law of total probability* from probability theory (see Section 6.2.4).

Example 4.10 (Bridge structure)

Consider the bridge structure in Figure 4.23. The structure function $\phi(\boldsymbol{x})$ of this structure can be determined by pivotal decomposition with respect to component 3.

$$\phi(\boldsymbol{x}) = x_3\phi(1_3,\boldsymbol{x}) + (1-x_3)\phi(0_3,\boldsymbol{x}).$$

Here, $\phi(1_3,\boldsymbol{x})$ is the structure function of the structure in Figure 4.27,

$$\phi(1_3,\boldsymbol{x}) = (x_1 \amalg x_2)(x_4 \amalg x_5) = (x_1 + x_2 - x_1x_2)(x_4 + x_5 - x_4x_5),$$

[7] Named after the US mathematician Claude E. Shannon (1916–2001), who is claimed to be "the father of information theory."

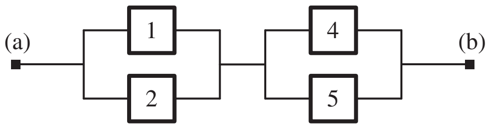


Figure 4.27 The structure $\phi(1_3, \boldsymbol{x})$ of the bridge structure.

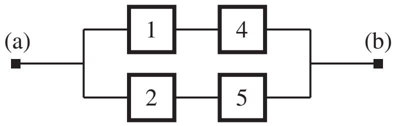


Figure 4.28 The structure $\phi(0_3, \boldsymbol{x})$ of the bridge structure.

where $\phi(0_3, \boldsymbol{x})$ is the structure function of the structure in Figure 4.28,

$$\phi(0_3, \boldsymbol{x}) = x_1x_4 \amalg x_2x_5 = x_1x_4 + x_2x_5 - x_1x_2x_4x_5.$$

Hence, the structure function of the bridge structure becomes

$$\begin{aligned}\phi(\boldsymbol{x}) &= x_3(x_1 + x_2 - x_1x_2)(x_4 + x_5 - x_4x_5)\\ &\quad + (1 - x_3)(x_1x_4 + x_2x_5 - x_1x_2x_4x_5).\end{aligned}$$

□

4.7.6 Modules of Coherent Structures

Consider the structure represented by the RBD in Figure 4.29. The structure may be split into three modules as shown by Figure 4.30, where the modules $\boxed{\text{I}}$, $\boxed{\text{II}}$, and $\boxed{\text{III}}$ are defined in Figure 4.31 The modules $\boxed{\text{I}}$, $\boxed{\text{II}}$, and $\boxed{\text{III}}$ may now be analyzed individually, and the results may be put together logically. Regarding this logical connection, it is important that the partitioning into subsystems is done in such a way that *each single component never appears within more than one of the modules.*

The term *coherent module* may be defined verbally as follows.

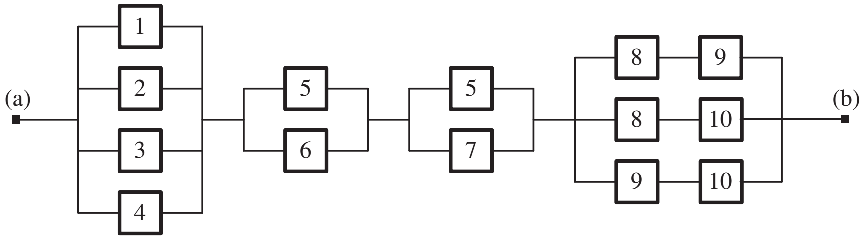


Figure 4.29 RBD.

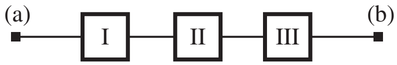


Figure 4.30 Structure of modules.

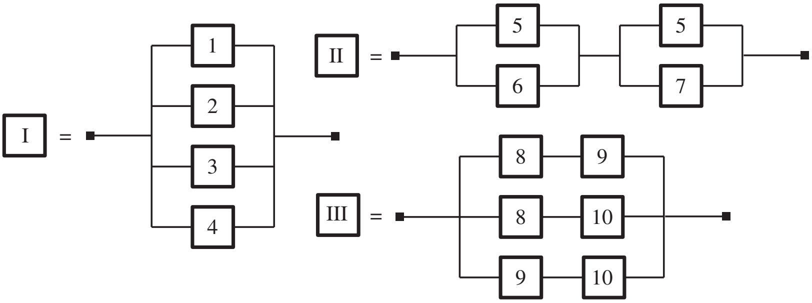


Figure 4.31 The three substructures.

Definition 4.7 (Coherent module – 1)
A coherent module (or subsystem) of a system is a subset of basic components of that system that are organized into a coherent structure of their own and that affect the system only through the performance of their components. Rephrasing: A coherent module is an assembly of components that can by itself be treated as a component of the system.[8] □

A more formal definition is given in Definition 4.8. When this partitioning is carried out in a specific way, described later, the procedure is called a *modular decomposition of the system.* In the following, we denote a system (C, ϕ), where C is the set of components and ϕ the structure function. Let A represent a subset of C,

$$A \subseteq C,$$

and A^c denote the complement of A with respect to C,

$$A^c = C - A.$$

We denote the elements in A by $i_1, i_2, \dots, i_\nu$, where $i_1 < i_2 < \cdots < i_\nu$. Let $\boldsymbol{x}^A$ be the state vector corresponding to the elements in A:

$$\boldsymbol{x}^A = (x_{i_1}, x_{i_2}, \dots, x_{i_\nu}),$$

and let

$$\chi(\boldsymbol{x}^A) = \chi(x_{i_1}, x_{i_2}, \dots, x_{i_\nu}),$$

be a binary function of $\boldsymbol{x}^A$. Obviously (A, χ) can be interpreted as a structure.

In our example, $C = \{1, 2, \dots, 10\}$. Let us choose $A = \{5, 6, 7\}$ and $\chi(\boldsymbol{x}^A) = (x_5 \amalg x_6)(x_5 \amalg x_7)$. (A, χ), then represents the substructure II. With this notation, a precise definition of a coherent module is

[8] Adapted from Birnbaum and Esary (1965).

Definition 4.8 (Coherent module – 2)
Let the coherent structure (C, ϕ) be given and let $A \subseteq C$. Then (A, χ) is said to be a coherent module of (C, ϕ), if $\phi(\boldsymbol{x})$ can be written as a function of $\chi(\boldsymbol{x}^A)$ and $\boldsymbol{x}^{A^c}$, $\psi(\chi(\boldsymbol{x}^A), \boldsymbol{x}^{A^c})$, where ψ is the structure function of a coherent structure. □

A is called a modular set of (C, ϕ), and if in particular $A \subset C$, (A, χ) is said to be a *proper* module of (C, ϕ).

What we actually do here is to consider all the components with index belonging to A as one "component" with state variable $\chi(\boldsymbol{x}^A)$. When we interpret the structure in this way, the structure function is

$$\psi(\chi(\boldsymbol{x}^A), \boldsymbol{x}^{A^c}).$$

In our example, we choose $A = \{5, 6, 7\}$. Then

$$\psi(\chi(\boldsymbol{x}^A), \boldsymbol{x}^{A^c}) = \chi(x_5, x_6, x_7)(\coprod_{i=1}^{4} x_i)(x_8 x_9 \amalg x_8 x_{10} \amalg x_9 x_{10}),$$

and because $A \subset C$, (A, χ) is a proper module of (C, ϕ). Now, let us define the concept of modular decomposition.

Definition 4.9 (Modular decomposition)
A modular decomposition of a coherent structure (C, ϕ) is a set of disjoint modules (A_i, χ_i), $i = 1, \ldots, r$, together with an organizing structure ω, such that

(1) $C = \cup_{i=1}^{r} A_i$; where $A_i \cap A_j = \emptyset$ for $i \neq j$
(2) $\phi(\boldsymbol{x}) = \omega[\chi_1(\boldsymbol{x}^{A_1}),\ \chi_2(\boldsymbol{x}^{A_2}), \ldots, \chi_r(\boldsymbol{x}^{A_r})]$ □

The "finest" partitioning into modules that we can have, is obviously to let each individual component constitute one module. The "coarsest" partitioning into modules is to let the whole system constitute one module. To be of practical use, a module decomposition should, if possible, be something between these two extremes. A module that cannot be partitioned into smaller modules without letting each components represent a module is called a *prime module.*

In our example, $\boxed{\text{III}}$ represents a prime module. But $\boxed{\text{II}}$ is not a prime module because it may be described as in Figure 4.32. and, hence, can be partitioned into two modules IIa and IIb as in Figure 4.33. This gives no guidance on how to determine individual prime modules in a system, but algorithms have been developed,

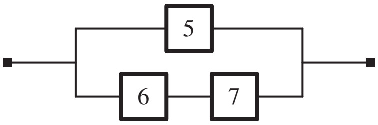
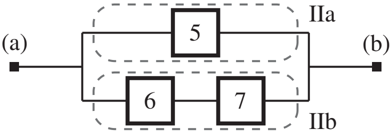


Figure 4.32 Module II.

Figure 4.33 Two prime modules.



for example by Chatterjee (1975), that can be used to find all the prime modules in a fault tree or in an RBD.

In Chapter 6, we justify the fact that it is natural to interpret the state vector as stochastic. In accordance with what we do in probability theory, we denote the state variables with *capital* letters from the end of the alphabet, for example $X_1, X_2, \dots, X_n$. Occasionally, two or more of these can be stochastically dependent. In such situations, it is advisable to try to "collect" the state variables in modules in such a way that dependency occurs only *within the modules*. If one succeeds in this, then the individual modules can be considered as being independent. This makes the further analysis simpler. This is elaborated further in Chapter 6.

4.8 Bayesian Networks

A *Bayesian network* (BN) is a graphical modeling tool that is used in many different application areas, including economy, medical diagnosis, and machine learning. The BN approach can also be used for system reliability analysis as an alternative to RBDs and fault trees. The term *Bayesian network* was coined by Judea Pearl in 1985, because the quantitative analyses of BNs are heavily based on Bayes' formula. Readers who are not familiar with Bayes' formula may consult Chapter 15.

A BN is a *directed acyclic graph* (DAG). Acyclic means that the BN cannot contain any cycles and "you" cannot come back to an earlier position. The network is made up of *nodes* and directed *arcs*, sometimes called *edges*. A node describes a state or condition, and an arc (or edge) indicates a *direct influence*. Because the arcs are directed, they can represent cause–effect relationships. In this book, the nodes are drawn as circles, and the directed arcs are drawn as arrows, but several other symbols for nodes are used in the literature and in computer programs.

As a modeling tool for system reliability, the nodes of the BN represent item states and the arcs illustrate how these states influence the states of other items. The simplest possible BN with to nodes and a single arc is shown in Figure 4.34. The directed arc from A to B indicates that A has a *direct influence* on B and that B is directly influenced by A. The arc from A to B is sometimes written as $\langle A, B \rangle$ and indicates that the state of B depends on the state of A.

In Figure 4.34, node A is called a *parent node* of node B, and node B is called a *child node* of node A. A node with no parents is called a *root node*. In this figure, A is therefore a root node. A node with no child (no descendant) is called a *leaf node*.

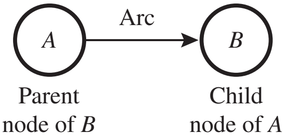


Figure 4.34 The main BN symbols.

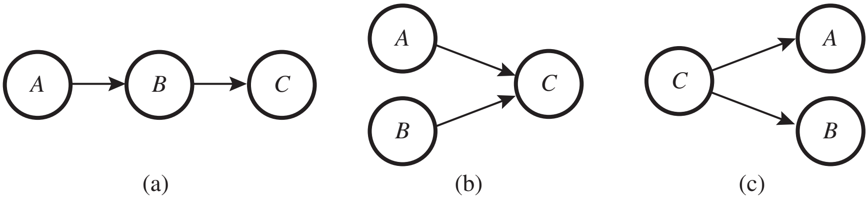


Figure 4.35 (a) Linear, (b) converging, and (c) diverging BN with three nodes.

In Figure 4.34, B is a leaf node. A BN (or a module of a BN) can be linear, converging, or diverging, as shown in Figure 4.35. The parents of a node X are sometimes written as pa(X). The parents of node C in the converging BN in Figure 4.35, hence, are $\text{pa}(C) = \{A, B\}$.

This section is delimited to the BN graph properties, whereas the probabilistic properties are treated in Section 6.9.

4.8.1 Illustrative Examples

We illustrate the application of BNs through three simple examples, a series structure of two components, a parallel structure of two components, and a 2oo3 structure. Each node has two possible states: 1 (=functioning) and 0 (= failed). The BN for a system of two components is shown in Figure 4.36 and is identical for a series and a parallel structure. The states of the parent nodes, A and B, influence the state of the child node, S. The structure of the influence is described by a *truth table*, which is different for a series structure and a parallel structure. The structure of the system is therefore determined by this truth table.

Example 4.11 (BN for a series structure)
Consider a series structure of two independent components A and B, illustrated by the BN in Figure 4.36. The state of the system S is seen to be directly influenced

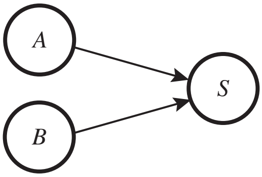


Figure 4.36 BN for a system S of two independent components A and B.

Table 4.7 Truth table for a series structure of two components.

Components		System state
A	*B*	*S*
0	0	0
0	1	0
1	0	0
1	1	1

by the state of component A and the state of component B. The two components A and B are independent and do not directly influence each other.

The properties of the series structure are specified in the truth table in Table 4.7. Table 4.7 shows that the system is functioning ($S = 1$) if and only if components A and B are both functioning (i.e. have state 1). □

Example 4.12 (BN for a parallel structure)
Consider a parallel structure of two independent components A and B, illustrated by the BN in Figure 4.36. The properties of the parallel structure are specified in the truth table in Table 4.8. Table 4.8 shows that the system state is 1 if at least one of the components is functioning. □

Example 4.13 (2oo3 structure)
Consider a 2oo3 structure of three components A, B, and C, illustrated by the BN in Figure 4.37. The properties of the 2oo3 structure is specified by the truth table in Table 4.9. Table 4.9 shows that the system state is 1 if at least two of the three components are functioning. □

Table 4.8 Truth table for a parallel structure of two components.

Components		System state
A	*B*	*S*
0	0	0
0	1	1
1	0	1
1	1	1

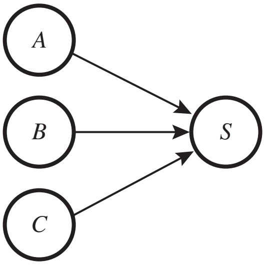


Figure 4.37 BN for a 2oo3 structure *S* of three components *A*, *B*, and *C*.

Table 4.9 Truth table for the 2oo3 structure.

Components			System state
A	***B***	***C***	***S***
0	0	0	0
0	0	1	0
1	0	0	0
0	1	0	0
0	1	1	1
1	0	1	1
1	1	0	1
1	1	1	1

A fault tree can directly be represented as a BN as illustrated by the simple example in Figure 4.38.

BNs can replace any RBD or fault tree. The graph is easy to establish and shares many of the positive features of RBDs and fault trees. As for RBDs and fault trees, the BN of a complicated system may be built by combining BNs for simpler parts. The resulting graph is intuitive and easy to communicate to people who are not experts in reliability analysis.

Both RBDs and fault trees are forced into the pure Boolean logic as input events can only be combined by AND and OR relations. The BN is more flexible because each node can have more than two states and because direct influences from parent nodes can be combined in more general ways. BNs can therefore be seen as an extension of RBDs and fault trees for reliability analysis.

Another extension is that BNs may be used to model many other influences than states of components. A BN may, for example be used to model how a machine is influenced by maintenance M, humidity H, type of lubrication L, and so on. To assess the influences, a limited number of values have to be specified for the influencing variables (M, H, L, etc.).

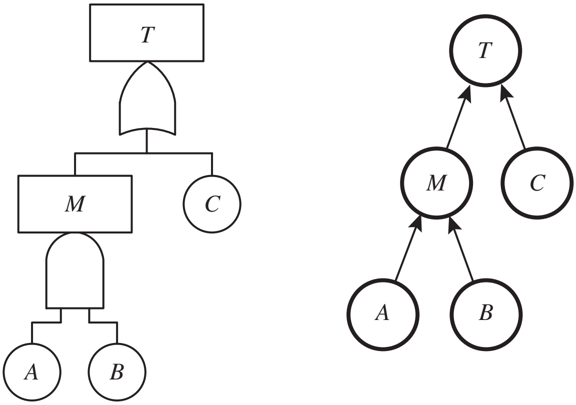


Figure 4.38 A simple fault tree and the corresponding BN.

Probabilistic Evaluation

As for RBDs and fault trees, node probabilities can be entered into the BN and used to find the probability of system failure or function. Probabilistic evaluation of BN is discussed in Chapter 6.

4.9 Problems

4.1 Assume that you have a bike and that the bike is important for you during most seasons of the year. In this problem, you are asked to perform a qualitative system reliability analysis considering the phases that were introduced in Chapters 2 and 4:

(a) System familiarization, including assumptions and illustrations to support your definition of system and system boundaries and interfaces.

(b) Functional analysis, using one of the techniques presented in Chapter 2.

(c) Failure analysis by FMECA as described in Section 4.2.3.

(d) Failure analysis by FTA as described in Section 4.3.3.

(e) For the FTA, you should identify the minimal cut sets and elaborate briefly about what this information gives you in terms of prioritizing for inspection and maintenance.

(f) Discuss briefly the value of information/insight you get from using FMECA compared to using FTA.

4.2 Consider the subsea shutdown valve in Figure 4.39. The valve is a spring-loaded, fail-safe close gate valve that is held open by hydraulic pressure. The gate is a solid block with a cylindrical hole with the same diameter as the pipeline. To open the valve, hydraulic pressure is applied on the upper

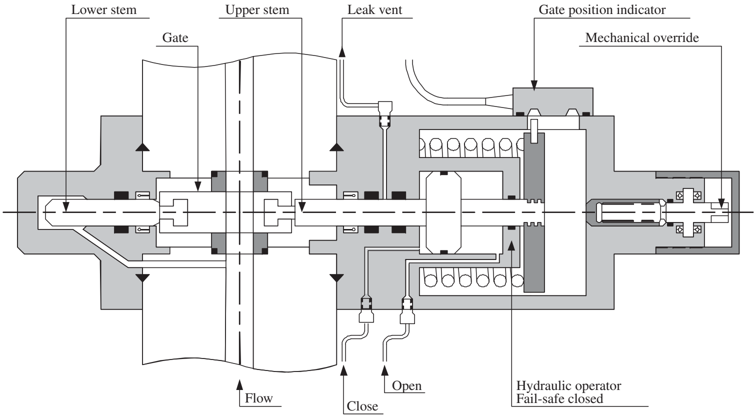


Figure 4.39 Hydraulically operated gate valve (Problem 4.2).

side of the piston. The pressure forces the piston, the piston rod, and the gate downwards until the hole in the gate is in line with the pipeline. When the pressure is bled off, the spring forces the piston upwards until the hole in the gate is no longer in contact with the pipeline conduct. The solid part of the gate is now pressed against the seat seal and the valve is closed.

(a) Carry out FMECA of the shutdown valve by following the procedure described in Section 4.2.3.

4.3 A loss-of-coolant accident (LOCA) is a serious accident in a nuclear power plant, and several protection systems are installed to prevent and/or mitigate this type of accidents. One of these protection systems is the emergency core cooling system (ECCS). The purpose of the ECCS is to remove residual heat from the reactor fuel rods in the event of a failure of the normal reactor cooling system. The ECCS has several subsystems, and one of these is the low-pressure coolant injection (LPCI) system. The main components of an LPCI system are three pressure transmitters (PT), a logic solver (LS), four low-pressure injection pumps (LPIPs), a refueling water storage tank (RWST), piping, and a sump. Each pump (LPIP) is driven by a dedicated diesel generator (EDG). In case of a LOCA incident, the reactor cooling system is depressurized and will empty quickly. At this point, the core is uncovered, and if no action is taken, the core will melt. In this situation, the purpose of the LPCI is to inject water into the reactor vessel to flood and cool the core. Refilling takes a few minutes. Three pressure transmitters (PT) are installed to detect low pressure in the reactor cooling system. When two of the three PTs detect low pressure, a signal to activate the LPCI is sent from the logic solver to the LPIPs and the EDGs. If the LPCI fails to refill and re-flood the tank, a severe accident (meltdown) will occur. Two of the four LPIPs need to work in order to successfully refill and re-flood the core. The piping and the sump are left out of the analysis.

(a) Establish an RBD for the LPCI with respect to the system's main function as a safety barrier.

(b) List the minimal cut sets of the system. What do we mean by the order of a minimal cut set?

(c) Construct a BN for the LPCI corresponding to a failure of its main function as a safety barrier.

4.4 Consider the RBD in Figure 4.40.

(a) Find the structure function of the structure by using pivotal decomposition.

(b) Find all the minimal path sets and all the minimal cut sets of the structure represented by the RBD.

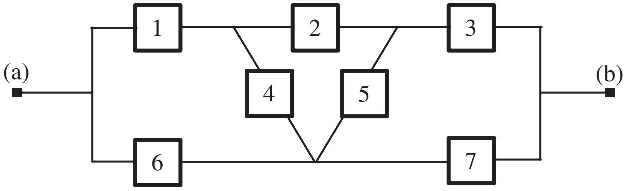


Figure 4.40 RBD for Problem 4.4.

4.5 A structure has the following minimal path sets: $P_1 = \{1,4\}, P_2 = \{2,3\}, P_3 = \{2,4\}$.

(a) Draw the corresponding RBD.

(b) Find the minimal cut sets of the structure.

(c) Establish the structure function for the structure.

4.6 In a chemical process plant, several compounds are mixed in a chemical reactor. Consider the pipeline where one of these compounds is fed into the reactor. If too much of this compound enters into the reactor, the mixture will come out of balance and the pressure in the reactor will increase. This is a very critical event and is controlled by the safety-instrumented system (SIS) illustrated in Figure 4.41. Three flow transmitters are installed in the pipeline. When at least two of the three flow transmitters detect and alarm "high flow." a signal is sent to the main logic solver that will transmit a signal to close the two shutdown valves in the pipeline. In addition, three pressure transmitters are installed in the reactor. When at least two of

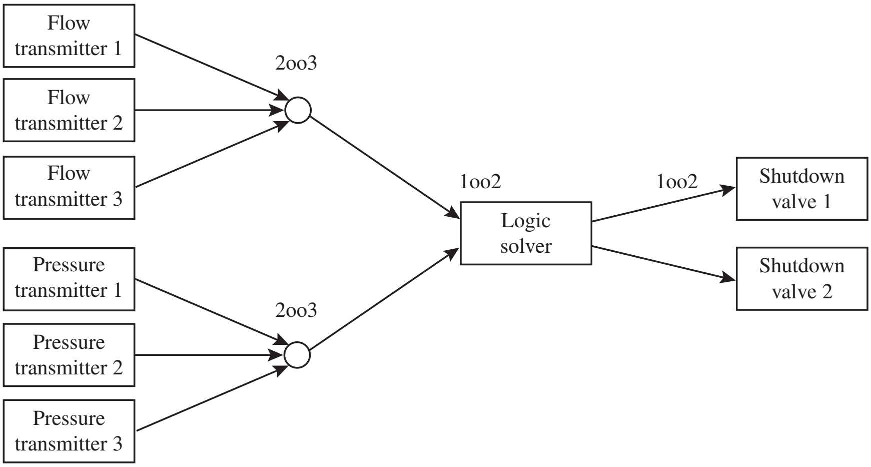


Figure 4.41 RBD for Problem 4.6.

the three pressure transmitters detect and alarm "high pressure," a signal enters the main logic solver that will transmit a signal to close the two shutdown valves in the pipeline – and stop the flow of the compound into the reactor. Any unplanned shutdown of the reactor may also lead to dangerous situations, and spurious shutdowns (i.e. caused by false alarms) should therefore be avoided. The three flow transmitters are of the same type and are, as illustrated in Figure 4.41, configured as a 2oo3 structure. In the same way, the three pressure transmitters are of the same type and also configured as a 2oo3 structure. The logic solver transmits a shutdown signal to the valves if it receives a signal from either the flow transmitters or the pressure transmitters. The main logic solver is therefore a 1oo2 configuration. It is sufficient that one of the two shutdown valves (of the same type) is able to close to stop the flow of the compound into the reactor. The shutdown valves are therefore a 1oo2 structure. The 2oo3 votings for the flow and pressure transmitters are physically modules of the logic solver, even if they are drawn as separate entities in Figure 4.41. The two shutdown valves are kept open in normal operation and should shut the flow in the pipeline when high flow or high pressure is "detected" by the transmitters.

(a) Establish a RBD for the whole system with respect to the system's main function as a safety barrier.

(b) Determine all minimal cut sets.

(c) Construct a BN for the whole system corresponding to failure of its main function as a safety barrier.

4.7 Figure 4.42 shows a sketch of the lubrication system on a ship engine. The separator separates water from the oil lubricant. The separator is functioning satisfactorily only when the oil is heated to a specified temperature. When the water content in the oil is too high, the quality of the lubrication becomes too low, and this may lead to damage or breakdown of the engine. The engine generally requires

- Sufficient throughput of oil/lubricant.
- Sufficient quality of the oil/lubricant.

The oil throughput is sufficient when at least one cooler is functioning, at least one filter is open (i.e. not clogged), and the pump is functioning. In addition, all necessary pipelines must be open, no valves must be unintentionally closed, the lubrication channels in the engine must be open (not clogged) and the lubrication system must not have significant leakages to the environment. We assume that the probabilities of these "additional" events are very low and that these events therefore may be neglected.

The quality of the oil is sufficient when

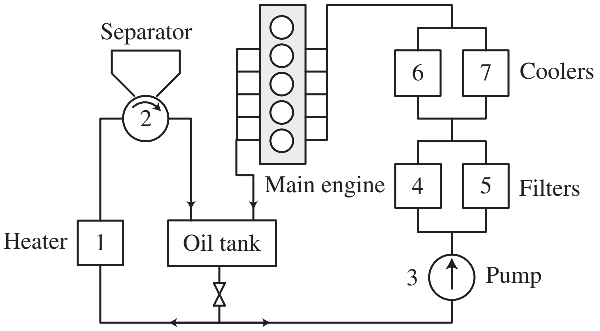


Figure 4.42 Lubrication system on a ship engine (Problem 4.7).

- Both coolers are functioning (with full throughput) such that the temperature of the oil to the engine is sufficiently low.
- None of the filters is clogged, and there is no holes in the filters.
- The separator system is functioning.

(a) Construct a fault tree with respect to the TOP event "Too low throughput of oil/lubricant."

(b) Construct a fault tree with respect to the TOP event "Too low quality of the oil/lubricant."

4.8 Use MOCUS to identify all the minimal cut sets of the fault tree in Figure 4.9.

4.9 Show that

(a) If ϕ represents a parallel structure, then:

$$\phi(\boldsymbol{x} \amalg \boldsymbol{y}) = \phi(\boldsymbol{x}) \amalg \phi(\boldsymbol{y}).$$

(b) If ϕ represents a series structure, then:

$$\phi(\boldsymbol{x} \cdot \boldsymbol{y}) = \phi(\boldsymbol{x}) \cdot \phi(\boldsymbol{y}).$$

4.10 The dual structure $\phi^D(\boldsymbol{x})$ to a given structure $\phi(\boldsymbol{x})$ is defined by

$$\phi^D(\boldsymbol{x}) = 1 - \phi(\mathbf{1} - \boldsymbol{x}),$$

where $(\mathbf{1} - \boldsymbol{x}) = (1 - x_1, 1 - x_2, \dots, 1 - x_n)$.

(a) Show that the dual structure of a $k$oo$n$ structure is a $(n - k + 1)$oo$n$ structure.

(b) Show that the minimal cut sets for ϕ are minimal path sets for ϕ^D, and vice versa.

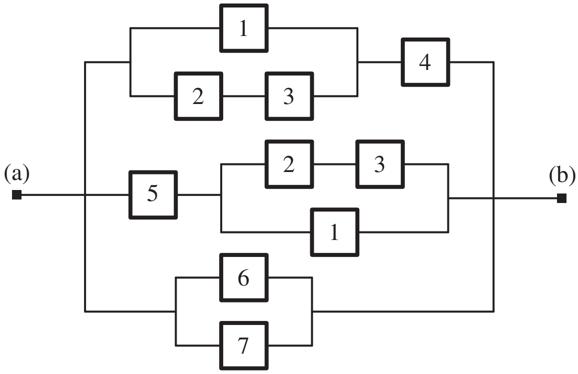


Figure 4.43 RBD for Problem 4.11.

4.11 Determine the structure function of the structure in Figure 4.43 by applying an appropriate modular decomposition.

4.12 Consider the fault tree in Figure 4.44.
(a) Use MOCUS to identify all the minimal path sets of the fault tree.
(b) Show that the system may be represented by the RBD in Figure 4.45.

4.13 Determine the structure function of the structure in Figure 4.46 by using pivotal decomposition.

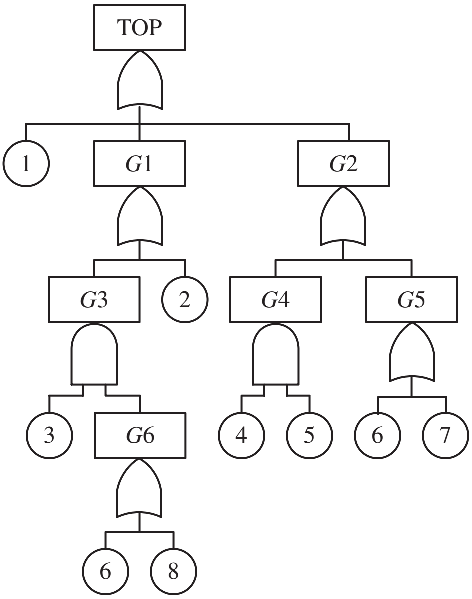


Figure 4.44 Fault tree for Problem 4.12.

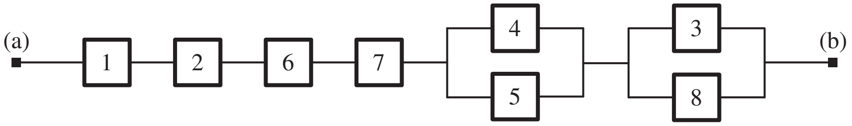


Figure 4.45 RBD for Problem 4.12.

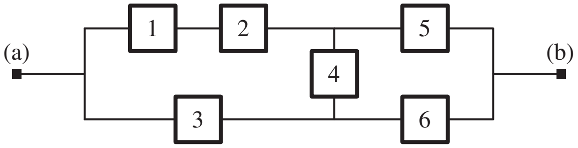


Figure 4.46 RBD for Problem 4.13.

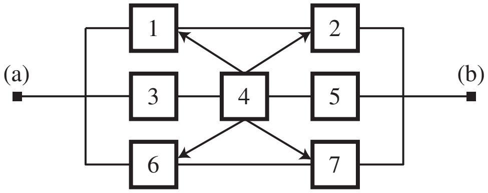


Figure 4.47 RBD for Problem 4.14.

4.14 Determine the structure function of the structure in Figure 4.47.

4.15 Construct a BN corresponding to the fault tree in Figure 4.44. Record the assumptions you make during the construction of the BN.

4.16 A list of the minimal cut sets of a structure or a fault tree can be used to determine the corresponding minimal path sets. Describe how this can be done and exemplify your approach by using the fault tree in Figure 4.12.

References

Barlow, R.E. and Lambert, H.E. (1975). Introduction to fault tree analysis. In: *Reliability and Fault Tree Analysis: Theoretical and Applied Aspects of System Reliability and Safety Assessment* (ed. R.E. Barlow, J.B. Fussell, and N.D. Singpurwalla). Philadelphia, PA: SIAM, 101–126.

Barlow, R.E. and Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing, Probability Models*. New York: Holt, Rinehart, and Winston.

Birnbaum, Z.W. and Esary, J.D. (1965). Modules of coherent binary systems. *Journal of the Society for Industrial and Applied Mathematics* 13 (2): 444–462.

CCPS (2008). *Guidelines for Hazard Evaluation Procedures*, 3e. Hoboken, NJ: Wiley.

Chatterjee, P. (1975). Modularization of fault trees; a method to reduce the cost of analysis. In: *Reliability and Fault Tree Analysis: Theoretical and Applied Aspects of System Reliability and Safety Assessment* (ed. R.E. Barlow, J.B. Fussell, and N.D. Singpurwalla). Philadelphia, PA: SIAM, 101–126.

Dugan, J.B. (2000). Galileo: a tool for dynamic fault tree analysis. In: *Computer Performance Evaluation. Modelling Techniques and Tools* (ed. B.R. Haverkort, H.C. Bohnenkamp, and C.U. Smith), 328–331. Berlin: Springer-Verlag.

Ford (2004). Failure mode and effects analysis handbook, *Handbook Version 4.1*. Dearborn, MI: Ford Design Institute.

Goble, W.M. and Brombacher, A.C. (1999). Using a failure modes, effects and diagnostic analysis (FMEDA) to measure diagnostic coverage in programmable electronic systems. *Reliability Engineering & System Safety* 66 (2): 145–148.

Haapanen, P. and Helminen, A. (2002). Failure Mode and Effects Analysis of Software-Based Automation Systems. *Technical Report STUK-YTO-TR 190*. Helsinki, Finland: Radiation and Nuclear Safety Authority.

IEC 60812 (2018). *Procedure for failure mode and effects analysis (FMEA and FMECA)*, *International standard*. Geneva: International Electrotechnical Commission.

IEC 61025 (2006). *Fault tree analysis (FTA)*, *International standard*. Geneva: International Electrotechnical Commission.

IEEE Std. 352 (2016). *IEEE guide for general principles of reliability analysis of nuclear power generating station protection station systems and other nuclear facilities*, *Standard*. New York: Institute of Electrical and Electronics Engineers.

MIL-STD-1629A (1980). *Procedures for performing a failure mode, effects, and criticality analysis*, *Military standard*. Washington, DC: U.S. Department of Defense.

NASA (2002). Fault tree handbook with aerospace applications, *Handbook*. Washington, DC: U.S. National Aeronautics and Space Administration.

NUREG-0492 (1981). Fault Tree Handbook, *Handbook NUREG-0492*. Washington, DC: U.S. Nuclear Regulatory Commission.

NUREG-75/014 (1975). Reactor Safety: An Assessment of Accident Risk in U.S. Commercial Nuclear Power Plants, *Report NUREG-75/014*. Washington, DC: U.S. Nuclear Regulatory Commission.

NUREG/CR-1278 (1983). Handbook of Human Reliability Analysis in Nuclear Power Plant Applications, *Handbook NUREG/CR-1278*. Washington, DC: U.S. Nuclear Regulatory Commission.

SAE ARP 5580 (2012). *Recommended failure modes and effects analysis (FMEA) practices for non-automobile applications*, *Recommended Practice ARP 5580*. Warrendale, PA: SAE International.

SAE J1739 (2009). *Potential failure mode and effects analysis in design (Design FMEA) and potential failure mode and effects analysis in manufacturing and assembly processes (Process FMEA)*, *Standard*. Warrendale, PA: SAE International.

Schmittner, C., Gruber, T., Puschner, P., and Schoitsch, E. (2014). Security application of failure mode and effect analysis (FMEA). In: *Computer Safety, Reliability, and Security. SAFECOMP 2014*, *LNCS 8666* (ed. A. Bondavalli and F.D. Giandomenico), 310–325. Springer.

Shooman, M.L. (1968). *Probabilistic Reliability: An Engineering Approach*. New York: McGraw-Hill.

Xu, H., Dugan, J.B., and Meshkat, L. (2006). A dynamic fault tree model of a propulsion system. In: *Proceedings of the 8th International Conference on Probabilistic Safety Assessment & Management (PSAM)* (ed. M.G. Stamatelatos and H.S. Blackman). ASME Press, https://doi.org/10.1115/1.802442.paper194, 1–10.

5

Probability Distributions in Reliability Analysis

5.1 Introduction

This chapter introduces the main time-to-failure distributions and the main probabilistic metrics for the reliability of a nonrepairable item. In some cases, the item may be literarily nonrepairable, meaning that it is discarded when the first failure occurs. In other cases, the item may be repaired, but we are not interested in what happens to the item after the first failure.

First, we introduce five reliability metrics for a nonrepairable item and illustrate how these can be understood by using probability theory. The five reliability metrics are

- The survivor function $R(t)$
- The failure rate function $z(t)$
- The mean time-to-failure (MTTF)
- The conditional survivor function
- The mean residual lifetime (MRL)

Thereafter, we introduce a number of probability distributions that may be used to model the time-to-failure of a nonrepairable item. The following time-to-failure distributions are covered:

- The exponential distribution
- The gamma distribution
- The Weibull distribution
- The normal distribution
- The lognormal distribution
- Three different extreme value distributions
- Time-to-failure distributions with covariates.

System Reliability Theory: Models, Statistical Methods, and Applications, Third Edition.
Marvin Rausand, Anne Barros, and Arnljot Høyland.

Companion Website: www.wiley.com/go/SystemReliabilityTheory3e

Next, three discrete distributions: the binomial, the geometric, and the Poisson distributions are introduced. The chapter is concluded by a brief survey of some broader classes of time-to-failure distributions.

5.1.1 State Variable

The state of the item at time t can be described by the state variable $X(t)$, where

$$X(t) = \begin{cases} 1 & \text{if the item is functioning at time } t \\ 0 & \text{if the item is in a failed state at time } t \end{cases}.$$

The state variable of a nonrepairable item is shown in Figure 5.1 and is a random variable.

5.1.2 Time-to-Failure

The *time-to-failure* or *lifetime* of an item is the time elapsing from when the item is put into operation until it fails for the first time. At least to some extent, the time-to-failure is subject to chance variations. It is therefore natural to interpret the time-to-failure as a random variable, T. We mainly use the term *time-to-failure* but will also use the term lifetime in some cases. The connection between the state variable $X(t)$ and the time-to-failure T is shown in Figure 5.1. Unless stated otherwise, it is always assumed that the item is new and in a functioning state when it is started up at time $t = 0$.

Observe that the time-to-failure T is not always measured in calendar time. It may also be measured by more indirect time concepts, such as

- Number of times a switch is operated.
- Number of kilometers driven by a car.
- Number of rotations of a bearing.
- Number of cycles for a periodically working item.

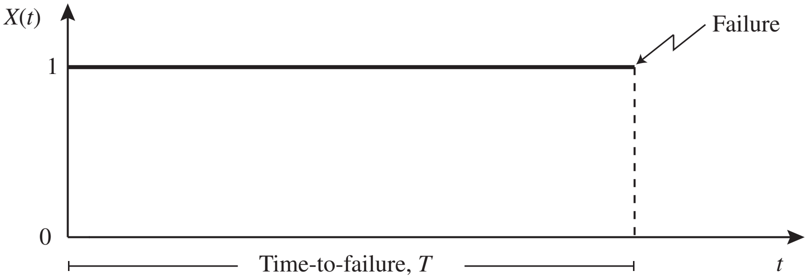


Figure 5.1 The state variable and the time-to-failure of an item.

From these examples, we observe that the time-to-failure may often be a discrete variable. A discrete variable can, however, be approximated by a continuous variable. Here, unless stated otherwise, we assume that the time-to-failure T is continuously distributed.

5.2 A Dataset

Consider an experiment where n identical and independent items are put into operation at time $t = 0$. We leave the n items without intervention and observe the times-to-failure of each item. The outcome of this experiment is the *dataset* $\{t_1, t_2, \dots, t_n\}$. Such a dataset is also called a *historical* dataset to point out that the dataset stems from past operations of items and that the times are recorded and known.

In probability and reliability theory, we accept that we cannot know in advance which outcome will be obtained for a future experiment, and we therefore try to predict the probability of occurrence for each possible outcome based on historical data. These probabilities make sense when

(1) The past and future experiments can be considered identical and independent (performed in the same way, under the same conditions, and without any dependencies between the outcomes).
(2) The past experiments have been repeated a large number of times.

Some common trends and some variations in the historical dataset can often be identified, and these allow us to make useful predictions with tractable uncertainties for future experiments. Generally speaking, this requires three main steps: (i) data analysis to extract relevant trends and variations, (ii) modeling to put relevant information into a format that allows probability calculations for new items, and (iii) quantification and probability calculation.

This section treats the last step, whereas data analysis is dealt with in Chapter 14. Modeling of a single nonrepairable item is discussed in the current chapter, but this is also a main topic in most of the chapters related to systems (i.e. several items in interaction). Here, a brief overview is given to make the reader understand the input that is provided from the data analysis and modeling phases for a single nonrepairable item. We present the quantities of interest that can be calculated, and how. As an example, consider the dataset of the $n = 60$ observed times-to-failure in Table 5.1.

5.2.1 Relative Frequency Distribution

From the dataset in Table 5.1, a *histogram* representing the number of failures within specified time intervals may be constructed. This histogram is called the

Table 5.1 Historical dataset.

23	114	125	459	468	472
520	558	577	616	668	678
696	700	724	742	748	761
768	784	785	786	792	811
818	845	854	868	870	871
878	881	889	892	912	935
964	965	970	971	976	1001
1006	1013	1020	1041	1048	1049
1065	1084	1102	1103	1139	1224
1232	1304	1310	1491	1567	1577

frequency distribution of the recorded times-to-failure. By dividing the number of failures in each interval by the total number of failures, the relative number of failures that occur in the intervals is obtained. The resulting histogram is called the *relative frequency distribution*. The histogram is made such that the area of each bar is equal to the percentage of all the failures that occurred in that time interval, the total area under the histogram is 100% (= 1).

The relative frequency distribution can be used to estimate and illustrate reliability quantities, such as the empirical mean, the empirical standard deviation (SD), and the probability of surviving a given time. The (empirical) mean is shown in Figure 5.2a.

The histogram may in practice have one or more maximum values and some reasonable standard deviation around these maximum values. Different failure modes and/or different failure causes may lead to more than one maximum value. The histogram in Figure 5.2 shows that the times-to-failure are spread around the mean and that there are some early failures that occurred short time after start-up.

5.2.2 Empirical Distribution and Survivor Function

Another way to present the dataset in Table 5.1 is to construct an *empirical survivor function*. This is done by sorting the times-to-failure, starting with the shortest and ending with the longest time-to-failure. For each time-to-failure, the proportion (i.e. percentage) of items that survived this time-to-failure is plotted. The obtained function is obviously decreasing from 1 to 0. The proportion of items that survived say, t_i, can be used to estimate the probability that an item will survive time t_i in a future experiment. The empirical survivor function for the dataset in Table 5.1 is shown in Figure 5.2b.

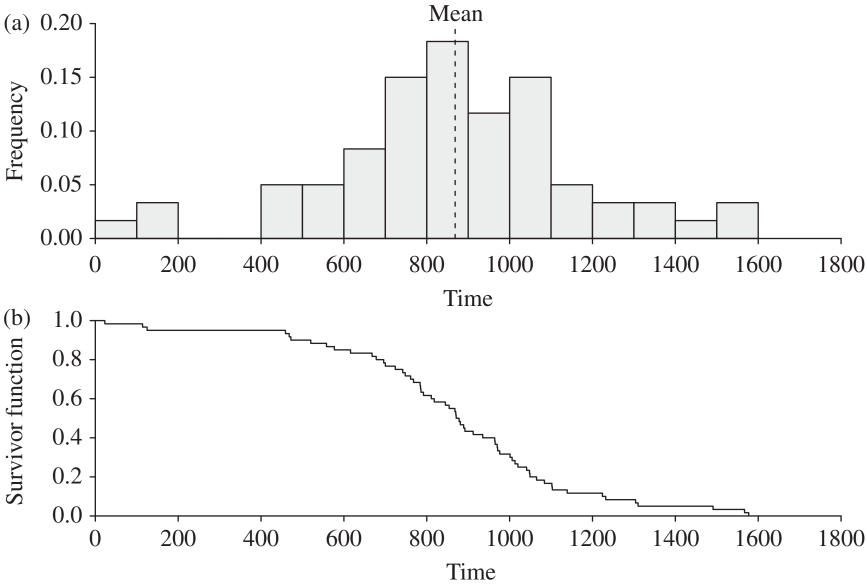


Figure 5.2 Relative frequency distribution (histogram) (a) and empirical survivor function (b) for the dataset in Table 5.1.

The empirical survivor function may be used to estimate the probabilities of interest for future experiments, but it is more common to fit a continuous function to the empirical function and to use this continuous function in the reliability studies.

5.3 General Characteristics of Time-to-Failure Distributions

Assume that the time-to-failure T is a continuously distributed random variable with *probability density function* $f(t)$ and *probability distribution function* $F(t)$.[1]

$$F(t) = \Pr(T \le t) = \int_0^t f(u)\, du \qquad \text{for } t > 0. \tag{5.1}$$

The event $T \le t$ occurs when the item fails before time t and $F(t)$ is therefore the probability that the item fails in the time interval $(0, t]$. The probability density function $f(t)$ is from (5.1) the derivative of $F(t)$.

$$f(t) = \frac{d}{dt}F(t) = \lim_{\Delta t \to 0} \frac{F(t+\Delta t) - F(t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{\Pr(t < T \le t + \Delta t)}{\Delta t}.$$

[1] $F(t)$ is also called the *cumulative distribution function*.

This implies that when Δt is small,

$$\Pr(t < T \leq t + \Delta t) \approx f(t)\Delta t. \tag{5.2}$$

When we, at time $t = 0$, look into the future, $\Pr(t < T \leq t + \Delta t)$ tells us the probability that the item will fail in the short interval $(t, t + \Delta t]$. When this probability is high (low), the probability density $f(t)$ is high (low), and this is the reason why $f(t)$ is also called the *failure density* function.

An example of a probability density curve is shown in Figure 5.3. The time unit used in Figure 5.3 is not given. It may, for example, be one year or 10 000 hours.

To be a proper probability density function, $f(t)$ must satisfy the two conditions

(1) $f(t) \geq 0$ for all $t \geq 0$

(2) $\int_0^\infty f(t)\, dt = 1$.

When a probability density function is specified, only its nonzero part is usually stated, and it is tacitly understood that the probability density function is zero over any unspecified region. Because the time-to-failure T cannot take negative values, $f(t)$ is only specified for nonnegative values of t.

For a continuous random variable, the probability that T is *exactly* equal to t is always zero, that is $\Pr(T = t) = 0$ for all specific values of t. This means that $\Pr(T \leq t) = \Pr(T < t)$ and $\Pr(T \geq t) = \Pr(T > t)$.

Because $f(t) \geq 0$ for all t, the probability distribution function must satisfy

(1) $0 \leq F(t) \leq 1$ because $F(t)$ is a probability

(2) $\lim_{t\to 0} F(t) = 0$

(3) $\lim_{t\to \infty} F(t) = 1$

(4) $F(t_1) \geq F(t_2)$ when $t_1 > t_2$, that is, $F(t)$ is a nondecreasing function of t.

The distribution function $F(t)$ and the probability density function $f(t)$ for the same distribution are shown in Figure 5.4. The probability density function (dashed line) is the same as in Figure 5.3, but the scale of the y-axis is changed.

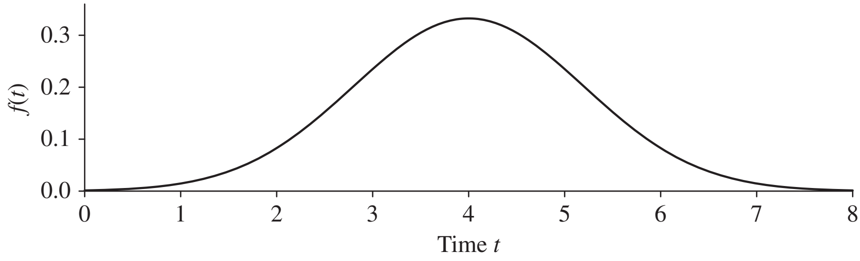


Figure 5.3 Probability density function, $f(t)$ for the time-to-failure T.

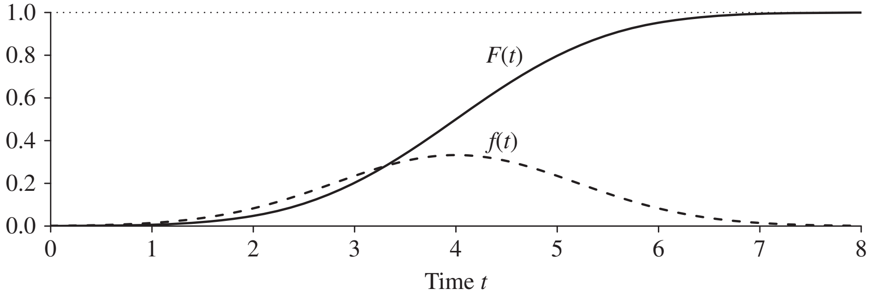


Figure 5.4 The distribution function $F(t)$ (fully drawn line) together with the corresponding probability density function $f(t)$ (dashed line).

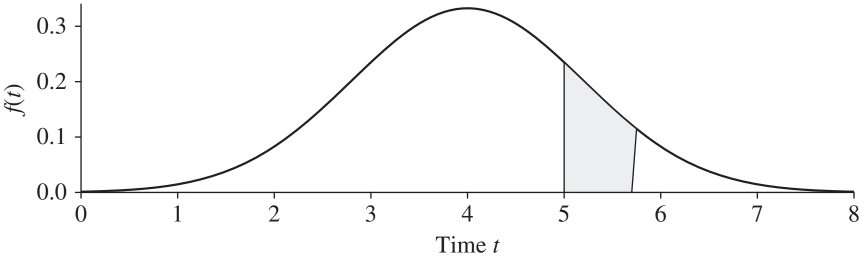


Figure 5.5 Illustration of the integral calculation of the probability to fail within $(t_1, t_2] = (5.0, 5.7]$.

The probability that a failure occurs in an interval $(t_1, t_2]$ is

$$\Pr(t_1 < T \le t_2) = F(t_2) - F(t_1) = \int_{t_1}^{t_2} f(u)\, du. \tag{5.3}$$

This quantity corresponds to the gray area below $f(t)$ on Figure 5.5 if $t_1 = 5$ and $t_2 = 5.7$ time units. Depending on the values of t_1, t_2 and $f(t)$ in $(t_1, t_2]$, the gray area will change and the probability to fail in $(t_1, t_2]$ will change as well.

5.3.1 Survivor Function

The *survivor function* of an item is defined by

$$R(t) = 1 - F(t) = \Pr(T > t) \tag{5.4}$$

or, equivalently

$$R(t) = 1 - \int_0^t f(u)\, du = \int_t^\infty f(u)\, du. \tag{5.5}$$

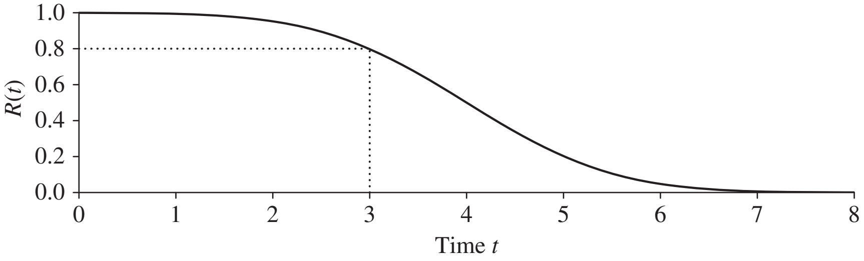


Figure 5.6 The survivor function $R(t)$.

Hence, $R(t)$ is the probability that the item does not fail in the time interval $(0, t]$, or in other words, the probability that the item survives the time interval $(0, t]$ and is still functioning at time t.

The survivor function is also called the *survival probability function*. Some authors define reliability by $R(t)$ and consequently call it the *reliability function*. This is also the reason for using the symbol $R(t)$. The survivor function that corresponds to the probability density function in Figure 5.3 is shown in Figure 5.6.

In Figure 5.6, the dotted line indicates that the probability that the item survives 3 time units is approximately 0.80 (= 80%). We may also read this in the opposite way, and find that the time corresponding to 80% survival is approximately 3 time units.

5.3.2 Failure Rate Function

The probability that an item will fail in the time interval $(t, t + \Delta t]$ when we know that the item is functioning at time t, is

$$\Pr(t < T \le t + \Delta t \mid T > t) = \frac{\Pr(t < T \le t + \Delta t)}{\Pr(T > t)} = \frac{F(t + \Delta t) - F(t)}{R(t)}.$$

By dividing this probability by the length of the time interval, Δt and letting $\Delta t \to 0$, we get the *failure rate function* $z(t)$ of the item

$$\begin{aligned} z(t) &= \lim_{\Delta t \to 0} \frac{\Pr(t < T \le t + \Delta t \mid T > t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \frac{1}{R(t)} = \frac{f(t)}{R(t)}. \end{aligned} \tag{5.6}$$

This implies that when Δt is small,

$$\Pr(t < T \le t + \Delta t \mid T > t) \approx z(t)\Delta t.$$

Because $R(t)$ is a probability and ≤ 1 for all t, (5.6) implies that $z(t) \ge f(t)$ for all $t \ge 0$.

Remark 5.1 (The difference between $f(t)$ and $z(t)$)
Observe the similarity and the difference between the probability density function $f(t)$ and the failure rate function $z(t)$.

$$\Pr(t < T \leq t + \Delta t) \approx f(t)\Delta t. \tag{5.7}$$

$$\Pr(t < T \leq t + \Delta t \mid T > t) \approx z(t)\Delta t. \tag{5.8}$$

Say that we start out with a new item at time $t = 0$ and at time $t = 0$ ask, "What is the probability that this item will fail in the interval $(t, t + \Delta t]$?" According to (5.7), this probability is approximately equal to the probability density function $f(t)$ at time t multiplied by the length of the interval Δt. Next consider an item that has survived until time t, and we then ask, "What is the probability that this item will fail in the next interval $(t, t + \Delta t]$?" This (conditional) probability is according to (5.8) approximately equal to the failure rate function $z(t)$ at time t multiplied by the length of the interval, Δt. □

If we put a large number of identical items into operation at time $t = 0$, then $z(t)\Delta t$ will roughly represent the relative proportion of the items still functioning at time t, that will fail in $(t, t + \Delta t]$.

Because

$$f(t) = \frac{d}{dt}F(t) = \frac{d}{dt}[1 - R(t)] = -R'(t),$$

then

$$z(t) = -\frac{R'(t)}{R(t)} = -\frac{d}{dt}\,\log R(t). \tag{5.9}$$

Because $R(0) = 1$, then

$$\int_0^t z(u)\,du = -\log R(t) \tag{5.10}$$

and

$$R(t) = e^{-\int_0^t z(u)\,du}. \tag{5.11}$$

The survivor function $R(t)$ and the distribution function $F(t) = 1 - R(t)$ are therefore uniquely determined by the failure rate function $z(t)$. From (5.6) and (5.11), the probability density function $f(t)$ can be written as

$$f(t) = z(t)\,e^{-\int_0^t z(u)\,du} \qquad \text{for } t > 0. \tag{5.12}$$

Some authors prefer the term *hazard rate* instead of failure rate, but because the term "failure rate" is so well established in applied reliability, we use this term even though we realize that this may lead to some confusion.

Remark 5.2 (The failure rate function versus ROCOF)
In actuarial statistics, the failure rate function is called the *force of mortality* (FOM). This term has been adopted by several authors of reliability textbooks to avoid the confusion between the failure rate function and the *rate of occurrence of failures* (ROCOF) of a repairable item. The failure rate function (FOM) is a function of the time-to-failure distribution of a single item and an indication of the "proneness to failure" of the item after time t has elapsed, whereas ROCOF is the occurrence rate of failures for a stochastic process; see Chapter 10. To be short, ROCOF is related to a counting process $N(t)$ that gives the cumulative number of failures from 0 to t and indicates at which speed this number is increasing or decreasing in average.

$$\text{ROCOF} = \frac{d}{dt}E[N(t)]. \tag{5.13}$$

For more details, see Ascher and Feingold (1984). □

The relationships between the functions $F(t), f(t)$, $R(t)$, and $z(t)$ are presented in Table 5.2.

The Bathtub Curve

The survivor function $R(t)$ is from (5.11) seen to be uniquely determined by the failure rate function $z(t)$. To determine the form of $z(t)$ for a given type of items, the following experiment may be carried out:

Put n identical and nonrepairable items into operation at time $t = 0$ and record the time each item fails. Assume that the last failure occurs at time $t_{\max}$. Split the

Table 5.2 Relationship between the functions $F(t), f(t), R(t)$, and $z(t)$.

Expressed by	$F(t)$	$f(t)$	$R(t)$	$z(t)$
$F(t) =$	–	$\int_0^t f(u)\,du$	$1 - R(t)$	$1 - e^{-\int_0^t z(u)\,du}$
$f(t) =$	$\frac{d}{dt}F(t)$	–	$-\frac{d}{dt}R(t)$	$z(t)\,e^{-\int_0^t z(u)\,du}$
$R(t) =$	$1 - F(t)$	$\int_t^\infty f(u)\,du$	–	$e^{-\int_0^t z(u)\,du}$
$z(t) =$	$\frac{dF(t)/dt}{1 - F(t)}$	$\frac{f(t)}{\int_t^\infty f(u)\,du}$	$-\frac{d}{dt}\log R(t)$	–

time axis into disjoint intervals of equal length Δt. Starting from $t = 0$, number the intervals as $j = 1, 2, \ldots$. For each interval record:

- The number of items, $n(j)$ that fail in interval j.
- The observed functioning times for the individual items $(t_{1j}, t_{2j}, \ldots, t_{nj})$ in interval j. Hence, t_{ij} is the time item i has been functioning in time interval j. t_{ij} is therefore equal to 0 if item j has failed before interval j, where $j = 1, 2, \ldots, m$.

Thus, $\sum_{i=1}^{n} t_{ij}$ is the total functioning time for the items in interval j. Now

$$z(i) = \frac{n(j)}{\sum_{i=1}^{n} t_{ij}}.$$

That is, the number of failures per unit functioning time in interval j. This is a natural estimate of the "failure rate" in interval j for the items that are functioning at the start of this interval.

Let $v(i)$ be the number of items that are functioning at the start of interval i. The failure rate in interval j is approximately

$$z(i) \approx \frac{n(i)}{v(i)\Delta t},$$

and hence,

$$z(i)\Delta t \approx \frac{n(i)}{v(i)}.$$

A histogram depicting $z(i)$ as a function of i typically is of the form in Figure 5.7. If n is large, we may use small time intervals. If we let $\Delta t \to 0$, is it expected that the step function $z(i)$ will tend toward a "smooth" curve, as shown in Figure 5.8, and is an estimate for the failure rate function $z(t)$.

This curve is usually called a *bathtub curve* after its characteristic shape. The failure rate is often high in the initial phase. This can be explained by the fact that there may be undiscovered defects in the items; these soon show up when

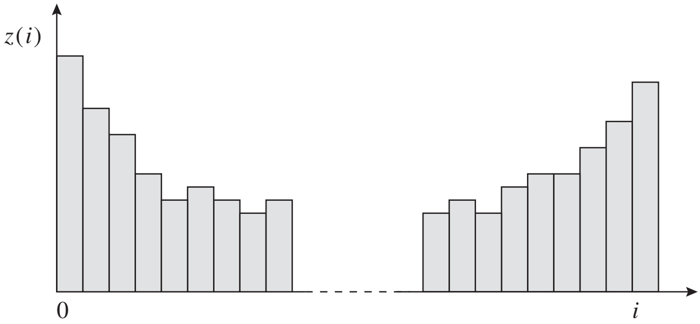


Figure 5.7 Empirical bathtub curve.

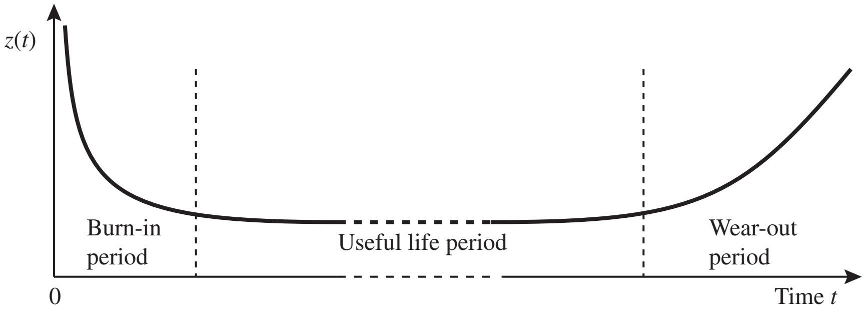


Figure 5.8 The bathtub curve.

the items are activated and the associated failures are called “infant mortality” failures. When the item has survived the “infant mortality” period, the failure rate often stabilizes at a level where it remains for a certain amount of time until it starts to increase as the items begin to wear out. From the shape of the bathtub curve, the time-to-failure of an item may be divided into three typical intervals: the *infant mortality* or *burn-in period*, the *useful life period*, and the *wear-out period*. The useful life period is also called the *chance failure period*. Sometimes, the items are tested at the factory before they are distributed to the users, and thus much of the “infant mortality” problems will be removed before the items are delivered for use. For the majority of mechanical items, the failure rate function will usually show a slightly increasing tendency in the useful life period.

Cumulative Failure Rate

The *cumulative failure rate* over $(0, t]$ is

$$Z(t) = \int_0^t z(u)\, du. \tag{5.14}$$

Equation (5.11) gives the following relationship between the survivor function $R(t)$ and $Z(t)$

$$R(t) = e^{-Z(t)} \quad \text{and} \quad Z(t) = -\log R(t). \tag{5.15}$$

The cumulative failure rate $Z(t)$ must satisfy

(1) $Z(0) = 0$

(2) $\lim_{t\to\infty} Z(t) = \infty$

(3) $Z(t)$ is a nondecreasing function of t.

Average Failure Rate

The *average failure rate* over the time interval (t_1, t_2) is

$$\overline{z}(t_1, t_2) = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} z(u)\, du = \frac{\log R(t_1) - \log R(t_2)}{t_2 - t_1}. \tag{5.16}$$

When the time interval is $(0, t)$, the average failure rate may be expressed as

$$\bar{z}(0,t) = \frac{1}{t}\int_0^t z(u)\,du = \frac{-\log R(t)}{t}. \tag{5.17}$$

Observe that this implies that

$$R(t) = e^{-\bar{z}(0,t)t}. \tag{5.18}$$

A Property of $z(t)$

Because $z(t) = -\frac{d}{dt}\log R(t)$, we have

$$\begin{aligned}\int_0^\infty z(t)\,dt &= -\int_0^\infty \frac{d[\log R(t)]}{dt}\,dt = -\int_0^\infty d\log R(t)\\ &= -\log R(t)\mid_0^\infty = \log R(0) - \log R(\infty) = \log 1 - \log 0 = \infty.\end{aligned} \tag{5.19}$$

The area under the failure rate curve is therefore infinitely large.

5.3.3 Conditional Survivor Function

The survivor function $R(t) = \Pr(T > t)$ was introduced under the assumption that the item was functioning at time $t = 0$. To make this assumption more visible, $R(t)$ may be written as

$$R(t \mid 0) = \Pr(T > t \mid T > 0).$$

Consider an item that is put into operation at time 0 and is still functioning at time x. The probability that the item of age x survives an *additional* interval of length t is

$$\begin{aligned}R(t \mid x) &= \Pr(T > t + x \mid T > x) = \frac{\Pr(T > t + x)}{\Pr(T > x)}\\ &= \frac{R(t+x)}{R(x)} \quad \text{for } 0 < x < t.\end{aligned} \tag{5.20}$$

$R(t \mid x)$ is called the *conditional survivor function* of the item at age x.

By using (5.12), $R(t \mid x)$ may be written

$$R(t \mid x) = \frac{R(t+x)}{R(x)} = \frac{e^{-\int_0^{t+x} z(u)\,du}}{e^{-\int_0^{x} z(u)\,du}} = e^{-\int_x^{t+x} z(u)\,du}. \tag{5.21}$$

The conditional probability density function $f(t \mid x)$ of an item that is still functioning at time x is

$$f(t \mid x) = -\frac{d}{dt}R(t \mid x) = -\frac{R'(t+x)}{R(x)} = \frac{f(t+x)}{R(x)}.$$

The associated failure rate function is

$$z(t \mid x) = \frac{f(t \mid x)}{R(t \mid x)} = \frac{f(t+x)}{R(t+x)} = z(t+x), \tag{5.22}$$

which is an obvious result because the failure rate function is a conditional rate, given that the item has survived up to the time where the rate is evaluated. This shows that when we have a failure rate function $z(t)$, such as for the bathtub curve in Figure 5.8, and consider the failure rate function for a used item of age x, we do not need any information about the form of $z(t)$ for $t \leq x$.

5.3.4 Mean Time-to-Failure

For the dataset in Table 5.1, the average time-to-failure is a metric for the central location of the failure times. It can be calculated empirically as the sum of the observed times-to-failure divided by the number n of failed items.

$$\bar{t} = \frac{1}{n} \sum_{i=1}^{n} t_i. \tag{5.23}$$

In probability theory, the *law of large numbers* says that if n tends to infinite, the empirical mean, $\bar{t}$, will stabilize around a constant value that does not depend on n any more. This value is called the expected, or mean value of T, and denoted $E(T)$. In reliability theory, it is named *MTTF*.

$$\text{MTTF} = E(T) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} t_i. \tag{5.24}$$

Law of Large Numbers

Let $X_1, X_2, \ldots$ be a sequence of independent random variables having a common distribution, and let $E(X_i) = \mu$. Then, with probability 1,

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \to \mu \quad \text{as } n \to \infty. \tag{5.25}$$

This definition is equivalent to the one in (5.26), which can be interpreted as a continuous version of the limit of the empirical mean: Each possible failure time t is multiplied by its frequency of occurrence $f(t)\,dt$ and the summation is replaced by an integral.

$$\text{MTTF} = E(T) = \int_0^\infty t f(t)\, dt. \tag{5.26}$$

Observe that MTTF only provides information about the central location of the failure times and no information about how failure times are dispersed around the *mean*. Therefore, the MTTF provides much less information than the histogram in Figure 5.2, but it gives useful input for a first screening and is very commonly used in reliability applications.

The MTTF can be derived from the other reliability metrics. Because $f(t) = -R'(t)$,

$$\text{MTTF} = -\int_0^\infty tR'(t)\,dt.$$

By partial integration

$$\text{MTTF} = -[tR(t)]_0^\infty + \int_0^\infty R(t)\,dt.$$

If MTTF $< \infty$, it can be shown that $[tR(t)]_0^\infty = 0$. In that case,

$$\text{MTTF} = \int_0^\infty R(t)\,dt. \tag{5.27}$$

It is often easier to determine MTTF by (5.27) than by (5.26).

Remark 5.3 (MTTF derived by Laplace transform)
The MTTF of an item may also be derived by using Laplace transforms. The Laplace transform of the survivor function $R(t)$ is (see Appendix B)

$$R^*(s) = \int_0^\infty R(t)\,e^{-st}\,dt. \tag{5.28}$$

When $s = 0$, we get

$$R^*(0) = \int_0^\infty R(t)\,dt = \text{MTTF}. \tag{5.29}$$

The MTTF may thus be derived from the Laplace transform $R^*(s)$ of the survivor function $R(t)$, by setting $s = 0$. □

5.3.5 Additional Probability Metrics

This section defines several additional metrics that may be used to describe a probability distribution.

Variance

The *variance* is related to the dispersion of the observed lifetimes around their mean value (see Chapter 12). The empirical variance is given by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(t_i - \bar{t})^2. \tag{5.30}$$

The empirical *standard deviation* is the square root of the variance

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(t_i - \bar{t})^2}. \tag{5.31}$$

The empirical variance indicates the average squared distance between the individual lifetimes of the dataset and the mean of the dataset. If n tends to infinite, this value converges to a constant called the *variance* defined as

$$\mathrm{var}(T) = \int_0^\infty [t - E(T)]^2 f(t)\, dt = E(T^2) - [E(T)]^2. \tag{5.32}$$

The associated *standard deviation* (SD) is defined as

$$\mathrm{SD}(T) = \sqrt{\mathrm{var}(T)}.$$

The variance and the standard deviation are not so much used in reliability, but they are often implicitly taken into account via the probability density function. We come back to the variance and the standard deviation in the sections dedicated to specific distributions. For further details, the reader may refer to Chapter 14.

Moments

The kth *moment* of T is defined as

$$\mu_k = E(T^k) = \int_0^\infty t^k f(t)\, dt = k\int_0^\infty t^{k-1} R(t)\, dt. \tag{5.33}$$

The first moment of T (i.e. for $k = 1$) is seen to be the mean of T.

Percentile Function

Because $F(t)$ is nondecreasing, the inverse function $F^{-1}(\cdot)$ exists and is called the *percentile function*.

$$F(t_p) = p \Rightarrow t_p = F^{-1}(p) \quad \text{for } 0 < p < 1, \tag{5.34}$$

where t_p is called the *p-percentile* of the distribution.

Median Lifetime

The MTTF is only one of several metrics of the “center” of a lifetime distribution. An alternative metric is the median lifetime t_m, defined by

$$R(t_m) = 0.50. \tag{5.35}$$

The median divides the distribution in two halves. The item will fail before time t_m with 50% probability, and will fail after time t_m with 50% probability. The median is the 0.50-percentile of the distribution.

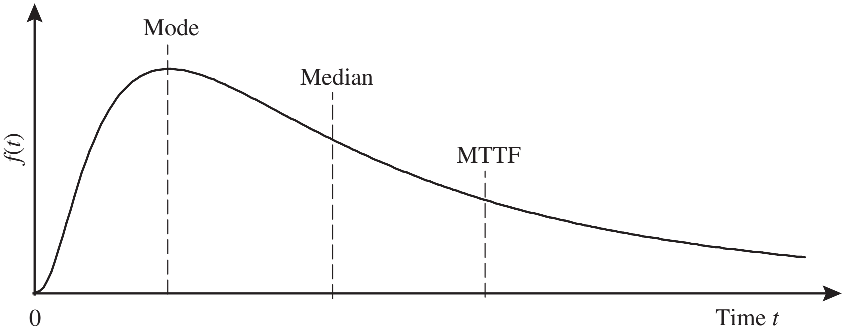


Figure 5.9 Location of the MTTF, the median lifetime, and the mode of a distribution.

Mode

The mode of a lifetime distribution is the most likely lifetime, that is, the time t_{mode} where the probability density function $f(t)$ attains its maximum.

$$f(t_{\text{mode}}) = \max_{0 \le t < \infty} f(t). \tag{5.36}$$

Figure 5.9 shows the location of the MTTF, the median lifetime t_m, and the mode t_{mode} for a distribution that is skewed to the right.

Example 5.1 Consider an item with survivor function

$$R(t) = \frac{1}{(0.2\ t + 1)^2} \qquad \text{for } t \ge 0,$$

where the time t is measured in months. The probability density function is

$$f(t) = -R'(t) = \frac{0.4}{(0.2\ t + 1)^3},$$

and the failure rate function is from (5.6)

$$z(t) = \frac{f(t)}{R(t)} = \frac{0.4}{0.2\ t + 1}.$$

The MTTF is from (5.27)

$$\text{MTTF} = \int_0^\infty R(t)\ dt = 5 \text{ mo}.$$

The functions $R(t), f(t)$, and $z(t)$ are shown in Figure 5.10. □

Additional metrics are discussed in Chapter 14.

5.3.6 Mean Residual Lifetime

Consider an item that is put into operation at time $t = 0$ and is still functioning at time x. The item fails at the random time T. The residual lifetime of the item,

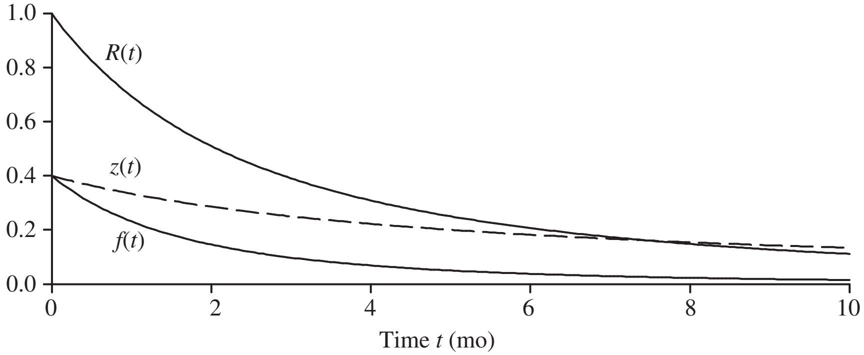


Figure 5.10 The survivor function $R(t)$, the probability density function $f(t)$, and the failure rate function $z(t)$ (dashed line) in Example 5.1.

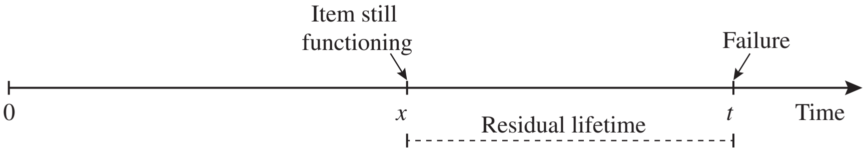


Figure 5.11 The residual lifetime of an item that is still functioning at time x.

when it is known that the item is still functioning at time x, is $T - x$, as shown in Figure 5.11.

The *mean residual* (or, *remaining) lifetime*, MRL(x), of the item at age x is

$$\text{MRL}(x) = E(T - x \mid T > x),$$

that is, the mean of the random variable $T - x$ when it is known that $T > x$. The mean value can be determined from the conditional survivor function in (5.27) as follows:

$$\text{MRL}(x) = \mu(x) = \int_x^\infty R(t \mid x)\, dt = \frac{1}{R(x)} \int_x^\infty R(t)\, dt. \tag{5.37}$$

Observe that MRL(x) is the *additional* MTTF, that is, the mean remaining lifetime of an item that has reached the age x. This means that when the item has reached age x, its mean age at failure is $x +$ MRL(x).

Also observe that MRL(x) applies to a *general* item that has reached the age x. We do not have access to any additional information about the particular item or its history in the interval $(0, x)$. Our knowledge about a possible degradation of the item is the same at age x as it was when the item was put into operation at time $t = 0$.

At time $t = 0$, the item is new, and we have $\mu(0) = \mu = \text{MTTF}$. It is sometimes of interest to study the function

$$g(x) = \frac{\text{MRL}(x)}{\text{MTTF}} = \frac{\mu(x)}{\mu}. \tag{5.38}$$

When an item has survived up to time x, then $g(x)$ gives the MRL(x) as a percentage of the initial MTTF. If, for example, $g(x) = 0.60$, then the mean residual lifetime, MRL(x) at time x, is 60% of the MRL at time 0.

Remark 5.4 (Remaining useful lifetime)
A concept similar to MRL(x) is the (mean of the) *remaining useful lifetime* (RUL) at age x. The main difference is that RUL(x) applies for a particular item, where we have access to performance and maintenance data from the period $(0, x)$ and/or information about possible changes in the future operational context. RUL is further discussed in Chapter 12. □

Example 5.2 (Mean residual lifetime)
Consider an item with failure rate function $z(t) = t/(t+1)$. The failure rate function is increasing and approaches 1 when $t \to \infty$. The corresponding survivor function is

$$R(t) = e^{-\int_0^t u/(u+1)\ du} = (t+1)\ e^{-t},$$

and the MTTF is

$$\text{MTTF} = \int_0^\infty (t+1)\ e^{-t}\ dt = 2.$$

The conditional survivor function is

$$R(t \mid x) = \Pr(T > t \mid T > x) = \frac{(t+1)\ e^{-t}}{(x+1)\ e^{-x}} = \frac{t+1}{x+1}\ e^{-(t-x)}.$$

The MRL is

$$\begin{aligned}
\text{MRL}(t) &= \int_x^\infty R(x \mid t)\ dx = \int_x^\infty \frac{t+1}{x+1}\ e^{-(t-x)}\ dt \\
&= \int_x^\infty \left(1 + \frac{t-x}{x+1}\right) e^{-(t-x)}\ dt \\
&= \int_x^\infty e^{-(t-x)}\ dt + \frac{1}{x+1}\int_x^\infty (t-x)e^{-(t-x)}\ dt \\
&= 1 + \frac{1}{x+1}.
\end{aligned}$$

Observe that MRL(x) is equal to 2 (=MTTF) when $x = 0$, that MRL(x) is a decreasing function of x, and that MRL(x) → 1 when $x \to \infty$. This means that the function $g(x)$ in (5.38) approaches 0.5 when x increases. The survivor functions

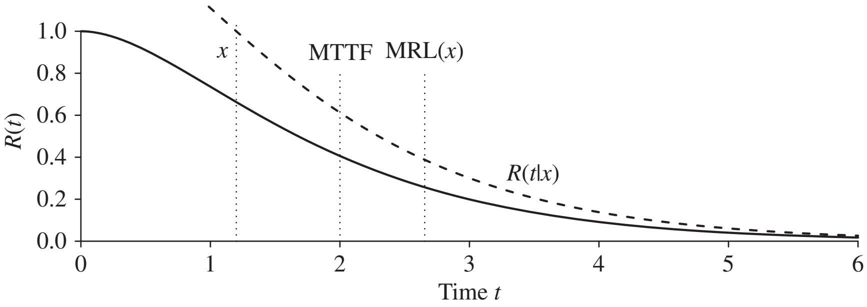


Figure 5.12 The survivor function $R(t)$ (fully drawn line), the conditional survivor function $R(t \mid x)$ for $x = 1.2$ (dashed line) together with the values of MTTF and MRL(x) in Example 5.2.

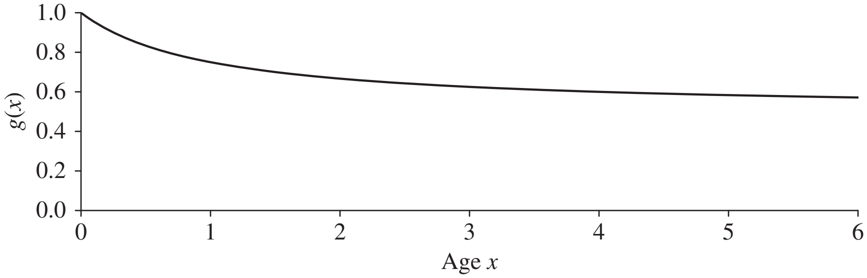


Figure 5.13 The $g(x)$ function (5.38) in Example 5.2.

and the MRL(x) are shown in Figure 5.12, whereas the $g(x)$ function is shown in Figure 5.13. □

5.3.7 Mixture of Time-to-Failure Distributions

Assume that the same type of items are produced at two different plants. The items are assumed to be independent with failure rate functions $z_1(t)$ and $z_2(t)$, respectively. The production process is slightly different at the two plants, and the items will therefore have different failure rates. Let $R_1(t)$ and $R_2(t)$ be the survivor functions associated with $z_1(t)$ and $z_2(t)$, respectively. The items are mixed up before they are sold. A fraction p is coming from plant 1, and the rest $(1 - p)$ is coming from plant 2.

If we pick one item at random, the survivor function for this item is

$$R(t) = p\, R_1(t) + (1 - p)\, R_2(t), \tag{5.39}$$

and the probability density function of the life distribution is

$$f(t) = -R'(t) = p\,f_1(t) + (1-p)\,f_2(t). \tag{5.40}$$

The failure rate function of the item is

$$\begin{aligned} z(t) &= \frac{f(t)}{R(t)} = \frac{p\,f_1(t) + (1-p)\,f_2(t)}{p\,R_1(t) + (1-p)\,R_2(t)} \\ &= \frac{p\,R_1(t)}{p\,R_1(t) + (1-p)\,R_2(t)} \left(\frac{f_1(t)}{R_1(t)}\right) + \frac{(1-p)\,R_1(t)}{p\,R_1(t) + (1-p)\,R_2(t)} \left(\frac{f_2(t)}{R_2(t)}\right). \end{aligned}$$

By introducing the factor

$$a_p(t) = \frac{p\,R_1(t)}{p\,R_1(t) + (1-p)\,R_2(t)}, \tag{5.41}$$

we can write the failure rate function as (by remembering that $z_i(t) = f_i(t)/R_i(t)$ for $i = 1, 2$)

$$z(t) = a_p(t)\,z_1(t) + [1 - a_p(t)]z_2(t). \tag{5.42}$$

The failure rate of the item chosen at random is therefore a weighted average of the two failure rates $z_1(t)$ and $z_2(t)$, but the weighing factor varies with the time t.

More details about life distributions are given by Rinne (2014) and O'Connor et al. (2016).

5.4 Some Time-to-Failure Distributions

This section introduces a number of parametric time-to-failure distributions:

(1) The exponential distribution
(2) The gamma distribution
(3) The Weibull distribution
(4) The normal (Gaussian) distribution
(5) The lognormal distribution.

In addition, an introduction to distributions with covariates and extreme value distributions is given.

5.4.1 The Exponential Distribution

Consider an item that is put into operation at time $t = 0$. The time-to-failure T of the item has *probability density function*

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{for } t > 0,\ \lambda > 0 \\ 0 & \text{otherwise} \end{cases}. \tag{5.43}$$

This distribution is called the *exponential distribution* with parameter λ, and we write $T \sim \exp(\lambda)$.

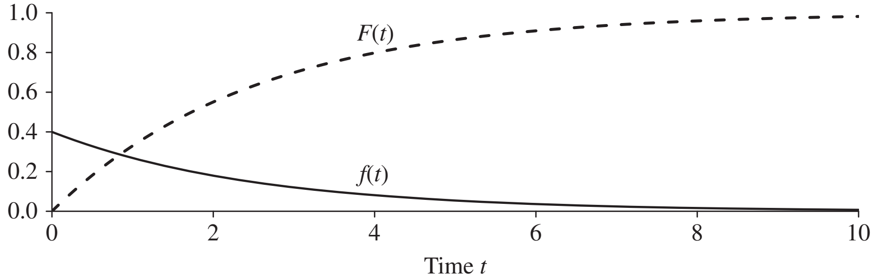


Figure 5.14 Probability density function $f(t)$ (fully drawn line) and distribution function $F(t)$ (dashed line) for the exponential distribution ($\lambda = 0.4$).

Survivor Function

The *survivor function* of the item is

$$R(t) = \Pr(T > t) = \int_t^\infty f(u)\, du = e^{-\lambda t} \quad \text{for } t > 0. \tag{5.44}$$

The probability density function $f(t)$ and the survivor function $R(t)$ for the exponential distribution are shown in Figure 5.14.

MTTF

The MTTF is

$$\text{MTTF} = \int_0^\infty R(t)\, dt = \int_0^\infty e^{-\lambda t}\, dt = \frac{1}{\lambda}, \tag{5.45}$$

and the *variance* of T is

$$\text{var}(T) = \frac{1}{\lambda^2}. \tag{5.46}$$

Observe that when the MTTF increases (or is reduced), the variance does the same. This is a limitation of the exponential distribution and makes it impossible to adapt independently the mean and the variance to fit a historical dataset.

The probability that an item survives its MTTF is

$$R(\text{ MTTF }) = R\left(\frac{1}{\lambda}\right) = e^{-1} \approx 0.3679 \quad \text{for all values of } \lambda.$$

Any item with exponential time-to-failure distribution will survive its MTTF with a probability that is approximately 36.8%

Failure Rate Function

The failure rate function is

$$z(t) = \frac{f(t)}{R(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda. \tag{5.47}$$

Hence, an item with exponential time-to-failure distribution has a failure rate function that is constant and independent of time. Because there is a one-to-one correspondence between the distribution and the failure rate function, any item with constant failure rate has an exponential time-to-failure distribution.

Figure 5.8 indicates that the exponential distribution may be a realistic time-to-failure distribution for an item during its useful life period, at least for certain types of items.

The results (5.45) and (5.47) compare well with the use of the concepts in everyday language. If an item on the average has $\lambda = 4$ failures/yr, the MTTF of the item is 1/4 year.

The corresponding cumulative failure rate function is $Z(t) = \lambda t$ and may be drawn as a straight line with slope λ.

Median Time-to-Failure

The median time-to-failure of the exponential distribution is determined from $R(t_m) = 0.50$ and is

$$t_m = \frac{\log 2}{\lambda} \approx \frac{0.693}{\lambda} = 0.693 \text{ MTTF}. \tag{5.48}$$

This means that for an item with constant failure rate, it is a 50% probability that the item will fail before it has reached 69.3% of its MTTF.

Changed Time Scale

Consider an item with time-to-failure, $T \sim \exp(\lambda)$. Assume that the time unit for measuring T is changed, for example, that we measure time in days instead of hours. This change of scale may be expressed by $T_1 = aT$, for some constant a. The survivor function of the time-to-failure T_1 in the new time scale is

$$R_1(t) = \Pr(T_1 > t) = \Pr(aT > t) = \Pr(T > t/a) = e^{-\lambda t/a}.$$

This means that $T_1 \sim \exp(\lambda/a)$ with MTTF

$$\text{MTTF}_1 = \frac{a}{\lambda} = a \text{ MTTF},$$

which is an obvious result. This shows that the exponential distribution is closed under change of scale, that is,

$$T \sim \exp(\lambda) \quad \Rightarrow \quad aT \sim \exp(\lambda/a) \quad \text{for all constants } a > 0. \tag{5.49}$$

Probability of Failure in a Short Time Interval

The Maclaurin series[2] of the exponential function is

$$e^{-\lambda t} = \sum_{x=0}^{\infty} \frac{(-\lambda t)^x}{x!} = 1 - \lambda t + \frac{(\lambda t)^2}{2} - \frac{(\lambda t)^3}{6} + \cdots .$$

[2] Named after the Scottish mathematician Colin Maclaurin (1698–1746). The Maclaurin series is a special case of a Taylor series.

When λt is small, $(\lambda t)^x$ for $x = 2, 3, \dots$ is negligible, and we may use the approximation

$$e^{-\lambda t} \approx 1 - \lambda t \quad \text{when } \lambda t \text{ is small.} \tag{5.50}$$

Consider a short time interval $(t, t + \Delta t]$. The probability that an item with time-to-failure $T \sim \exp(\lambda)$ fails in this interval is

$$\begin{aligned} \Pr(t < T \le t + \Delta t) &= \Pr(T \le t + \Delta t) - \Pr(T \le t) = 1 - e^{-\lambda(t+\Delta t)} - (1 - e^{-\lambda t}) \\ &= e^{-\lambda t} - e^{-\lambda(t+\Delta t)} \approx \lambda \Delta t. \end{aligned} \tag{5.51}$$

The probability that at an item with constant failure rate λ fails in a short time interval of length Δt is approximately $\lambda \Delta t$. The approximation is sufficiently accurate when Δt is very small.

Series Structure of Independent Components

Consider a series structure of n independent components with constant failure rates $\lambda_1, \lambda_2, \dots, \lambda_n$. A series structure fails when the first component failure occurs such that time-to-failure T_s of the series structure is

$$T_s = \min\{T_1, T_2, \dots, T_n\} = \min_{i=1.2,\dots,n} T_i.$$

The survivor function of the series structure is

$$\begin{aligned} R_s(t) &= \Pr(T_s > t) = \Pr\left(\min_{i=1.2,\dots,n} T_i > t\right) = \Pr\left(\bigcap_{i=1}^{n} T_i > t\right) \\ &= \Pr[(T_1 > t) \cap (T_2 > t) \cap \cdots \cap (T_n > t)] = \prod_{i=1}^{n} \Pr(T_i > t) \\ &= \prod_{i=1}^{n} e^{-\lambda_i t} = e^{-\left(\sum\limits_{i=1}^{n} \lambda_i\right)\, t} = e^{-\lambda_s t}, \end{aligned} \tag{5.52}$$

where $\lambda_s = \sum_{i=1}^{n} \lambda_i$. This shows that the time-to-failure, T_s, of the series structure is exponentially distributed with failure rate $\lambda_s = \sum_{i=1}^{n} \lambda_i$.

For the special case when the n independent components are identical, such that $\lambda_i = \lambda$ for $i = 1, 2, \dots, n$, the time-to-failure of the series structure is exponentially distributed with failure rate $\lambda_s = n\lambda$. The MTTF of the series structure is

$$\text{MTTF}_s = \frac{1}{\lambda_s} = \frac{1}{n}\,\frac{1}{\lambda},$$

that is the MTTF of the series structure is equal to the MTTF of a single component divided by the number of components in the structure.

Conditional Survivor Function and Mean Residual Lifetime

The conditional survivor function of an item with time-to-failure $T \sim \exp(\lambda)$ is

$$\begin{aligned} R(t \mid xt) &= \Pr(T > t + x \mid T > xt) = \frac{\Pr(T > t + x)}{\Pr(T > xt)} \\ &= \frac{e^{-\lambda(t+x)}}{e^{-\lambda xt}} = e^{-\lambda tx} = \Pr(T > t) = R(tx). \end{aligned} \tag{5.53}$$

The survivor function of an item that has been functioning for x time units, is therefore equal to the survivor function of a new item. A new item and a used item (that is still functioning), therefore, have the same probability of surviving a time interval of length t. The MRL, for the exponential distribution is therefore

$$\text{MRL}(xt) = \int_0^\infty R(t \mid xt)\, dtx = \int_0^\infty R(tx)\, dtx = \text{MTTF}.$$

The mean residual lifetime, MRL(t), of an item with exponential time-to-failure distribution is hence equal to its MTTF irrespective of the age xt of the item. The item is therefore *as-good-as-new* as long as it is functioning, and we often say that the exponential distribution has *no memory*.

Assuming an exponentially distributed time-to-failure implies that

- A used item is stochastically as-good-as-new, so there is no reason to replace a functioning item.
- For the estimation of the survivor function, the MTTF, and so on, it is sufficient to collect data on the number of hours of observed time in operation and the number of failures. The age of the items is of no interest in this connection.

The exponential distribution is the most commonly used time-to-failure distribution in applied reliability analysis. The reason for this is its mathematical simplicity and that it leads to realistic time-to-failure models for certain types of items.

The Difference Between a Random Variable and a Parameter

A stochastic experiment is usually carried out in order to observe and measure one or more random variables, such as the time-to-failure T. Observing T gives a number, such as 5000 hours. Identical experiments lead to different numbers. The variation, or uncertainty, in these numbers can be described by a statistical distribution $F(t)$. As a basis for the experiment, the distribution is usually not specified fully, but depends on one or more variables known

(Continued)

(Continued)

as *parameters*. Parameters are often represented in the distribution by Greek letters. An example is the parameter λ of the exponential distribution.

A parameter in statistics is a variable that cannot be measured directly from an experiment, but needs to be estimated based on observed values (numbers) of the random variable. After an experiment, we measure the random variable, but *estimate* the parameter. Different experiments will usually give slightly different estimates of the parameter. The rule, or formula, used to estimate a parameter is called an *estimator* of the parameter and can be assessed by its mean value and its standard deviation. Estimators are discussed further in Chapter 14.

Example 5.3 (Rotary pump)

A rotary pump has a constant failure rate $\lambda = 4.28 \times 10^{-4}\ \text{h}^{-1}$. The probability that the pump survives one month ($t = 730$ hours) in continuous operation is

$$R(t) = e^{-\lambda t} = e^{-4.28\times 10^{-4}\cdot 730} \approx 0.732.$$

The MTTF is

$$\text{MTTF} = \frac{1}{\lambda} = \frac{1}{4.28 \times 10^{-4}}\ \text{h} \approx 2336\ \text{h} \approx 3.2\ \text{mo}.$$

Suppose that the pump has been functioning without failure during its first two months ($t_1 = 1460$ hours) in operation. The probability that the pump will fail during the next month ($t_2 = 730$ hours) is

$$\Pr(T \le t_1 + t_2 \mid T > t_1) = \Pr(T \le t_2) = 1 - e^{-4.28\times 10^{-4}\cdot 730} \approx 0.268.$$

because the pump is as-good-as-new when we know that it is still functioning at time t_1. □

Example 5.4 (Probability of one item failing before the other)

Consider a structure of two independent components with failure rates λ_1 and λ_2, respectively. The probability that component 1 fails before component 2 is

$$\begin{aligned}\Pr(T_2 > T_1) &= \int_0^\infty \Pr(T_2 > t \mid T_1 = t) f_{T_1}(t)\, dt \\ &= \int_0^\infty e^{-\lambda_2 t}\ \lambda_1 e^{-\lambda_1 t}\, dt \\ &= \lambda_1 \int_0^\infty e^{-(\lambda_1+\lambda_2)t}\, dt = \frac{\lambda_1}{\lambda_1 + \lambda_2}.\end{aligned}$$

This result can easily be generalized to a structure of n independent components with failure rates $\lambda_1, \dots, \lambda_n$. The probability that component j is the first component to fail is

$$\Pr(\text{ Component } j \text{ fails first}) = \frac{\lambda_j}{\sum_{i=1}^{n} \lambda_i}. \tag{5.54}$$

□

Mixture of Exponential Distributions

Assume that the same type of items is produced at two different plants. The items are assumed to be independent and have constant failure rates. The production process is slightly different at the two plants, and the items will therefore have different failure rates. Let λ_i be the failure rate of the items coming from plant i, for $i = 1, 2$. The items are mixed up before they are sold. A fraction p is coming from plant 1, and the rest $(1 - p)$ is coming from plant 2. If we pick one item at random, the survivor function of this item is

$$R(t) = pR_1(t) + (1 - p)R_2(t) = p\, e^{-\lambda_1 t} + (1 - p)\, e^{-\lambda_2 t}.$$

The MTTF is

$$\text{MTTF} = \frac{p}{\lambda_1} + \frac{1 - p}{\lambda_2},$$

and the failure rate function is

$$z(t) = \frac{p\lambda_1\, e^{-\lambda_1 t} + (1 - p)\lambda_2\, e^{-\lambda_2 t}}{p\, e^{-\lambda_1 t} + (1 - p)\, e^{-\lambda_2 t}}. \tag{5.55}$$

The failure rate function, which is shown in Figure 5.15, is seen to be decreasing. If we assume that $\lambda_1 > \lambda_2$, early failures should have a failure rate close to λ_1. After a while, all the "weak" components have failed, and we are left with components with a lower failure rate λ_2.

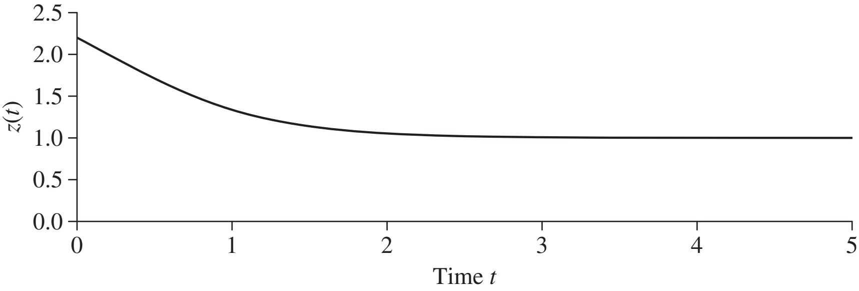


Figure 5.15 The failure rate function of the mixture of two exponential distributions ($\lambda_1 = 1$, $\lambda_2 = 3$, and $p = 0.4$).

Stepwise Constant Failure Rate

Consider an item that is running in distinct intervals only. When not running, it remains in a standby mode that may be energized or nonenergized. An example of such an item is a household heat-pump.[3] When the room temperature is low, the heat-pump is started on demand from a thermostat and when the room temperature is high, the heat-pump is stopped and enters a standby mode. The item may fail to start (on demand) with a probability p. When running, it has a constant failure rate λ_r and in standby mode it has a constant failure rate λ_s. The failure rate function $z(t)$ of the item becomes as shown in Figure 5.16.

If we can record the number n of start demands per time unit (e.g. per week) and the fraction ν of time the item is running, we may calculate an average failure rate λ_t of the item as

$$\lambda_t = \lambda_d + \nu\lambda_r + (1-\nu)\lambda_s, \tag{5.56}$$

where $\lambda_d = np$ is the number of start failures per time unit.

5.4.2 The Gamma Distribution

The time-to-failure T of an item is said to be *gamma distributed* when its probability density function is

$$f(t) = \frac{\lambda}{\Gamma(\alpha)}(\lambda t)^{\alpha-1}\, e^{-\lambda t} \qquad \text{for } t > 0, \tag{5.57}$$

where $\Gamma(\cdot)$ is the *gamma function*, $\alpha > 0$ and $\lambda > 0$ are parameters, and t is the time. The gamma distribution is often written $T \sim \text{gamma}(\alpha, \lambda)$. The probability density function $f(t)$ is sketched in Figure 5.17 for selected values of α. The gamma distribution is not a widely used time-to-failure distribution, but is considered to

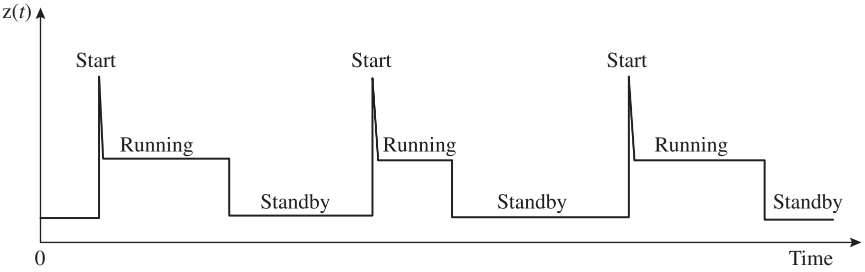


Figure 5.16 The failure rate function of an item with stepwise constant failure rates and start problems.

[3] This example is inspired by a similar example found on the Internet, unfortunately without any author's name or any other references.

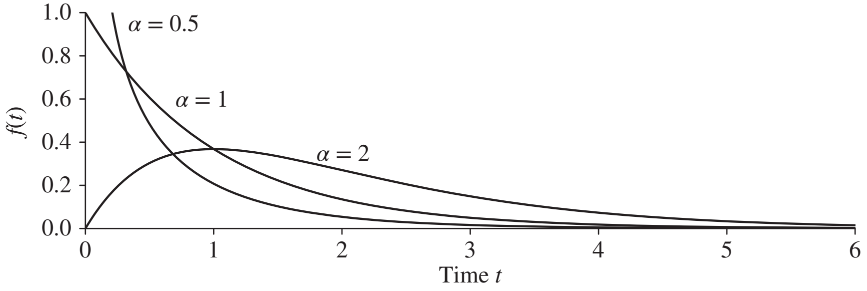


Figure 5.17 The gamma probability density for selected values of α, $\lambda = 1.0$.

be adequate in cases where partial failures can exist and where a specific number of partial failures must occur before the item fails. In spite of this limited usage, the gamma distribution is an important distribution in reliability because it is used in other situations as illustrated later in this book (e.g. see Chapter 15).

The gamma function is available in R by the command `gamma(x)`, for example, `gamma(2.7)= 1.544686`. In R, the parameter α is called `shape` and λ is called `rate`. We may alternatively use the parameter $\theta = 1/\lambda$, which is called `scale` in R, as the second parameter. The probability density functions (e.g. for $\alpha = 2$ and $\lambda = 1$ can be plotted by the R script:

```
t<-seq(0,6,length=300)  # Set time axis
# Set the parameters
a<-2   # shape
rate<-1   # rate
# Calculate the gamma density (y) for each t
y<-dgamma(t,a,rate,log=F)
plot(t,y,type="l")
```

Observe that we have to write `rate`= to specify λ in the script. We could, alternatively, have written `scale`= to specify the scale parameter $\theta(= 1/\lambda)$. The `scale` parameter is the default parameter in R and if we write only the number, it is interpreted as `scale`.

From (5.57) we find that

$$\text{MTTF} = \frac{\alpha}{\lambda} = \alpha\theta. \tag{5.58}$$

$$\text{var(T)} = \frac{\alpha}{\lambda^2} = \alpha\theta^2. \tag{5.59}$$

The parameter α is a dimensionless number, whereas θ is measured in time units (e.g. hours). For a specified value of α, the MTTF is proportional to θ.

The distribution function $F(t)$ is available in R by the command `pgamma` and $R(t)$ is than obtained as `1-pgamma`. The survivor function $R(t)$ (e.g. for $\alpha = 2$ and $\lambda = 1$) can be plotted by the R script:

```
t<-seq(0,6,length=300)   # Set time axis
# Set the parameter
a<-2    # shape
rate <- 1 #rate
# Calculate the survivor function (y) for each t
y<-1-pgamma(t,a,rate,log=F)
plot(t,y,type="l")
```

A sketch of $R(t)$ is given in Figure 5.18 for some values of α.

The failure rate function (e.g. for $\alpha = 2$ and $\lambda = 1$) may be calculated and plotted by the R script

```
t<-seq(0, 6, length=300)  # Set time axis
# Set the parameter
a<-2    # shape
rate <- 1
# Calculate the failure rate function (y) for each t
y<-dgamma(t,a,rate,log=F)/(1-pgamma(t,a,rate,log=F))
plot(t,y,type="l")
```

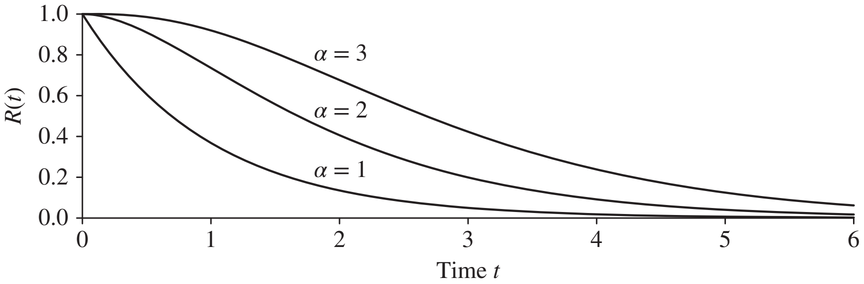


Figure 5.18 Survivor function for the gamma distribution for selected values of α, $\lambda = 1.0$.

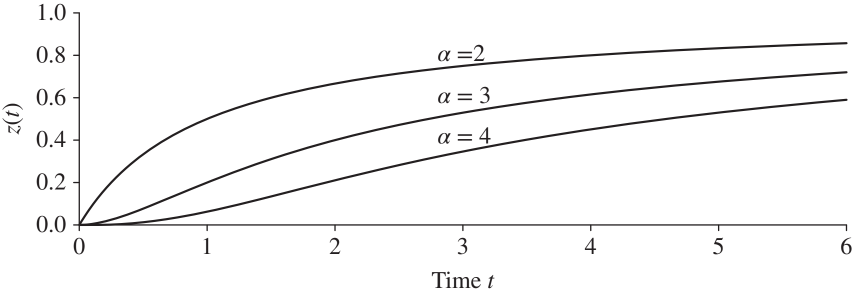


Figure 5.19 Failure rate function of the gamma distribution for selected values of α, $\lambda = 1$.

The“behavior” of the failure rate function can now be studied by running the above script for various values of α. Observe that

$$\begin{array}{lll} \text{for } 0 < \alpha < 1, & z(t) \rightarrow \infty & \text{when } t \rightarrow 0 \\ \text{for } \alpha > 1, & z(t) \rightarrow 0 & \text{when } t \rightarrow 0 \end{array}.$$

The function $z(t)$ is hence not continuous as a function of the shape parameter α for $\alpha = 1$. We must therefore be careful when specifying α near 1.

The failure rate function $z(t)$ is shown in Figure 5.19 for some integer values of α.

Let T_1 and T_2 be independent and gamma distributed (α_1, λ) and (α_2, λ), respectively. It is then easy to show (see Problem 5.13) that $T_1 + T_2$ is gamma distributed with parameters $(\alpha_1 + \alpha_2, \lambda)$. Gamma distributions with a common λ are therefore *closed under addition*.

For integer values of α, the gamma distribution can be deduced from the homogeneous Poisson process (HPP), as shown in Section 5.8.5.

Special Cases

For special values of the parameters α and λ, the gamma distribution is known under other names:

(1) When $\alpha = 1$, we have the *exponential distribution* with failure rate λ.
(2) When $\alpha = n/2$, n is an integer, and $\lambda = 1/2$, the gamma distribution coincides with the well-known *chi-square* (χ^2) *distribution* with n degrees of freedom.
(3) When α is an integer, the gamma distribution is called an *Erlangian distribution* with parameters α and λ.

The χ^2 Distribution

The χ^2 distribution is a very important distribution in many branches of statistics. A main feature is its relation to the standard normal distribution $\mathcal{N}(0, 1)$. If $U_1, U_2, \dots, U_n$ are independent and standard normal variables, $X = \sum_{i=1}^{n} U_i^2$ is a χ^2

distributed variable with *n degrees of freedom*, with probability density function

$$f_n(x) = \frac{1}{\Gamma(n/2)2^{n/2}} x^{n/2-1} e^{-x/2} \qquad \text{for } x > 0.$$

The χ^2 distribution has mean $E(X) = n$ and variance $\text{var}(X) = 2n$. The χ^2 distribution is not a relevant time-to-failure distribution, but is important in some data-analyses. The χ^2 distribution is available in R where, for example, the density of the χ^2 distribution with `df` degrees of freedom is calculated by the command `dchisq(x,df,log=F)`.

Example 5.5 (Mixture of exponential distributions)

This example illustrates another application of the gamma distribution. Assume that items of a specific type are produced in a plant where the production process is unstable such that the failure rate λ of the items varies with time. If we pick an item at random, the conditional probability density function of the time-to-failure T, given λ, is

$$f(t \mid \lambda) = \lambda e^{-\lambda t} \qquad \text{for } t > 0.$$

Assume that the variation in λ can be modeled by assuming that the failure rate is a random variable Λ that is gamma distributed with parameters k and α. The probability density function of Λ is

$$\pi(\lambda) = \frac{\alpha^k}{\Gamma(k)} \lambda^{k-1} \, e^{-\alpha\lambda} \qquad \text{for } \lambda > 0, \;\; \alpha > 0, \;\; k > 0.$$

The unconditional probability density of T is thus

$$f(t) = \int_0^\infty f(t \mid \lambda)\pi(\lambda) \, d\lambda = \frac{k\alpha^k}{(\alpha + t)^{k+1}}. \tag{5.60}$$

The survivor function is

$$R(t) = \Pr(T > t) = \int_t^\infty f(u) \, du = \frac{\alpha^k}{(\alpha + t)^k} = \left(1 + \frac{t}{\alpha}\right)^{-k}. \tag{5.61}$$

The MTTF is

$$\text{MTTF} = \int_0^\infty R(t) \, dt = \frac{\alpha}{k-1} \qquad \text{for } k > 1.$$

Observe that MTTF does not exist for $0 < k \leq 1$. The failure rate function is

$$z(t) = \frac{f(t)}{R(t)} = \frac{k}{\alpha + t}, \tag{5.62}$$

and hence is monotonically *decreasing* as a function of t. This may be illustrated by the following case:

A factory is producing a specific type of gas detectors. Experience has shown that the *mean* failure rate of the detectors is $\lambda_m = 1.15 \times 10^{-5}$ h^{-1}. The corresponding mean MTTF is $1/\lambda_m \approx 9.93$ years, but the production is unstable and the standard deviation of the failure rate is estimated to be 4×10^{-6} h^{-1}. As above, we assume that the failure rate is a random variable Λ with a gamma(k, α) distribution. From (5.59), we have $E(\Lambda) = k/\alpha = 1.15 \times 10^{-5}$, and $\text{var}(\Lambda) = k/\alpha^2 = [4 \times 10^{-6}]^2$. We can now solve for k and α and get

$$k \approx 8.27 \qquad \text{and} \qquad \alpha \approx 7.19 \times 10^6.$$

The MTTF is then

$$\text{MTTF} = \frac{\alpha}{k-1} \approx 9.9 \times 10^5 \text{ h} \approx 11.3 \text{ yr}.$$

The corresponding failure rate function $z(t)$ may be found from (5.62). Similar examples are discussed in Chapter 15. □

Remark 5.5 (Mixed distributions)
Example 5.5 is similar to the situation illustrated in Figure 5.15, where we by mixing two different exponential distributions got a decreasing failure rate (DFR) function. The results from these examples are very important for collection and analysis of field data. Suppose that the failure rate of a specific item is equal to λ. When we collect data from different installations and from different operational contexts, the failure rate λ will vary. If we pool all the data into one single dataset and analyze the data, we conclude that the failure rate function is decreasing. □

5.4.3 The Weibull Distribution

The Weibull distribution is one of the most widely used time-to-failure distributions in reliability analysis. The distribution is named after the Swedish professor Waloddi Weibull (1887–1979), who developed the distribution for modeling the strength of materials. The Weibull distribution is very flexible, and can, through an appropriate choice of parameters, model many types of failure rate behaviors.

Two-Parameter Weibull Distribution

The time-to-failure T of an item is said to be *Weibull distributed* with parameters $\alpha(> 0)$ and $\theta(> 0)$ if the distribution function is given by

$$F(t) = \Pr(T \le t) = \begin{cases} 1 - e^{-\left(\frac{t}{\theta}\right)^\alpha} & \text{for } t > 0 \\ 0 & \text{otherwise} \end{cases}. \tag{5.63}$$

The two-parameter Weibull distribution is often written as $T \sim \text{Weibull}(\alpha, \theta)$. The corresponding probability density is

$$f(t) = \frac{d}{dt}F(t) = \frac{\alpha}{\theta}\left(\frac{t}{\theta}\right)^{\alpha-1} e^{-\left(\frac{t}{\theta}\right)^\alpha} \qquad \text{for } t > 0, \tag{5.64}$$

where θ is a *scale* parameter measured in time units, and α is a dimensionless constant called the *shape* parameter. Observe that when $\alpha = 1$, the Weibull distribution is equal to the exponential distribution with $\lambda = 1/\theta$.

Remark 5.6 (Choice of parameters)
The parameters in (5.64) are chosen because this is the default parameterization in R. Many authors prefer instead the parameters α and λ ($= 1/\theta$), in which case the distribution function is written $F(t) = 1 - e^{-(\lambda t)^\alpha}$. This way, the special case for $\alpha = 1$ directly becomes the exponential distribution $\exp(\lambda)$. Both parameterizations give the same results, and it is therefore a matter of habit and convenience which one to use. Later in this book, you will see both versions, and we hope this will not be too confusing. Both θ and λ are referred to as *scale* parameters. □

A plot of the probability density function (`dweibull`) of the Weibull distribution with shape parameter $\alpha = 2.5$ and scale parameter $\theta = 300$ is, for example, obtained by the following R script.

```
t<-seq(0,1000,length=300)  # Set time axis
# Set the parameters
a<-2.5   # shape parameter (alpha)
th<-200   # scale parameter (theta)
# Calculate the Weibull density (y) for each t
y<-dweibull(t,a,th,log=F)
plot(t, y, type="l")
```

The probability density function $f(t)$ is shown in Figure 5.20 for selected values of α.

Survivor Function

The *survivor function* of $T \sim \text{Weibull}(\alpha, \theta)$ is

$$R(t) = \Pr(T > 0) = e^{-\left(\frac{t}{\theta}\right)^\alpha} \qquad \text{for } t > 0. \tag{5.65}$$

Failure Rate Function

The *failure rate function* of $T \sim \text{Weibull}(\alpha, \theta)$ is

$$z(t) = \frac{f(t)}{R(t)} = \frac{\alpha}{\theta}\left(\frac{t}{\theta}\right)^{\alpha-1} \qquad \text{for } t > 0. \tag{5.66}$$

Observe that the failure rate may be written as

$$z(t) = \alpha\ \theta^{-\alpha}\ t^{\alpha-1} \qquad \text{for } t > 0.$$

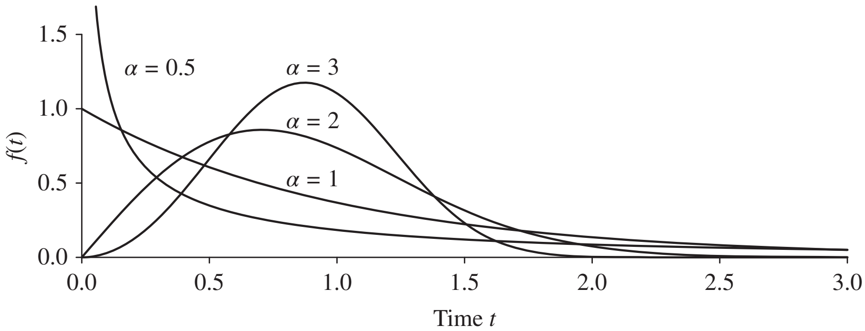


Figure 5.20 The probability density function of the Weibull distribution for selected values of the shape parameter α ($\theta = 1$).

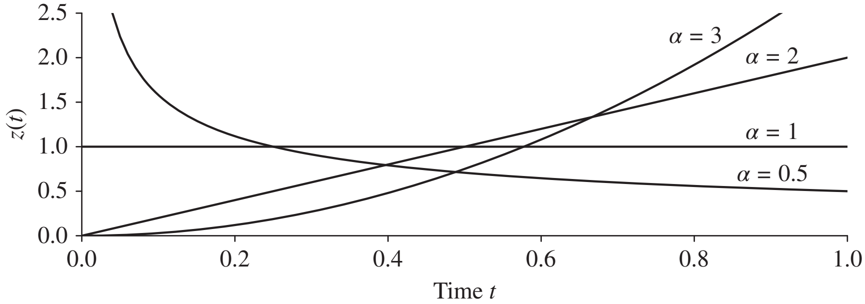


Figure 5.21 Failure rate function of the Weibull distribution, $\theta = 1$ and four different shape parameter (α) values.

When $\alpha = 1$, the failure rate is constant, when $\alpha > 1$, the failure rate function is increasing, and when $0 < \alpha < 1$, $z(t)$ is decreasing. When $\alpha = 2$ (such that the failure rate function is linearly increasing, see Figure 5.21), the resulting distribution is known as the *Rayleigh distribution*. The failure rate function $z(t)$ of the Weibull distribution is shown in Figure 5.21 for some selected values of α. The Weibull distribution is seen to be flexible and may be used to model time-to-failure distributions, where the failure rate function is decreasing, constant, or increasing.

Observe that

$$\begin{aligned} \alpha < 1 &\Rightarrow z(t) \text{ is a decreasing function of time} \\ \alpha = 1 &\Rightarrow z(t) \text{ is constant} \\ \alpha > 1 &\Rightarrow z(t) \text{ in an increasing function of time} \end{aligned} .$$

Remark 5.7 (A warning)
The failure rate function is seen to be discontinuous as a function of the shape parameter α at $\alpha = 1$. It is important to be aware of this discontinuity in numerical calculations, because, for example, $\alpha = 0.999$, $\alpha = 1.000$, and $\alpha = 1.001$ give significantly different failure rate functions for small values of t. □

Assume that $T \sim \text{Weibull}(\alpha, \theta)$ and consider the variable T^α. The survivor function of T^α is

$$\Pr(T^\alpha > t) = \Pr(T > t^{1/\alpha}) = \exp\left(-\frac{t}{\theta^\alpha}\right),$$

which means that T^α is exponentially distributed with constant failure rate $\lambda = 1/\theta^\alpha$.

The parameter θ is called the *characteristic lifetime* of the Weibull distribution. From (5.65), it follows that

$$R(\theta) = e^{-1} = \frac{1}{e} \approx 0.368, \qquad \text{for all } \alpha > 0.$$

This means that for any choice of the shape parameter α, the item will survive time θ with probability 36.8%.

MTTF

The MTTF of the two-parameter Weibull distribution is

$$\text{MTTF} = \int_0^\infty R(t)\, dt = \theta\, \Gamma\left(1 + \frac{1}{\alpha}\right). \tag{5.67}$$

The MTTF is equal to the characteristic lifetime θ multiplied with a factor that depends on the shape parameter α. This factor, $\Gamma(1 + 1/\alpha)$, varies with α as shown in Figure 5.22, which shows that MTTF is slightly less than θ when $\alpha \geq 1$.

The median time-to-failure t_m of the Weibull distribution is

$$R(t_m) = 0.50 \quad \Rightarrow \quad t_m = \theta(\log 2)^{1/\alpha}. \tag{5.68}$$

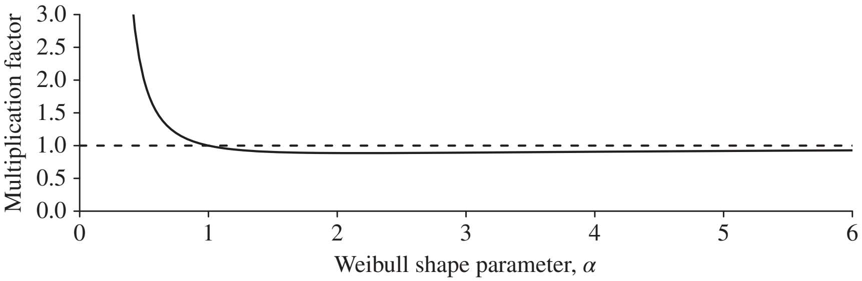


Figure 5.22 The proportionality factor of MTTF as a function of α.

The variance of T is

$$\mathrm{var(T)} = \theta^2 \left[\Gamma\left(1 + \frac{2}{\alpha}\right) - \Gamma^2\left(1 + \frac{1}{\alpha}\right)\right]. \tag{5.69}$$

Observe that $\mathrm{MTTF}/\sqrt{\mathrm{var(T)}}$ is independent of θ.

The Weibull distribution also arises as a limit distribution for the smallest of a large number of independent, identically distributed, nonnegative random variables. The Weibull distribution is therefore often called the *weakest link* distribution. This is further discussed in Section 5.5.3.

The Weibull distribution has been widely used in reliability analysis of semiconductors, ball bearings, engines, spot weldings, biological organisms, and so on. The Weibull distribution is discussed in detail by Murthy et al. (2003) and McCool (2012). The Weibull distribution is a basic distribution in R and is covered in several R packages. Interested readers may have a look at the package `Weibull-R`.

Example 5.6 (Choke valve)

The time-to-failure T of a variable choke valve is assumed to have a Weibull distribution with shape parameter $\alpha = 2.25$ and scale parameter $\theta = 8695$ hours. The valve will survive six months ($t = 4380$ hours) in continuous operation with probability

$$R(t) = \exp\left[-\left(\frac{t}{\theta}\right)^{\alpha}\right] = \exp\left[-\left(\frac{4380}{8695}\right)^{2.25}\right] \approx 0.808.$$

The MTTF is

$$\mathrm{MTTF} = \theta\, \Gamma\left(1 + \frac{1}{\alpha}\right) = 8695\ \Gamma(1.44)\ \mathrm{h}\ \approx 7701\ \mathrm{h},$$

and the median time-to-failure is

$$t_m = \theta\ (\log 2)^{1/\alpha} \approx 7387\ \mathrm{h}.$$

A valve that has survived the first six months ($t_1 = 4380$ hours), will survive the next six months ($t_2 = 4380$ hours) with probability

$$R(t_1 + t_2 \mid t_1) = \frac{R(t_1 + t_2)}{R(t_1)} = \frac{\exp\left[-\left(\frac{t_1+t_2}{\theta}\right)^{\alpha}\right]}{\exp\left[-\left(\frac{t_1}{\theta}\right)^{\alpha}\right]} \approx 0.448,$$

that is, significantly less than the probability that a new valve will survive six months.

The MRL when the valve has been functioning for six months ($x = 4380$ hours) is

$$\mathrm{MRL}(x) = \frac{1}{R(x)} \int_0^{\infty} R(t + x)\, dt \approx 4448\ \mathrm{h}.$$

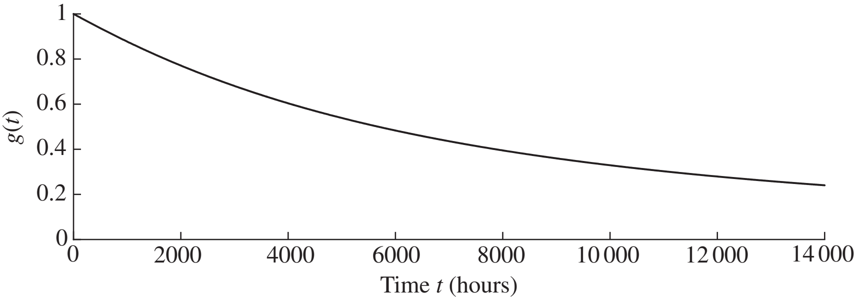


Figure 5.23 The scaled mean residual lifetime function $g(t) = \text{MRL}(x)/\text{MTTF}$ for the Weibull distribution with parameters $\alpha = 2.25$ and $\theta = 8760$ hours.

The MRL(x) cannot be given a simple closed form in this case and was therefore calculated by using a computer. The function $g(x) = \text{MRL}(x)/\ \text{MTTF}$ is shown in Figure 5.23. □

Series Structure of Independent Components

Consider a series structure of n components. The times-to-failure $T_1, T_2, \dots, T_n$ of the n components are assumed to be independent and Weibull distributed:

$$T_i \sim \text{Weibull}\ (\alpha, \theta_i) \quad \text{for } i = 1, 2, \dots, n.$$

A series structure fails as soon as the first component fails. The time-to-failure of the structure, T_s is thus

$$T_s = \min\{T_1, T_2, \dots, T_n\}.$$

The survivor function of this series structure becomes

$$\begin{aligned} R_s(t) &= \Pr(T > t) = \Pr\left(\min_{1\le i\le n} T_i > t\right) = \prod_{i=1}^{n} \Pr(T_i > t) \\ &= \prod_{i=1}^{n} \exp\left[-\left(\frac{t}{\theta_i}\right)^{\alpha}\right] = \exp\left[-\sum_{i=1}^{n}\left(\frac{t}{\theta_i}\right)^{\alpha}\right] = \exp\left[-\sum_{i=1}^{n}\left(\frac{1}{\theta_i}\right)^{\alpha} t^{\alpha}\right]. \end{aligned}$$

A series structure of independent components with Weibull time-to-failure distribution with the same shape parameter α, again has a Weibull time-to-failure distribution, with scale parameter $\theta_s = 1/\sum_{i=1}^{n}(1/\theta_i)^{1/\alpha}$ and with the shape parameter being unchanged.

Identical Components

When all the n components have the same distribution, such that $\theta_i = \theta$ for $i = 1, 2, \dots, n$, the series structure has a Weibull time-to-failure distribution with scale parameter $\theta/(n^{1/\alpha})$ and shape parameter α.

Example 5.7 (Numerical example)
Consider a series structure of n independent and identical components with Weibull distributed times-to-failure the same parameters as in Example 5.6, $\alpha = 2.25$ and $\theta = 8695$ hours. The MTTF of the series structure is

$$\mathrm{MTTF}_s = \theta_s \Gamma\left(1 + \frac{1}{\alpha}\right),$$

where

$$\theta_s = \frac{\theta}{n^{1/\alpha}}.$$

For a series structure of $n = 5$ components, the mean time-to-failure is

$$\mathrm{MTTF}_s = \frac{8695}{5^{1/2.25}}\,\Gamma\left(1 + \frac{1}{2.25}\right)\ \mathrm{h} = 3766.3\ \mathrm{h}.$$

In Figure 5.24, MTTF_s is shown as a function of n, the number of identical components in the series structure. □

Three-Parameter Weibull Distribution

The Weibull distribution we have discussed so far is a two-parameter distribution with shape parameter $\alpha > 0$ and scale parameter $\theta > 0$. A natural extension of this distribution is the *three-parameter Weibull distribution* (α, θ, ξ) with distribution function

$$F(t) = \Pr(T \le t) = \begin{cases} 1 - e^{-\left(\frac{t-\xi}{\theta}\right)^{\alpha}} & \text{for } t > \xi \\ 0 & \text{otherwise} \end{cases}. \tag{5.70}$$

The corresponding density is

$$f(t) = \frac{d}{dt}F(t) = \frac{\alpha}{\theta}\left(\frac{t-\xi}{\theta}\right)^{\alpha-1} e^{-\left(\frac{t-\xi}{\theta}\right)^{\alpha}} \qquad \text{for } t > \xi.$$

The third parameter ξ is sometimes called the *guarantee* or *threshold* parameter because the probability that a failure occurs before time ξ is 0.

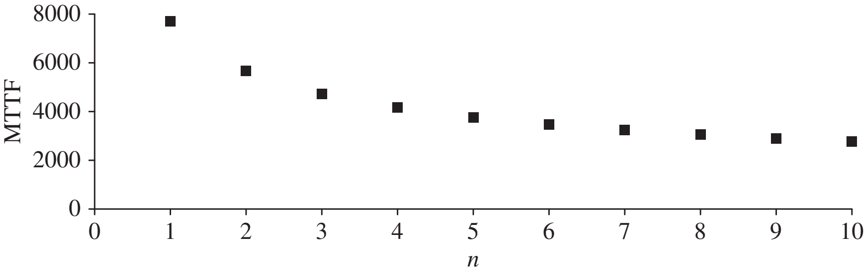


Figure 5.24 MTTF_s as a function of n, the number of independent and identical components in a series structure (Example 5.7).

Because $(T - \xi)$ obviously has a two-parameter Weibull distribution (α, θ), the mean and variance of the three-parameter Weibull distribution (α, θ, ξ) follows from (5.67) and (5.69).

$$\begin{aligned} \text{MTTF} &= \xi + \theta\, \Gamma\left(1 + \frac{1}{\alpha}\right). \\ \text{var}(T) &= \theta^2 \left[\Gamma\left(1 + \frac{2}{\alpha}\right) - \Gamma^2\left(1 + \frac{1}{\alpha}\right)\right]. \end{aligned}$$

In reliability applications, reference to the Weibull distribution usually means the two-parameter family, unless otherwise specified.

5.4.4 The Normal Distribution

The most commonly used distribution in statistics is the normal (Gaussian[4]) distribution. A random variable T is said to be normally distributed with mean ν and standard deviation τ, $T \sim \mathcal{N}(\nu, \tau^2)$, when the probability density of T is

$$f(t) = \frac{1}{\sqrt{2\pi}\tau} e^{-(t-\nu)^2/2\tau^2} \quad \text{for} - \infty < t < \infty. \tag{5.71}$$

The probability density function of $\mathcal{N}(\nu, \tau^2)$ may be plotted in R by the script

```
t<-seq(0,20,length=300)  # Set the time axis
# Set the parameters
nu<-10
tau<-2
# Calculate the normal density y for each t
y<-dnorm(t,nu,tau,log=F)
plot(t,y,type="l")
```

The resulting plot is shown in Figure 5.25. The $\mathcal{N}(0, 1)$ distribution is called the *standard normal distribution*. The distribution function of the standard normal distribution is usually denoted by $\Phi(\cdot)$. The probability density of the standard normal distribution is

$$\phi(t) = \frac{1}{\sqrt{2\pi}}\, e^{-t^2/2}. \tag{5.72}$$

The distribution function of $T \sim \mathcal{N}(\nu, \tau^2)$ may be written as

$$F(t) = \Pr(T \le t) = \Phi\left(\frac{t - \nu}{\tau}\right). \tag{5.73}$$

The normal distribution is sometimes used as a time-to-failure distribution, even though it allows negative values with positive probability.

[4] Named after the German mathematician Johann Carl Friedrich Gauss (1777–1855).

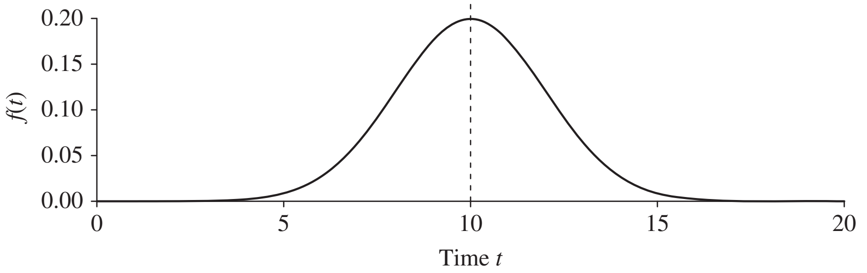


Figure 5.25 The normal distribution with mean $\mu = 10$ and standard deviation $\sigma = 2$.

Survivor Function

The survivor function of $T \sim \mathcal{N}(\nu, \tau^2)$ is

$$R(t) = 1 - \Phi\left(\frac{t-\nu}{\tau}\right). \tag{5.74}$$

Failure Rate Function

The failure rate function of $T \sim \mathcal{N}(\nu, \tau^2)$ is

$$z(t) = -\frac{R'(t)}{R(t)} = \frac{1}{\tau}\frac{\phi[(t-\nu)/\tau]}{1 - \Phi[(t-\nu)/\tau]}. \tag{5.75}$$

The failure rate function may be plotted in R by the script

```
t<-seq(-2,10,length=300)  # Set the time axis
# Set the parameters
nu<-10
tau<-2
# Calculate the failure rate function for each t
y<-dnorm(t,nu,tau,log=F)/(1-pnorm(t,nu,tau,log=F))
plot(t,y,type="l")
```

If $z_\Phi(t)$ is the failure rate function of the standard normal distribution, the failure rate function of $\mathcal{N}(\nu, \tau^2)$ is seen to be

$$z(t) = \frac{1}{\tau} z_\Phi\left(\frac{t-\nu}{\tau}\right).$$

The failure rate function of the standard normal distribution, $\mathcal{N}(0, 1)$, is shown in Figure 5.26. The failure rate function is *increasing* for all t and approaches $z(t) = t$ when $t \to \infty$.

When a random variable has a normal distribution but with an upper bound and/or a lower bound for the values of the random variable, the resulting distribution is called a *truncated normal distribution*. When there is only a lower bound,

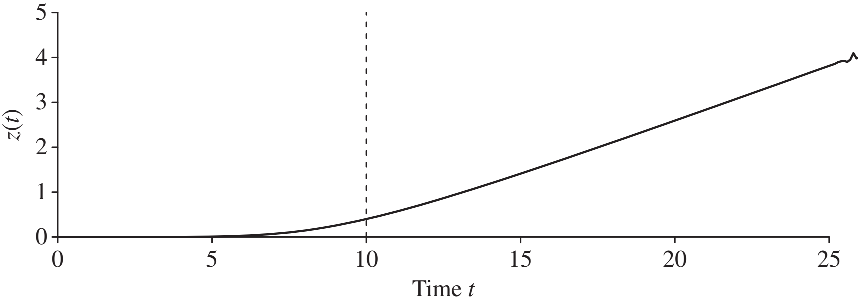


Figure 5.26 Failure rate function of the standard normal distribution wit mean $\mu = 10$ and standard deviation $\sigma = 2$.

the distribution is said to be left truncated. When there is only an upper bound, the distribution is said to be right truncated. Should there be an upper as well as a lower bound, it is said to be doubly truncated.

A normal distribution, left truncated at 0, is sometimes used as a time-to-failure distribution. This left truncated normal distribution has survivor function

$$R(t) = \Pr(T > t \mid T > 0) = \frac{\Phi[(\nu - t)/\tau]}{\Phi(\nu/\tau)} \qquad \text{for } t \geq 0. \tag{5.76}$$

The corresponding failure rate function becomes

$$z(t) = \frac{-R'(t)}{R(t)} = \frac{1}{\tau}\frac{\phi[(t-\nu)/\tau]}{1 - \Phi[(t-\nu)/\tau]} \qquad \text{for } t \geq 0.$$

Observe that the failure rate function of the left truncated normal distribution is identical to the failure rate function of the (untruncated) normal distribution when $t \geq 0$.

Example 5.8 (Wear-out of car tires)

A specific type of car tires has an average wear-out "time" T of 50 000 km, and 5% of the tires last for at least 70 000 km. We assume that T is normally distributed with mean $\nu = 50\ 000$ km, and that $\Pr(T > 70\ 000) = 0.05$. Let τ be the standard deviation of T. The variable $(T - 50\ 000)/\tau$ then has a standard normal distribution. Standardizing, we get

$$\Pr(T > 70\ 000) = 1 - \Pr\left(\frac{T - 50\ 000}{\tau} \leq \frac{70\ 000 - 50\ 000}{\tau}\right) = 0.05.$$

Therefore,

$$\Phi\left(\frac{20\ 000}{\tau}\right) = 0.95 \approx \Phi(1.645)$$

and

$$\frac{20\ 000}{\tau} \approx 1.645 \qquad \Rightarrow \qquad \tau \approx 12\ 158.$$

The probability that a tire will last more than 60 000 km is now

$$\Pr(T > 60\ 000) = 1 - \Pr\left(\frac{T - 50\ 000}{12\ 158} \le \frac{60\ 000 - 50\ 000}{12\ 158}\right)$$
$$\approx 1 - \Phi(0.795) \approx 0.205.$$

The probability of a “negative” time-to-failure is in this case

$$\Pr(T < 0) = \Pr\left(\frac{T - 50\ 000}{12\ 158} < \frac{-50\ 000}{12\ 158}\right) \approx \Phi(-4.11) \approx 0.$$

The effect of using a truncated normal distribution instead of a normal distribution is therefore negligible. □

5.4.5 The Lognormal Distribution

The time-to-failure T of an item is said to be *lognormally distributed* with parameters ν and τ, $T \sim \text{lognorm}(\nu, \tau)$, if $Y = \log T$ is normally (Gaussian) distributed with mean ν and standard deviation τ [i.e. $Y \sim \mathcal{N}(\nu, \tau^2)$]. The probability density function of T is

$$f(t) = \begin{cases} \frac{1}{\sqrt{2\pi}\ \tau\ t}\ e^{-\frac{1}{2\tau^2}(\log t - \nu)^2} & \text{for } t > 0 \\ 0 & \text{otherwise} \end{cases}. \tag{5.77}$$

The probability density function of the lognormal distribution may be plotted in R by the script

```
t<-seq(0,10,length=300) # Set the time axis
# Set the parameters:
nu<-5
tau<-2
# Calculate the lognormal density y for each t
y<-dlnorm(t,nu,tau,log=F)
plot(t,y,type="l")
```

The reader is encouraged to make the plot for various values of ν and τ and to study how the shape of the density varies with the parameter values. The lognormal probability density is sketched in Figure 5.27.

The MTTF is

$$\text{MTTF} = e^{\nu + \tau^2/2}, \tag{5.78}$$

the median time-to-failure (i.e. satisfying $R(t_m) = 0.5$) is

$$t_m = e^{\nu}, \tag{5.79}$$

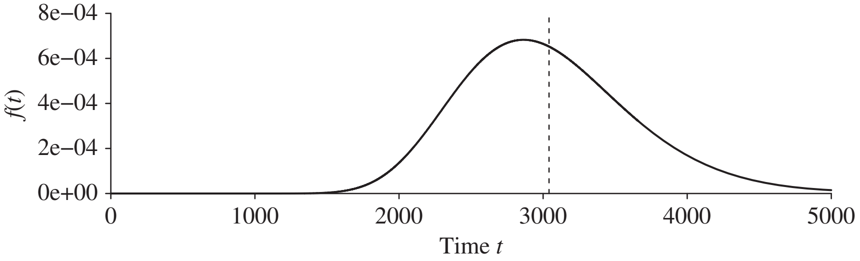


Figure 5.27 Probability density of the lognormal distribution with $\nu = 8$ and $\tau = 0.2$. The mean value is indicated by the dotted line.

and the mode of the distribution is

$$t_{\text{mode}} = e^{\nu-\tau^2}.$$

Observe that the MTTF may be written

$$\text{MTTF} = t_m\, e^{\tau^2/2},$$

and that the mode may be written

$$t_{\text{mode}} = t_m\, e^{-\tau^2}.$$

It is therefore easy to see that

$$t_{\text{mode}} < t_m < \text{MTTF}, \qquad \text{for } \tau > 0.$$

The variance of T is

$$\text{var}(T) = e^{2\nu}(e^{2\tau^2} - e^{\tau^2}). \tag{5.80}$$

Survivor Function

The survivor function of $T \sim \text{lognorm}(\nu, \tau)$ is

$$\begin{aligned} R(t) &= \Pr(T > t) = \Pr(\log T > \log t) \\ &= \Pr\left(\frac{\log T - \nu}{\tau} > \frac{\log t - \nu}{\tau}\right) = \Phi\left(\frac{\nu - \log t}{\tau}\right), \end{aligned} \tag{5.81}$$

where $\Phi(\cdot)$ is the distribution function of the standard normal distribution.

Failure Rate Function

The failure rate function of $T \sim \text{lognorm}(\nu, \tau)$ is

$$z(t) = -\frac{d}{dt}\left[\log \Phi\left(\frac{\nu - \log t}{\tau}\right)\right] = \frac{\phi[(\nu - \log t)/\tau)]/\tau t}{\Phi[(\nu - \log t)/\tau]/\tau}, \tag{5.82}$$

where $\phi(t)$ is the probability density of the standard normal distribution.

The failure rate function of the lognormal distribution may be plotted in R by the script

```
t<-seq(0,12000,1)  # Set the time axis
# Set the parameters:
nu<-8
tau<-0.2
# Calculate the failure rate y for each t:
y<-dlnorm(t,nu,tau)/(1-plnorm(t,nu,tau)
plot(x,y,type="l")
```

The shape of $z(t)$ is discussed in detail by Sweet (1990) who describes an iterative procedure to compute the time t for which the failure rate function attains its maximum value. He proves that $z(t) \to 0$ when $t \to \infty$.

Let $T_1, T_2, \ldots, T_n$ be independent lognormally distributed functions with parameters ν_i and τ_i^2 for $i = 1, 2, \ldots, n$. The product $T = \prod_{i=1}^{n} T_i$ is then lognormally distributed with parameters $\sum_{i=1}^{n} \nu_i$ and $\sum_{i=1}^{n} \tau_i^2$.

Repair Time Distribution

The lognormal distributed is commonly used as a distribution for repair time. The *repair rate* is defined analogous to the failure rate. When modeling the repair time, it is natural to assume that the repair rate is increasing, at least in a first phase. This means that the probability of completing the repair action within a short interval increases with the elapsed repair time. When the repair has been going on for a rather long time, this indicates serious problems, for example that there are no spare parts available on the site. It is therefore natural to believe that the repair rate is decreasing after a certain period of time, namely, that the repair rate function has the same shape as the failure rate function of the lognormal distribution shown in Figure 5.28.

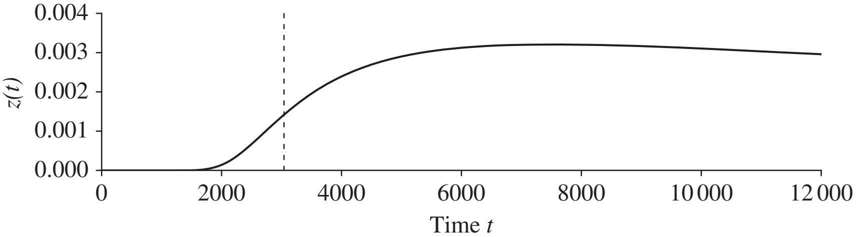


Figure 5.28 Failure rate function of the lognormal distribution with $\nu = 8$ and $\tau = 0.2$. The MTTF is indicated by the dotted line.

Median and Error Factor

In some cases, we may be interested to find an interval (t_L, t_U) such that $\Pr(t_L < T \leq t_U) = 1 - 2\alpha$, for example. If the interval is symmetric in the sense that $\Pr(T \leq t_L) = \alpha$ and $\Pr(T > t_U) = \alpha$, it is easy to verify that $t_L = e^{-u_\alpha \tau}$ and $t_U = e^{u_\alpha \tau}$, where u_α is the upper α% percentile of the standard normal distribution [i.e. $\Phi(u_\alpha) = 1 - \alpha$]. By introducing the median $t_m = e^{\nu}$ and $k = e^{u_\alpha \tau}$, the lower limit t_L and the upper limit t_U may be written

$$t_L = \frac{t_m}{k} \quad \text{and} \quad t_U = k\, t_m. \tag{5.83}$$

The factor k is often called the $(1 - 2\alpha)$ *error factor*. α is usually chosen to be 0.05.

Uncertainty in Failure Rate Estimate

In many situations, the (constant) failure rate λ may vary from one item to another. In the Reactor Safety Study (NUREG-75/014), the variation (uncertainty) in λ was modeled by a lognormal distribution, that is, the failure rate λ is regarded as a random variable Λ with a lognormal distribution.

In the Reactor Safety Study, the lognormal distribution was determined by the median λ_m and a 90% error factor k such that

$$P\left(\frac{\lambda_m}{k} < \Lambda < k\lambda_m\right) = 0.90.$$

If we, as an example, choose the median to be $\lambda_m = 6.0 \times 10^{-5}$ failures/h, and an error factor $k = 3$, then the 90% interval is equal to $(2.0 \times 10^{-5}, 1.8 \times 10^{-4})$. The parameters ν and τ of the lognormal distribution can now be determined from (5.79) and (5.83).

$$\nu = \log(\lambda_m) = \log 6.0 \times 10^{-5} \approx -9.721.$$
$$\tau = \frac{1}{1.645} \log k = \frac{1}{1.645} \log 3 \approx 0.668.$$

With these parameter values, the MTTF is equal to

$$\text{MTTF} = e^{\nu + \tau^2/2} \approx 1.47 \times 10^{-4} \text{ h}.$$

Example 5.9 (Fatigue analysis)

The lognormal distribution is commonly used in the analysis of fatigue failures. Considering the following simple situation: A smooth, polished test rod of steel is exposed to sinusoidal stress cycles with a given stress range (double amplitude) s. We want to estimate the time-to-failure of the test rod (i.e. the number of stress cycles N, until fracture occurs). In this situation, it is usually assumed that N is lognormally distributed. The justification for this is partly physical and partly mathematical convenience. A fatigue crack will always start in an area with local yield, normally caused by an impurity in the material. It seems reasonable that in the beginning the failure rate function increases with the number of stress cycles. If the test rod survives a large number of stress cycles, this indicates that there are very few impurities in the material. It is therefore to be expected that the failure rate function will decrease when the possibility for impurities in the material is reduced.

It is known that within a limited area of the stress range s, the number N of cycles to failure will roughly satisfy the equation

$$Ns^b = c, \tag{5.84}$$

where b and c are constants depending on the material and the geometry of the test rod. They may also depend on the surface treatment and the environment in which the rod is used.

By taking the logarithms of both sides of (5.83), we get

$$\log N = \log c - b \ \log s. \tag{5.85}$$

If we introduce $Y = \log N, \alpha = \log c, \beta = -b$ and $x = \log s$, it follows from (5.84) that Y roughly can be expressed by the relation

$$Y = \alpha + \beta x + \text{random error.}$$

If N is assumed to be lognormally distributed, then $Y = \log N$ will be normally distributed, and the usual theory for linear regression models applies when estimating the expected number of cycles to failure for a given stress range s. Equation (5.84) represents the Wöhler[5] or s–N diagram for the test rod. Such a diagram is shown in Figure 5.29. When the stress range is below a certain value s_0, the test rod will not fracture, irrespective of how many stress cycles it is exposed to. Equation (5.84) is therefore valid only for stress values above s_0.

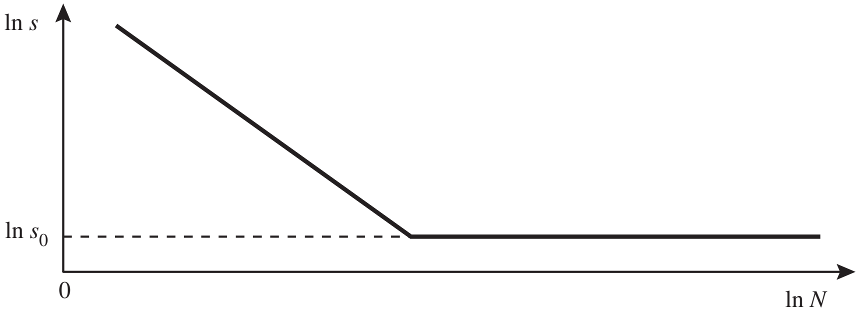


Figure 5.29 Wöhler or s–N diagram.

[5] Named after the German engineer August Wöhler (1819–1914).

The stress range s_0 is called the fatigue limit. For certain materials such as aluminum, the Wöhler curve has no horizontal asymptote. Such materials therefore have no fatigue limit. In a corrosive environment, such as salt water, neither does steel have any fatigue limit. □

5.4.6 Additional Time-to-Failure Distributions

Section 6.4 has, so far, presented the most common time-to-failure distributions used in practical reliability analyses. There are several other time-to-failure distributions that are not covered. Two of the most important are

Birnbaum–Saunders distribution. This distribution was developed as a time-to-failure distribution for fatigue failures in aircrafts (see Birnbaum and Saunders 1969). For a brief survey of the main properties of the Birnbaum–Saunders distribution, see https://en.wikipedia.org/wiki/Birnbaum-Saunders_distribution.

Inverse Gaussian distribution. This distribution is sometimes used as time-to-failure distribution for fatigue failures (e.g. see Chhikara and Folks 1989). The inverse Gaussian distribution resembles the lognormal distribution, but its failure rate function does not approach zero when the time increases. For a brief survey of the main properties of the inverse Gaussian distribution, see https://en.wikipedia.org/wiki/Inverse_Gaussian_distribution.

5.5 Extreme Value Distributions

Extreme value distributions play an important role in reliability analysis. They arise in a natural way, for example, in the analysis of engineering systems, made up of n identical items with a series structure, and in the study of corrosion of metals, of material strength, and of breakdown of dielectrics.

Let $T_1, T_2, \dots, T_n$ be independent, identically distributed random variables (not necessarily times-to-failure) with a continuous distribution function $F_T(t)$, for the sake of simplicity assumed to be strictly increasing for $F_T^{-1}(0) < t < F_T^{-1}(1)$. Then

$$T_{(1)} = \min\{T_1, T_2, \dots, T_n\} = U_n \tag{5.86}$$

$$T_{(n)} = \max\{T_1, T_2, \dots, T_n\} = V_n \tag{5.87}$$

are called the *extreme values*.

The distribution functions of U_n and V_n are easily expressed by $F_T(\cdot)$ in the following way (e.g. see Cramér 1946; Mann et al. 1974):

$$F_{U_n}(u) = 1 - [1 - F_T(u)]^n = L_n(u) \tag{5.88}$$

and

$$F_{V_n}(v) = F_T(v)^n = H_n(v). \tag{5.89}$$

Despite of the simplicity of (5.88) and (5.89), these formulas are usually not easy to work with. If $F_T(t)$, say, represents a normal distribution, one is lead to work with powers of $F_T(t)$, which may be cumbersome.

In many practical reliability applications, n is very large. Hence, one is lead to look for asymptotic techniques, which under general conditions on $F_T(t)$ may lead to simple representations of $F_{U_n}(u)$ and $F_{V_n}(v)$.

Cramér (1946) suggests the following approach: Introduce

$$Y_n = nF_T(U_n),$$

where U_n is defined as in (5.86). Then for $y \geq 0$,

$$\begin{aligned}\Pr(Y_n \leq y) &= P\left(F_T(U_n) \leq \frac{y}{n}\right)\\ &= P\left[U_n \leq F_T^{-1}\left(\frac{y}{n}\right)\right]\\ &= F_{U_n}\left[F_T^{-1}\left(\frac{y}{n}\right)\right]\\ &= 1 - \left[1 - F_T\left(F_T^{-1}\left(\frac{y}{n}\right)\right)\right]\\ &= 1 - \left(1 - \frac{y}{n}\right)^n.\end{aligned} \tag{5.90}$$

As $n \to 0$

$$\Pr(Y_n \leq y) \to 1 - e^{-y} \quad \text{for } y > 0. \tag{5.91}$$

Because the right hand side of (5.91) is the distribution function for the exponential distribution with parameter $\lambda = 1$, it is continuous for $y > 0$, this implies that Y_n converges in distribution to a random variable Y, with distribution function

$$F_Y(y) = 1 - e^{-y} \quad \text{for } y > 0. \tag{5.92}$$

It follows from (5.88) that the distribution of U_n becomes more and more similar to the distribution of the random variable $F_T^{-1}(Y/n)$ when n increases. Therefore,

$$\Pr(U_n \leq x) \approx \Pr\left(F_T^{-1}\left[\frac{Y}{n}\right] \leq x\right) \quad \text{when } n \text{ is "large".} \tag{5.93}$$

Similarly, let

$$Z_n = n[1 - F_T(V_n)], \tag{5.94}$$

where V_n is defined in (5.87). By an analogous argument, it can be shown that for $z > 0$.

$$\Pr(Z_n \leq z) = 1 - \left(1 - \frac{z}{n}\right)^n, \tag{5.95}$$

which implies that the distribution of V_n becomes more and more similar to the distribution of the random variable $F_T^{-1}(1 - Z/n)$ when n increases. Therefore,

$$\Pr(V_n \le x) \approx \Pr\left(F_T^{-1}\left[1 - \frac{Z}{n}\right] \le x\right) \quad \text{when } n \text{ is ``large''}, \tag{5.96}$$

where Z has distribution function

$$\Pr(Z \le z) = 1 - e^{-z} \qquad \text{for } z > 0. \tag{5.97}$$

It is to be expected that the limit distribution of U_n and V_n will depend on the type of distribution $F_T(\cdot)$, but it turns out that there are only three possible types of limiting distributions for the minimum extreme U_n, and only three possible types of limiting distributions for the maximum extreme V_n.

For a comprehensive discussion of the application of extreme value theory to reliability analysis, see Mann et al. (1974), Lawless (1982), and Johnson and Kotz (1970). Here, we content ourselves with mentioning three of the possible types of limiting distributions, and indicate areas where they are applied.

5.5.1 The Gumbel Distribution of the Smallest Extreme

If the probability density $f_T(t)$ of the T_is approaches zero exponentially as $t \to \infty$, then the limiting distribution of $U_n = T_{(1)} = \min\{T_1, T_2, \dots, T_n\}$ is of the form

$$F_{T_{(1)}}(t) = 1 - \exp(-e^{(t-\vartheta)/\alpha}) \quad \text{for} - \infty < t < \infty, \tag{5.98}$$

where $\alpha > 0$ and ϑ are constants. α is the mode, and ϑ is a scale parameter.

The corresponding "survivor" function is

$$R_{T_{(1)}}(t) = 1 - F_{T_{(1)}}(t) = \exp(-e^{(t-\vartheta)/\alpha}) \quad \text{for} - \infty < t < \infty. \tag{5.99}$$

Gumbel (1958) calls this distribution the Type I asymptotic distribution of the smallest extreme. It is now called the *Gumbel distribution of the smallest extreme*.[6] If standardized variables

$$Y = \frac{T - \vartheta}{\alpha} \tag{5.100}$$

are introduced, the distribution function takes the form

$$F_{Y_{(1)}}(y) = 1 - \exp(-e^y) \quad \text{for} - \infty < y < \infty,$$

with probability density

$$f_{Y_{(1)}}(y) = e^y \exp(-e^y) \quad \text{for} - \infty < y < \infty. \tag{5.101}$$

The corresponding "failure rate" is

$$z_{Y_{(1)}}(y) = \frac{f_{Y_{(1)}}(y)}{1 - F_{Y_{(1)}}(y)} = e^y \quad \text{for} - \infty < y < \infty. \tag{5.102}$$

[6] Named after the German mathematician Emil Julius Gumbel (1891–1966).

The mean value of $T_{(1)}$ is (see Lawless 1982, p. 19)

$$E(T_{(1)}) = \vartheta - \alpha\gamma,$$

where $\gamma = 0.5772\ldots$ is known as Euler's constant.

Because $T_{(1)}$ can take negative values, (5.101) is not a valid time-to-failure distribution, but a valid time-to-failure distribution is obtained by left-truncating (5.101) at $t = 0$. In this way, we get the *truncated* Gumbel distribution of the smallest extreme, which is given by the survivor function

$$\begin{aligned} R^0_{T_{(1)}}(t) &= \Pr(T_{(1)} > t \mid T > 0) = \frac{\Pr(T_{(1)} > t)}{\Pr(T_{(1)} > 0)} \\ &= \frac{\exp(-e^{(t-\vartheta)/\alpha})}{\exp(-e^{\vartheta/\alpha})} = \exp(-e^{-(\vartheta/\alpha)(e^{t/\alpha}-1)}) \qquad \text{for } t > 0. \end{aligned} \tag{5.103}$$

By introducing new parameters $\beta = e^{-\vartheta/\alpha}$ and $\varrho = 1/\alpha$, the truncated Gumbel distribution of the smallest extreme is given by the survivor function

$$R^0_{T_{(1)}}(t) = \exp[-\beta(e^{\varrho t} - 1)] \qquad \text{for } t > 0. \tag{5.104}$$

The failure rate function of the truncated distribution is

$$z^0_{T_{(1)}}(t) = -\frac{d}{dt}\log R^0_{T_{(1)}}(t) = \frac{d}{dt}\,\beta(e^{\varrho t} - 1) = \beta\varrho e^{\varrho t} \qquad \text{for } t \geq 0. \tag{5.105}$$

5.5.2 The Gumbel Distribution of the Largest Extreme

If the probability density $f_T(t)$ approaches zero exponentially as $t \to \infty$, then the limiting distribution of $V_n = T_{(n)} = \max\{T_1, T_2, \ldots, T_n\}$ is of the form

$$F_{T_{(n)}}(t) = e^{-e^{-(t-\vartheta)/\alpha}} \qquad \text{for} -\infty < t < \infty,$$

where $\alpha > 0$ and ϑ are constants. Gumbel (1958) calls this distribution the Type I asymptotic distribution of the largest extreme. It is now known as the *Gumbel distribution of the largest extreme.*

If standardized variables are introduced, the distribution takes the form

$$F_{Y_{(n)}}(y) = \exp(-e^{-y}) \quad \text{for} -\infty < y < \infty, \tag{5.106}$$

with probability density

$$f_{Y_{(n)}}(y) = e^{-y}\exp(-e^{-y}) \quad \text{for} -\infty < y < \infty. \tag{5.107}$$

5.5.3 The Weibull Distribution of the Smallest Extreme

Another limiting distributions for the smallest extreme is the Weibull distribution

$$F_{T_{(1)}}(t) = 1 - \exp(-[(t - \vartheta)/\eta]^{\beta}) \qquad \text{for } t \geq \vartheta, \tag{5.108}$$

where $\beta > 0$, $\eta > 0$, and $\vartheta > 0$ are constants.

Introducing standardized variables [see (5.100)],

$$F_{Y_{(1)}}(y) = 1 - \exp(-y^{\beta}) \qquad \text{for } y > 0 \text{ and } \beta > 0. \tag{5.109}$$

This distribution is also called the Type III asymptotic distribution of the smallest extreme.

Example 5.10 (Pitting corrosion)

Consider a steel pipe with wall thickness θ which is exposed to corrosion. Initially, the surface has a certain number n of microscopic pits. Pit i has a depth D_i, for $i = 1, 2, \ldots, n$. Due to corrosion, the depth of each pit will increase with time. Failure occurs when the first pit penetrates the surface, that is when $\max\{D_1, D_2, \ldots, D_n\} = \theta$.

Let T_i be the time pit i will need to penetrate the surface, for $i = 1, 2, \ldots, n$. The time-to-failure T of the item is

$$T = \min\{T_1, T_2, \ldots, T_n\}.$$

Assume that the time to penetration T_i is proportional to the remaining wall thickness, that is $T_i = k(\theta - D_i)$. We further assume that k is independent of time, which implies that the corrosion rate is constant.

Assume next that the random initial depths of the pits $D_1, \ldots, D_n$ are independent and identically distributed with a right truncated exponential distribution. Then the distribution function of D_i is

$$\begin{aligned} F_{D_i}(d) &= \Pr(D_i \le d \mid D_i \le \theta) = \frac{\Pr(D_i \le d)}{\Pr(D_i \le \theta)} \\ &= \frac{1 - e^{-\eta d}}{1 - e^{-\eta\theta}} \quad \text{for } 0 \le d \le \theta. \end{aligned}$$

The distribution function of the time to penetration, T_i, is thus

$$\begin{aligned} F_{T_i}(t) &= \Pr(T_i \le t) = \Pr(k(\theta - D_i) \le t) = P\left(D_i \ge \theta - \frac{t}{k}\right) \\ &= 1 - F_{D_i}\left(\theta - \frac{t}{k}\right) = \frac{e^{\eta t/k} - 1}{e^{\eta\theta} - 1} \quad \text{for } 0 \le t \le k\theta, \end{aligned} \tag{5.110}$$

and the survivor function $R(t)$ of the item becomes

$$R(t) = \Pr(T > t) = [1 - F_{T_i}(t)]^n \qquad \text{for } t \ge 0.$$

If we assume that the number n of pits is very large, then as $n \to \infty$, we get

$$R(t) = [1 - F_{T_i}(t)]^n \approx e^{-nF_{T_i}(t)} \qquad \text{for } t \ge 0.$$

By using (5.110)

$$R(t) \approx \exp\left(-n\frac{e^{\eta t/k} - 1}{e^{\eta\vartheta}}\right) \qquad \text{for } t \ge 0.$$

By introducing new parameters $\beta = n/(e^{\eta\vartheta} - 1)$ and $\varrho = \eta/k$, we get

$$R(t) \approx \exp(-\beta(e^{\varrho t} - 1)) \qquad \text{for } t \geq 0,$$

which is equal to (5.104), namely the time-to-failure caused by pitting corrosion has approximately a truncated Gumbel distribution of the smallest extreme.

A similar example is discussed by Mann et al. (1974), Lloyd and Lipow (1962), and Kapur and Lamberson (1977). □

5.6 Time-to-Failure Models With Covariates

The reliability of items is often found to be influenced by one or more *covariates*. A *covariate* is a variable, condition, or property that can influence the time-to-failure T of an item, either because it has a direct causal relationship to the time-to-failure or because it influences the survival time in a noncausal way. Examples of covariates that can influence T are temperature, humidity, voltage, and vibrations. The covariates may be continuous or discrete variables. In some cases, it is relevant to use binary variables to distinguish two types of items or two types of operation (e.g. active or standby). A covariate is also called a *concomitant variable*, an *explanatory variable*, or a *stressor*.

In most applications, the items are exposed to several covariates $\boldsymbol{s} = (s_1, s_2, \dots, s_k)$, where $\boldsymbol{s}$ is called a *covariate vector*. Each covariate can take several different levels and in a sample of times-to-failure, each time-to-failure may be associated with a specific set of values for the k covariates.

So far in the book, we have tacitly assumed that all the covariates are kept constant. Many situations require that we consider the reliability of items operating under different conditions, where items are influenced by different covariate vectors. This is, for example, relevant when

- We have adequate knowledge about the reliability of the item in a *baseline* situation with known and constant covariates, but wonder what will happen if some of the covariates (e.g. temperature, voltage) are changed.
- We have datasets of field data from several, slightly different application conditions (i.e. covariate vectors) and wonder how we can use all the datasets to assess the reliability for a particular covariate vector.
- We have identical items that are used under different conditions and want to identify the covariates with the strongest influence on the reliability (i.e. that need to be controlled).

To identify the relevant covariates requires a thorough understanding of the potential failure modes of the item and of which factors that may influence the occurrence of the failure. Some main covariates influencing a shutdown valve are listed in Example 5.11.

Example 5.11 (Covariates for a shutdown valve)
Typical covariates for a shutdown valve in a process plant may include

- Corrosiveness of the fluid (flowing through the valve)
- Erosiveness of the fluid (i.e. presence and amount of particles in the fluid)
- Flow-rate of the fluid
- Pressure of the fluid
- Test principle used when proof-testing the valve
- Test frequency of proof-tests.

□

The influence of the covariate vector on the reliability can be modeled in many different ways. This section deals with three different models:

- The accelerated failure time (AFT) model
- The Arrhenius model
- The proportional hazards (PH) model.

5.6.1 Accelerated Failure Time Models

Consider an item that has been used in a baseline application where its reliability characteristics are known. A new application is planned with a constant covariate vector $\boldsymbol{s} = (s_1, s_2, \dots, s_k)$, where s_i is measured as the deviation from the baseline application, for $i = 1, 2, \dots, k$. Our aim is to describe how the reliability of the item is influenced by the new covariate vector $\boldsymbol{s}$. The time-to-failure of an item operating with the baseline covariate vector is denoted T_0, with corresponding survivor function $R_0(t)$ and failure rate function $z_0(t)$.

The AFT model assumes that the influence of the covariates $\boldsymbol{s}$ can be modeled by a factor $h(\boldsymbol{s}) > 0$ that directly scales the time-to-failure, such that T has the same distribution as $T_0/h(\boldsymbol{s})$.

The deterioration of the item is therefore accelerated by increasing $h(\boldsymbol{s})$. When $h(\boldsymbol{s}) < 1$, an item with covariate vector $\boldsymbol{s}$ deteriorates slower than the baseline item, and when $h(\boldsymbol{s}) > 1$, the item with covariate vector $\boldsymbol{s}$ deteriorates faster than the baseline item.

The survivor function of T with covariate vector $\boldsymbol{s}$ is

$$\begin{aligned} R(t \mid \boldsymbol{s}) &= \Pr(T > t \mid \boldsymbol{s}) = \Pr\left(\frac{T_0}{h(\boldsymbol{s})} > t\right) \\ &= \Pr(T_0 > h(\boldsymbol{s})\, t) = R_0[h(\boldsymbol{s})\, t]. \end{aligned} \tag{5.111}$$

The probability density function of T with covariate vector $\boldsymbol{s}$ is

$$f(t \mid \boldsymbol{s}) = \frac{d}{dt} R(t \mid \boldsymbol{s}) = h(\boldsymbol{s})\, f_0[h(\boldsymbol{s})t]. \tag{5.112}$$

The failure rate function becomes

$$z(t \mid \boldsymbol{s}) = h(\boldsymbol{s})\; z_0[h(\boldsymbol{s})\; t]. \tag{5.113}$$

Because T has the same distribution as $T_0/h(\boldsymbol{s})$, $\text{MTTF}_{\boldsymbol{s}}$ is obviously

$$\text{MTTF}_{\boldsymbol{s}} = \frac{\text{MTTF}_0}{h(\boldsymbol{s})}. \tag{5.114}$$

Example 5.12 (Constant failure rate)
Consider identical items with a constant failure rate, where the constant failure rate assumption is realistic within a certain range of stress (covariate) levels. The failure rate has been estimated for baseline stress to be λ_0. Identical items are to be used with stress (covariate) level $\boldsymbol{s}$, where $\boldsymbol{s}$ is measured as the difference from the baseline stress level. We want to determine the failure rate $\lambda_{\boldsymbol{s}}$ for this stress level. If we know the scaling factor $h(\boldsymbol{s})$, $\lambda_{\boldsymbol{s}}$ can be found from (5.119), because $\text{MTTF}_{\boldsymbol{s}} = 1/\lambda_{\boldsymbol{s}}$, as

$$\lambda_{\boldsymbol{s}} = h(\boldsymbol{s})\; \lambda_0.$$

For items with constant failure rate, the AFT model implies that the failure rate at a specified stress level can be determined as the failure rate for normal stress multiplied by a constant (that is determined by the increases stress). □

For two identical items operating with two different covariate vectors $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, the ratio

$$\text{AF}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \frac{\text{MTTF}_1}{\text{MTTF}_2} = \frac{g(\boldsymbol{s}_2)}{g(\boldsymbol{s}_1)} \tag{5.115}$$

is called the *acceleration factor* (AF) of covariate $\boldsymbol{s}_1$ with respect to covariate $\boldsymbol{s}_2$.

5.6.2 The Arrhenius Model

The Arrhenius model[7] is one of the earliest acceleration models. Initially, Arrhenius studied how a *chemical reaction rate* varies as a function of temperature τ and found that

$$\nu(\tau) = A_0 \exp\left[-\frac{E_a}{k\tau}\right], \tag{5.116}$$

[7] Named after the Swedish scientist Svante Arrhenius (1859–1927).

where

- τ is the temperature measured in degrees Kelvin (the Celsius temperature plus 273.16°).
- $\nu(\tau)$ is the chemical reaction rate at temperature τ, that is, the amount of a reactant reacted per time unit.
- A_0 is a constant scaling factor (for each chemical reaction).
- E_a is the activation energy for the reaction.
- k is the universal gas constant that is equal to the Boltzmann constant, but expressed in another unit.

Taking the natural logarithm of (5.116) yields:

$$\log \nu(\tau) = \log A_0 - \frac{E_a}{k}\frac{1}{\tau}.$$

Rearranging gives

$$\log \nu(\tau) = \frac{-E_a}{k}\left(\frac{1}{\tau}\right) + \log A_0, \tag{5.117}$$

which is the equation for a straight line $y = ax + b$, where x is $1/\tau$, and where the slope and the intercept can be used to determine E_a and A_0.

The Arrhenius Model for Times-to-Failure

The Arrhenius model has been adapted to model accelerated times-to-failure for electronic and especially semiconductor components (and some other items) with respect to temperature. The Arrhenius failure time model is similar to the Arrhenius chemical reaction model in (5.121) and is given as

$$L(\tau) = A \exp\left[\frac{E_a}{k\tau}\right]. \tag{5.118}$$

The main differences between this model and (5.116) are

- A is a constant determined by the material properties of the item (whereas A_0 was determined by the properties of the chemical reaction).
- $L(\tau)$ is an expression for the time-to-failure of the item at temperature τ ($L(\tau)$ may, for example be a percentile of the distribution of the time-to-failure T. For constant failure rates, a natural choice for $L(\tau)$ is the MTTF (τ)).
- k is the Boltzmann constant (8.6171×10^{-5} eV/K°).
- The reaction rate $\nu(\tau)$ in (5.116) is the amount of reactant reacted per time unit. Assume that the reaction has reached a critical level. This event may be considered a "failure" of the reaction and the mean time until this event occurs for temperature τ is reciprocal to the reaction rate $\nu(\tau)$. The same idea is used for failure times, but instead of presenting (5.118) for the failure rate, it is presented with respect to $L(\tau)$, that can be considered the reciprocal of the failure rate and the minus sign has therefore disappeared in (5.118).

Consider a semiconductor that is tested at two different temperatures $\tau_1 < \tau_2$. The *acceleration factor* due to the change of temperature is

$$A(\tau_1, \tau_2) = \frac{L(\tau_1)}{L(\tau_2)} = \exp\left[\left(\frac{E_a}{k}\right)\left(\frac{1}{\tau_1} - \frac{1}{\tau_2}\right)\right]. \tag{5.119}$$

The acceleration factor is seen to be zero when $\tau_1 = \tau_2$, positive when $\tau_1 < \tau_2$, and negative when $\tau_1 > \tau_2$. Observe that the constant A has disappeared.

Example 5.13 (Constant failure rate)

Reconsider the situation in Example 5.14 where identical items with constant failure rate are operating in temperature τ. For items with constant failure rate, it is recommended to use MTTF (τ) as a measure of the time-to-failure $L(\tau)$. Because MTTF $= 1/\lambda$, the *survivor function* for temperature τ can from (5.123) be written as

$$R(t \mid \tau) = \exp(-\lambda t) = \exp\left(-\frac{t}{\text{MTTF}(\tau)}\right) = \exp\left(-\frac{t}{A\exp\left[\frac{E_a}{k\tau}\right]}\right).$$

If we are able to determine A and E_a, the survivor function may be determined as a function of the temperature τ.

Consider two different temperatures $\tau_1 < \tau_2$. The *acceleration factor* in (5.119) can be written as

$$A(\tau_1, \tau_2) = \frac{\text{MTTF }(\tau_1)}{\text{MTTF }(\tau_2)} = \frac{1/\lambda_1}{1/\lambda_2} = \frac{\lambda_2}{\lambda_1},$$

such that

$$\lambda_2 = A(\tau_1, \tau_2)\lambda_1.$$

For items with constant failure rate, the Arrhenius model is seen to give the same result as the PH model; the failure rate at increased stress is equal to the failure rate at initial stress multiplied with a constant. □

Example 5.14 (Weibull distribution)

Consider identical items with Weibull distributed times-to-failure. The items are operating in two different temperatures $\tau_1 < \tau_2$. For the Weibull distribution, it is usually recommended to use the median as metric for the time-to-failure $L(\tau)$. For temperature τ_1, the Weibull parameters are α_1 and θ_1, and for temperature τ_2, the Weibull parameters are α_2 and θ_2. The acceleration factor in (5.119) can now be written as

$$A(\tau_1, \tau_2) = \frac{\text{median}(\tau_1)}{\text{median}(\tau_2)} = \frac{\theta_1(\log 2)^{\frac{1}{\alpha_1}}}{\theta_2(\log 2)^{\frac{1}{\alpha_2}}} = \frac{\theta_1}{\theta_2}(\log 2)^{\left(\frac{1}{\alpha_1} - \frac{1}{\alpha_2}\right)}.$$

If we assume that the shape parameter is unchanged $\alpha_1 = \alpha_2$, the acceleration factor is simply

$$A(\tau_1, \tau_2) = \frac{\theta_1}{\theta_2}.$$

□

5.6.3 Proportional Hazards Models

PH models are the most popular models that include covariates. To put it briefly, a PH model splits the failure rate $z(t \mid \boldsymbol{s})$ of an item into two separate parts: (i) a part $z_0(t)$ that is a function of the time t but not of the stress vector $\boldsymbol{s} = (s_1, s_2, \dots, s_k)$ and (ii) a second part $g(\boldsymbol{s})$ that depends on the stress vector $\boldsymbol{s}$, but not on the time t. The PH model can be written as

$$z(t \mid \boldsymbol{s}) = z_0(t)\; g(\boldsymbol{s}).$$

The PH model is further discussed in Section 14.8.

5.7 Additional Continuous Distributions

This section introduces two continuous distributions: the uniform distribution and the beta distribution. None of these are commonly used as time-to-failure distributions, but they are used for several other purposes in reliability analyses.

5.7.1 The Uniform Distribution

A random variable X has a uniform distribution over an interval $[a, b]$ when

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{otherwise} \end{cases} . \tag{5.120}$$

That X has a uniform distribution is often written as $X \sim \text{unif}(a, b)$, for $a < b$. The uniform distribution is available in R where the uniform density, for example is available by the command `dunif(x,min=a,max=b,log=F)`. The probability density function of unif(0,1) is illustrated in Figure 5.30. The mean value of $X \sim$ unif(a, b) is

$$E(X) = \int_a^b x f_X(x)\, dx = \frac{a+b}{2}, \tag{5.121}$$

and the variance of X is

$$\text{var}(X) = \frac{(b-a)^2}{12}. \tag{5.122}$$

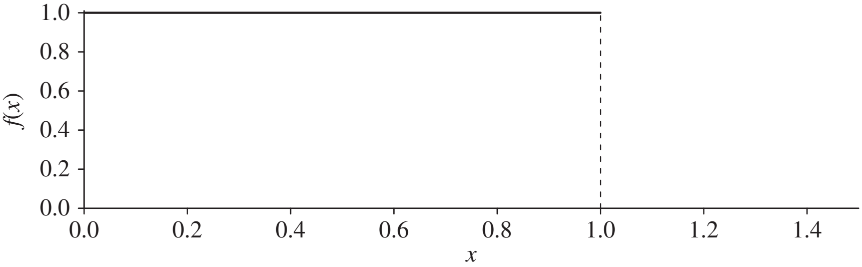


Figure 5.30 The probability density of $X \sim \text{unif}(0, 1)$.

The derivation is left to the reader as an exercise. In many applications, the interval $[a, b]$ is equal to $[0, 1]$.

5.7.2 The Beta Distribution

A random variable X has a beta distribution with parameters α and β over the interval $[0, 1]$ when

$$f_X(x) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1} & \text{for } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}, \tag{5.123}$$

where $\alpha > 0$ and $\beta > 0$. The statement X has a beta distribution with parameters α and β is often written as $X \sim \text{beta}(\alpha, \beta)$. The beta distribution is available in R where the beta density, for example, is available by the command `dbeta(x,shape1,shape2,log=F)`. The probability density function of the beta distribution is illustrated in Figure 5.31 for some selected values of α and β.

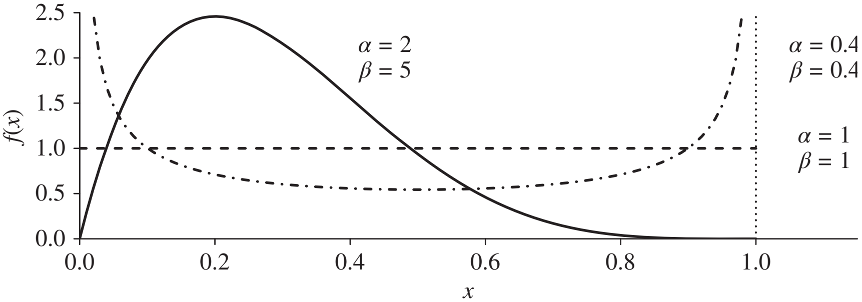


Figure 5.31 The probability density of $X \sim \text{beta}(\alpha, \beta)$ for some selected values of α and β.

A simple R script to plot the density of the beta distribution is as follows

```
x<-seq(0,1,length=300) # Set the values for the x-axis
# Set the parameters a (=alpha) and b (=beta)
a<-2
b<-5
# Calculate the beta density y for each x
y<-dbeta(x,a,b,log=F)
plot(x,y,type="l",xlab="x",ylab="f(x)")
```

The reader is encouraged to run this script for different sets of parameters to become familiar with the possible shapes of the beta density function.

The mean value of $X \sim \text{beta}(\alpha, \beta)$ is

$$E(X) = \frac{\alpha}{\alpha + \beta}, \tag{5.124}$$

and the variance is

$$\text{var}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \tag{5.125}$$

The beta distribution is used for several purposes in reliability analyses. In Chapter 15, the beta distribution is an important prior distribution for parameters. Observe that when $\alpha = \beta = 1$, the beta distribution is equal to the uniform distribution over $[0, 1]$, that is, $\text{beta}(1, 1) = \text{unif}(0, 1)$.

5.8 Discrete Distributions

This section introduces three *discrete* distributions: the *binomial distribution*, the *geometric distribution*, and the *negative binomial distribution*. All these are frequently used in reliability models and in connection with time-to-failure models. All distributions are used in the same setup called the *binomial situation*. In addition, the HPP is introduced.

5.8.1 Binomial Situation

The binomial situation must fulfill three requirements:

(1) n independent trials are carried out.
(2) Each trial has two possible outcomes A and A^*.
(3) The probability $\Pr(A) = p$ is the same in all the n trials.

The n trials are sometimes referred to as *Bernoulli trials*. A trial may, for example, be to start a fire pump (or some other engine). Outcome A may be that the start is successful and outcome A^* may be that the fire pump does not start. Each trial must be independent with the same probability $\Pr(A)$. This means that the fire pumps must be of the same type and that the startup procedure must be the same for each trial.

5.8.2 The Binomial Distribution

Consider a binomial situation and let X be the number of the n trials that have outcome A. X is then a discrete random variable with probability mass function

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad \text{for } x = 0, 1, \dots, n, \tag{5.126}$$

where $\binom{n}{x}$ is the binomial coefficient

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}.$$

The distribution (5.126) is called the *binomial distribution* (n, p) and is shown in Figure 5.32 for a simulated dataset with $n = 20$ and $p = 0.3$. For brevity, the binomial distribution is often written as $X \sim \text{binom}(n, p)$. The mean value and the variance of X are

$$E(X) = np. \tag{5.127}$$

$$\text{var}(X) = np(1-p). \tag{5.128}$$

5.8.3 The Geometric Distribution

Assume that we carry out a sequence of Bernoulli trials and want to find the number Z of trials until the first trial with outcome A. If $Z = z$, this means that the first $z - 1$ trials have outcome A^*, and that the first A will occur in trial z.

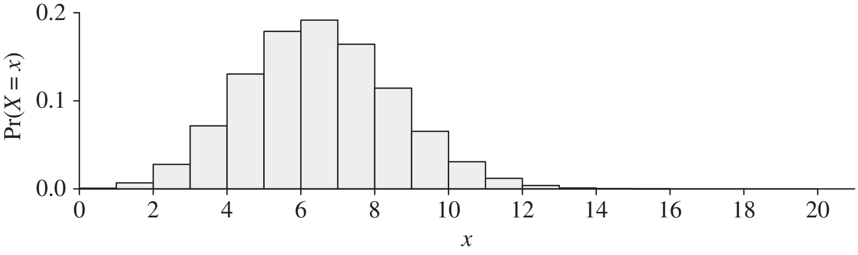


Figure 5.32 The binomial distribution (20, 0.3).

The probability mass function of Z is

$$\Pr(Z = z) = (1-p)^{z-1}p \qquad \text{for } z = 1, 2, \ldots. \tag{5.129}$$

The distribution (5.129) is called the *geometric distribution* with parameter p and is often written as $Z \sim \text{geom}(p)$. We have that

$$\Pr(Z > z) = (1-p)^z.$$

The mean value and the variance of Z are

$$E(Z) = \frac{1}{p}. \tag{5.130}$$

$$\text{var}(Z) = \frac{1-p}{p^2}. \tag{5.131}$$

5.8.4 The Negative Binomial Distribution

Again, assume that we carry out a series of independent Bernoulli trials (see Section 5.8.1). Let Z_r be the number of trials until a predefined number r of outcomes A have occurred. If $Z_r = z$, this means that during the first $z - 1$ trials, we got $r - 1$ outcomes A and in trial number z the rth outcome A occurred. The probability mass function of Z_r is

$$\begin{aligned}\Pr(Z_r = z) &= \binom{z-1}{r-1} p^{r-1}(1-p)^{z-r} \cdot p \\ &= \binom{z-1}{r-1} p^r(1-p)^{z-r} \qquad \text{for } z = r, r+1, r+2, \ldots.\end{aligned} \tag{5.132}$$

When $r = 1$, the negative binomial distribution becomes a geometric distribution.

The negative binomial distribution is sometimes defined by the random variable Y_r = the number of occurrences of A^* before the rth occurrence of A. By this formulation $Y_r = Z_r - r$ we can obtain the probability mass function of Y by a simple transformation of the variables. The probability mass function of Y becomes

$$\Pr(Y_r = y) = \binom{r+y-1}{y} p^r(1-p)^y \qquad \text{for } y = 0, 1, 2, \ldots. \tag{5.133}$$

Remark 5.8 (The name of the distribution)
The negative binomial distribution has got its name from the relationship

$$\binom{r+y-1}{y} = (-1)^y \binom{-r}{y} = (-1)^y \frac{(-r)(-r-1)\ \cdots (-r-y-1)}{y(y-1)\ \cdots 2 \cdot 1},$$

which defines the binomial coefficient with negative integers. □

The mean value of Y_r is (see Problem 5.32)

$$E(Y_r) = \sum_{y=0}^{\infty} y \binom{r+y-1}{y} p^r(1-p)^y = \frac{r(1-p)}{p}. \tag{5.134}$$

Because $Y_r = Z_r - r$, the mean value of Z is

$$E(Z_r) = E(Y_r) + r = \frac{r}{p}. \tag{5.135}$$

5.8.5 The Homogeneous Poisson Process

The HPP[8] is a stochastic process that may be used to model occurrences of a specific event E in the course of a given time interval. The event E may, for example, be a failure, or an accident. The HPP is discussed in more detail in Chapter 10. The following conditions are assumed to be fulfilled:

(1) The event E may occur at any time in the interval, and the probability of E occurring in the interval $(t, t + \Delta t]$ is independent of t and may be written as $\lambda\Delta t + o(\Delta t)$,[9] where λ is a positive constant.
(2) The probability of more than one event E in the interval $(t, t + \Delta t]$ is $o(\Delta t)$.
(3) Let $(t_{11}, t_{12}], (t_{21}, t_{22}], \dots$ be any sequence of disjoint intervals in the time period in question. Then the events "E occurs in $(t_{j1}, t_{j2}]$," for $j = 1, 2, \dots$, are independent.

Without loss of generality, we let $t = 0$ be the starting point of the process.

Let $N(t)$ be the number of times the event E occurs during the interval $(0, t]$. The stochastic process $\{N(t), t \geq 0\}$ is then an HPP with rate λ. The rate λ is sometimes called the *intensity* of the process, or the *frequency* of events E. A consequence of assumption (1) is that the rate of events E is constant and does not change with time. The HPP can therefore not be used to model processes where the rate of events changes with time, for example processes that have a long-term trend, or are exposed to seasonal variations.

The time t may be measured as calendar time or operational time. In many cases, several subprocesses are running in parallel and the time t must then be measured as total, or accumulated, time in service. This is, for example the case when we observe failures in a population of repairable items.

The probability that E occurs exactly n times in the time interval $(0, t]$ is

$$\Pr(N(t) = n) = \frac{(\lambda t)^n}{n!}\, e^{-\lambda t} \quad \text{for } n = 0, 1, 2, \dots. \tag{5.136}$$

The distribution (5.136) is called the *Poisson distribution*, and we sometimes write $N(t) \sim \text{Poisson}(\lambda t)$. When we observe the occurrence of events E in an interval $(s, s + t]$, the probability that E occurs exactly n times in $(s, s + t]$ is

$$\Pr(N(s + t) - N(s) = n) = \frac{(\lambda t)^n}{n!}\, e^{-\lambda t} \quad \text{for } n = 0, 1, 2, \dots,$$

[8] Named after the French mathematician Siméon Denis Poisson (1781–1840).

[9] $o(\Delta t)$ denotes a function of Δt with the property that $\lim_{\Delta t \to 0} \frac{o\,(\Delta t)}{\Delta t} = 0$.

that is, the same probability as we found in (5.132). The important quantity is therefore the length t of the time interval we are observing the process, not when this interval starts.

Consider a time interval $(t, t + \Delta t]$ that is so short that at most one event E can occur within the interval. Because $\lambda \Delta t$ is so small that $(\lambda \Delta t)^x$, for $x = 2, 3, \ldots$, become negligible, the probability of observing one event E in the interval is approximately

$$\Pr(N(\Delta t) = 1) = \lambda \Delta t \; e^{-\lambda \Delta t} \approx \lambda \Delta t(1 - \lambda \Delta t) \approx \lambda \Delta t, \tag{5.137}$$

which is in line with the above assumption (1) for the HPP.

The mean number of events in $(0, t]$ is

$$E[N(t)] = \sum_{n=0}^{\infty} n \Pr(N(t) = n) = \lambda t, \tag{5.138}$$

and the variance is

$$\text{var}[N(t)] = \lambda t. \tag{5.139}$$

From Eq. (5.138), the parameter λ may be written as $\lambda = E(N(t))/t$, that is, the mean number of events per time unit. This is why λ is called the *rate* of the HPP. When the event E is a failure, λ is called the *ROCOF* of the HPP.

A natural unbiased estimator of λ is

$$\hat{\lambda} = \frac{N(t)}{t} = \frac{\text{No. of events observed in an interval of length } t}{\text{Length } t \text{ of the interval}}. \tag{5.140}$$

Let T_1 be the time when E occurs for the first time, and let $F_{T_1}(t)$ be the distribution function of T_1. Because the event $(T_1 > t)$ means that no event has occurred in the interval $(0, t]$, we get

$$\begin{aligned} F_{T_1}(t) &= \Pr(T_1 \le t) = 1 - \Pr(T_1 > t) \\ &= 1 - \Pr(N(t) = 0) = 1 - e^{-\lambda t} \qquad \text{for } \; t \ge 0. \end{aligned} \tag{5.141}$$

The time T_1 to the first E is seen to be *exponentially* distributed with parameter λ. It may be shown, see Chapter 10, that the times between events, $T_1, T_2, \ldots$ are independent, and exponentially distributed with parameter λ. The times between events $T_1, T_2, \ldots$ are called the *interoccurrence times* of the process.

Example 5.15 (Repairable item)

Consider a repairable item that is put into operation at time $t = 0$. The first failure (event E) occurs at time T_1. When the item has failed, it is replaced with a new item of the same type. The replacement time is so short that it can be neglected. The second failure occurs at time T_2, and so on. We thus get a sequence of failure times $T_1, T_2, \ldots$. The number of failures, $N(t)$ in the time interval $(0, t]$ is assumed to be Poisson distributed with rate (ROCOF) λ. The interoccurrence times $T_1, T_2, \dots$ are then independent and exponentially distributed with failure rate λ. Observe the important difference in meaning between the two concepts "failure rate" and ROCOF. □

Let us consider an HPP with rate λ and assume that we are interested in determining the distribution of the time S_k, where E occurs for the kth time (k is accordingly an integer). We let t be an arbitrarily chosen point of time on the positive real axis. The event $(T_k > t)$ is then obviously synonymous with the event that E is occurring at most $(k-1)$ times in the time interval $(0, t]$. Therefore,

$$\Pr(S_k > t) = \Pr(N(t) \le k-1) = \sum_{j=0}^{k-1} \frac{(\lambda t)^j}{j!}\, e^{-\lambda t}.$$

Hence,

$$F_{S_k}(t) = 1 - \sum_{j=0}^{k-1} \frac{(\lambda t)^j}{j!}\, e^{-\lambda t}, \tag{5.142}$$

where $F_{S_k}(t)$ is the distribution function for S_k. The probability density function $f_{S_k}(t)$ is obtained by differentiating $F_{S_k}(t)$ with respect to t:

$$\begin{aligned} f_{S_k}(t) &= -\sum_{j=1}^{k-1} \frac{j\lambda(\lambda t)^{j-1}}{j!}\, e^{-\lambda t} + \lambda \sum_{j=0}^{k-1} \frac{(\lambda t)^j}{j!}\, e^{-\lambda t} \\ &= \lambda e^{-\lambda t} \left(\sum_{j=0}^{k-1} \frac{(\lambda t)^j}{j!} - \sum_{j=1}^{k-1} \frac{(\lambda t)^{j-1}}{(j-1)!} \right) \\ &= \lambda e^{-\lambda t} \left(\sum_{j=0}^{k-1} \frac{(\lambda t)^j}{j!} - \sum_{j=0}^{k-2} \frac{(\lambda t)^j}{j!} \right) \\ &= \frac{\lambda}{(k-1)!} (\lambda t)^{k-1}\, e^{-\lambda t} \qquad \text{for } t \ge 0 \qquad \text{and } \lambda > 0, \end{aligned} \tag{5.143}$$

where k is a positive integer. This distribution is the *gamma* distribution with parameters k and λ. The gamma distribution is discussed in Section 5.4.2. We can therefore conclude that the waiting time until the kth occurrence of E in an HPP with rate λ, is gamma distributed (k, λ). HPP is discussed in more detail in Chapter 10.

5.9 Classes of Time-to-Failure Distributions

This section defines four categories or families of time-to-failure distribution.

5.9.1 IFR and DFR Distributions

We say that a distribution $F(t)$ is an *increasing failure rate* (IFR) distribution if its failure rate function $z(t)$ increases as a function of t, for $t > 0$.[10]

A more general definition is to say that $F(t)$ is an IFR distribution if $-\log R(t)$ is a convex function of t. This is because a differentiable convex function has an increasing derivative.

Similarly, a distribution $F(t)$ is said to be a *DFR* distribution if $z(t)$ decreases as a function of t, for $t > 0$, or more generally when $-\log R(t)$ is a concave function of t. This follows because a differentiable concave function has a decreasing derivative.

In the following examples, we consider some common time-to-failure distributions and check whether they are IFR, DFR, or neither of these.

Example 5.16 (The uniform distribution over $(0, b)$)
Let T be uniformly distributed over $(0, b]$. Then

$$F(t) = \frac{t}{b} \quad \text{for } 0 < t \le b$$

$$f(t) = \frac{1}{b} \quad \text{for } 0 < t \le b.$$

Hence,

$$z(t) = \frac{1/b}{1 - (t/b)} = \frac{1}{b - t} \quad \text{for } 0 < t \le b \tag{5.144}$$

is strictly increasing for $0 < t \le b$. The uniform distribution is accordingly IFR.

The same conclusion follows by considering $-\log R(t)$, which in this case becomes $-\log[1 - (t/b)]$ and hence is convex for $0 < t \le b$. □

Example 5.17 (The exponential distribution)
Let T be exponentially distributed with probability density

$$f(t) = \lambda e^{-\lambda t} \qquad \text{for } t > 0.$$

Then

$$z(t) = \lambda \qquad \text{for } t > 0,$$

$z(t)$ is thus constant, that is, both nonincreasing and nondecreasing.

The exponential distribution therefore belongs to the IFR family as well as the DFR family. Alternatively, one could argue that $-\log R(t) = \lambda t$, that is convex and concave as well. □

[10] In this section "increasing" and "decreasing" are used in place of "nondecreasing" and "nonincreasing," respectively.

Hence, the families of IFR distributions and DFR distributions are not disjoint. The exponential distribution can be shown to be the only continuous distribution that belongs to both families (see Barlow and Proschan 1975, p. 73).

Example 5.18 (The Weibull distribution)
The distribution function of the Weibull distribution with parameters $\alpha > 0$ and $\theta > 0$ is given by

$$F(t) = 1 - \exp\left[-\left(\frac{t}{\theta}\right)^{\alpha}\right] \qquad \text{for } t \geq 0.$$

It follows that

$$-\log R(t) = -\log\left(\exp\left[-\left(\frac{t}{\theta}\right)^{\alpha}\right]\right) = \left(\frac{t}{\theta}\right)^{\alpha}. \tag{5.145}$$

Because $(t/\theta)^{\alpha}$ is convex in t when $\alpha \geq 1$ and concave in t when $\alpha \leq 1$, the Weibull distribution is IFR for $\alpha \geq 1$ and DFR for $\alpha \leq 1$. For $\alpha = 1$, the distribution is "reduced" to an exponential distribution with failure rate $\lambda = 1/\theta$, and hence is IFR as well as DFR. □

Example 5.19 (The gamma distribution)
The gamma distribution is defined by the probability density

$$f(t) = \frac{\lambda}{\Gamma(\alpha)}(\lambda t)^{\alpha-1}e^{-\lambda t} \qquad \text{for } t > 0,$$

where $\alpha > 0$ and $\lambda > 0$. To determine whether the gamma distribution (α, λ) is IFR, DFR, or neither of these, we consider the failure rate function.

$$z(t) = \frac{[\lambda(\lambda t)^{\alpha-1}e^{-\lambda t}]/\Gamma(\alpha)}{\int_t^{\infty}[\lambda(\lambda u)^{\alpha-1}e^{-\lambda u}]/\Gamma(\alpha)\, du}.$$

Dividing the denominator by the numerator yields

$$z(t)^{-1} = \int_t^{\infty}\left(\frac{u}{t}\right)^{\alpha-1}e^{-\lambda(u-t)}\, du.$$

Introducing $v = (u - t)$ as a new variable of integration gives

$$z(t)^{-1} = \int_0^{\infty}\left(1 + \frac{v}{t}\right)^{\alpha-1}e^{-\lambda v}\, dv. \tag{5.146}$$

First suppose that $\alpha \geq 1$. Then $[1 + (v/t)]^{\alpha-1}$ is nonincreasing in t. Accordingly, the integrand is a decreasing function of t. Thus, $z(t)^{-1}$ is decreasing in t. When $\alpha \geq 1$, $z(t)$ is in other words increasing in t, and the gamma distribution (α, λ) is IFR. This is, in particular, the case when α is an integer (the Erlangian distribution).

Next suppose $\alpha \leq 1$. Then by an analogous argument $z(t)$ is decreasing in t, which means that the gamma distribution (α, λ) is DFR.

For $\alpha = 1$, the gamma distribution (α, λ) is reduced to an exponential distribution with parameter λ. □

The plot of the failure rate function in Figure 5.28 for a lognormal distribution indicates that this distribution is neither IFR nor DFR.

If time-to-failure distribution is DFR and continuous, $z(t) = f(t)/[1 - F(t)]$ must be decreasing. Knowing that $1 - F(t)$ is decreasing in t, then $f(t)$ must decrease by at least as much as $1 - F(t)$ in order for $z(t)$ to be decreasing. These arguments lead to the useful result: If a continuous time-to-failure distribution is to be DFR, its probability density $f(t)$ must be nonincreasing.

5.9.2 IFRA and DFRA Distributions

Chapter 6 shows that the time-to-failure distribution of a system of components is not necessarily IFR even if all the components have IFR distributions. We therefore introduce a less demanding class of distributions and say that the distribution $F(t)$ is an *increasing failure rate average* (IFRA) distribution if the average value of its failure rate function $z(t)$ increases in the sense that

$$\frac{1}{t}\int_0^t z(u)\,du \qquad \text{increases as a function of } t.$$

A more general definition is to say that $F(t)$ is an IFRA distribution if

$$-\frac{1}{t}\log R(t) \qquad \text{increases with } t \geq 0.$$

Similarly, the distribution $F(t)$ is said to be a *decreasing failure rate average* (DFRA) distribution if the average failure rate decreases with time, or slightly more general that $[-\log R(t)]/t$ decreases as a function of t.

Let $t_1 \leq t_2$ and assume that $F(t)$ is an IFR distribution. This implies that $-\log R(t)$ is a convex function of t. For $t = 0$, we have $R(0) = 1$ and hence, $-\log R(0) = 0$. If we draw the convex curve $-\log R(t)$, we immediately see that

$$-\log R(t_1) \leq \frac{t_1}{t_2}[-\log R(t_2)],$$

which implies that

$$\frac{1}{t_1}\int_0^{t_1} z(u)\,du \leq \frac{1}{t_2}\int_0^{t_2} z(u)\,du,$$

and we have shown that if $F(t)$ is IFR, it is also IFRA. To show that if $F(t)$ is DFR implies that it is also DFRA is done by similar arguments.

5.9.3 NBU and NWU Distributions

A distribution $F(t)$ is said to be a *new better than used* (NBU) distribution if

$$R(t \mid x) \leq R(t) \qquad \text{for } t \geq 0, x \geq 0, \tag{5.147}$$

where $R(t \mid x) = \Pr(T > t + x \mid T > x)$ is the conditional survivor function that is introduced in Section 5.3.3. Equation (5.147) may also be written

$$\Pr(T > t + x \mid T > x) = \frac{\Pr(T > t)}{\Pr(T > x)} \leq \Pr(T > t),$$

which implies that

$$\Pr(T > t + x) \leq \Pr(T > t) \Pr(T > x). \tag{5.148}$$

If an item with NBU distribution is to work for a period of length $t + x$, the reliability would increase if we replace the item at some time x during this interval.

Similarly, $F(t)$ is said to be a *new worse than used* (NWU) distribution if

$$R(t \mid x) \geq R(t) \qquad \text{for} \;\; t \geq 0, x \geq 0.$$

For an item with NWU distribution that should work for a period of length $t + x$, it would be stupid to replace the item with a new one.

5.9.4 NBUE and NWUE Distributions

The MRL of an item at age x was defined in Section 5.3.6 as

$$\text{MRL}(x) = \int_0^\infty R(t \mid x)\, dt. \tag{5.149}$$

When $x = 0$, we start out with a new item and consequently MRL(0) = MTTF.

Definition 5.1 (New better/worse than used in expectation)
A time-to-failure distribution $F(t)$ is said to be a *new better than used in expectation* (NBUE) distribution if

(1) F has a finite mean μ
(2) $\text{MRL}(x) \leq \mu$ for $x \geq 0$.

A time-to-failure distribution $F(t)$ is said to be a *new worse than used in expectation* (NWUE) distribution if

(1) F has a finite mean μ
(2) $\text{MRL}(x) \geq \mu$ for $x \geq 0$.

□

5.9.5 Some Implications

The families of time-to-failure distributions presented above are further discussed, for example by Barlow and Proschan (1975) and Gertsbakh (1989) who proves the following chain of implications:

$$\begin{array}{ccccccc}
\text{IFR} & \Rightarrow & \text{IFRA} & \Rightarrow & \text{NBU} & \Rightarrow & \text{NBUE} \\
\text{DFR} & \Rightarrow & \text{DFRA} & \Rightarrow & \text{NWU} & \Rightarrow & \text{NWUE}
\end{array}$$

Table 5.3 Summary of time-to-failure distributions and parameters.

Distribution	**Probability density** $f(t)$	**Survivor function** $R(t)$	**Failure rate** $z(t)$	**MTTF**
Exponential	$\lambda e^{-\lambda t}$	$e^{-\lambda t}$	λ	$1/\lambda$
Gamma	$\frac{\lambda}{\Gamma(k)}\,(\lambda t)^{k-1}e^{-\lambda t}$	$\sum_{x=0}^{k-1}\frac{(\lambda t)^x}{x!}\;e^{-\lambda t}$	$\frac{f(t)}{R(t)}$	k/λ
Weibull	$\alpha\lambda(\lambda t)^{\alpha-1}e^{-(\lambda t)^{\alpha}}$	$e^{-(\lambda t)^{\alpha}}$	$\alpha\lambda(\lambda t)^{\alpha-1}$	$\frac{1}{\lambda}\Gamma\left(\frac{1}{\alpha}+1\right)$
Lognormal	$\frac{1}{\sqrt{2\pi}}\frac{1}{\tau}\frac{1}{t}e^{-(\log t - \nu)^2/2\tau^2}$	$\Phi\left(\frac{\nu-\log t}{\tau}\right)$	$\frac{f(t)}{R(t)}$	$e^{\nu+\tau^2/2}$

5.10 Summary of Time-to-Failure Distributions

A number of time-to-failure distributions have been introduced in this chapter. Some characteristics of the main distributions are presented in Table 5.3 to provide a brief reference.

5.11 Problems

5.1 Show that to say "the item has a constant failure rate" is equivalent to saying "the time-to-failure of the item is exponentially distributed."

5.2 An item with time-to-failure T has constant failure rate

$$z(t) = \lambda = 3.5 \times 10^{-6}\ \mathrm{h}^{-1}.$$

(a) Determine the probability that the item survives a period of six months in continuous operation without failure.
(b) Find the MTTF of the item.
(c) Find the probability that the item fails in the interval (t_1, t_2), where $t_1 =$ 16 months and $t_2 = 17$ months.

5.3 A machine with constant failure rate λ survives a period of 4000 hours without failure, with probability 0.95.
(a) Determine the failure rate λ.
(b) Find the probability that the machine survives 5000 hours without failure.

(c) Determine the probability that the machine fails within 5000 hours, when you know that the machine was functioning at 3500 hours.

5.4 A safety valve is assumed to have constant failure rate with respect to all failure modes. A study has shown that the total MTTF of the valve is 2450 days. The safety valve is in continuous operation, and the failure modes are assumed to occur independent of each other.

(a) Determine the total failure rate of the safety valve.

(b) Determine the probability that the safety valve survives a period of three months without any failure.

(c) 48% of all failures are assumed to be *critical* failure modes. Determine the mean time to a critical failure, $\text{MTTF}_{\text{crit}}$.

5.5 The time-to-failure T of an item is assumed to have an exponential distribution with failure rate λ. Show that the rth moment of T is

$$E(T^r) = \frac{\Gamma(r+1)}{\lambda^r}. \tag{5.150}$$

5.6 Let T_1 and T_2 be two independent times-to-failure with constant failure rates λ_1 and λ_2, respectively. Let $T = T_1 + T_2$.

(a) Show that the survivor function of T is

$$R(t) = \Pr(T > t) = \frac{1}{\lambda_2 - \lambda_1}(\lambda_2 \, e^{-\lambda_1 t} - \lambda_1 \, e^{-\lambda_2 t}) \qquad \text{for } \lambda_1 \neq \lambda_2.$$

(b) Find the corresponding failure rate function $z(t)$, and make a sketch of $z(t)$ as a function of t for selected values of λ_1 and λ_2.

5.7 Show that $f(t) \leq z(t)$ for all $t \geq 0$ and for all life distributions.

5.8 Let X be a random variable with a binomial distribution with parameters (n, p). Find $E(X)$ and var(X).

5.9 Let N be a random variable with value set $0, 1, \ldots$. Show that

$$E(N) = \sum_{n=1}^{\infty} \Pr(N \geq n).$$

5.10 Consider the time-to-failure T with cumulative failure rate function $Z(t)$ and show that the transformed variable $Z(T) \sim \exp(1)$.

5.11 Let Z have a geometric distribution with probability p, and determine

(a) The mean value, $E(Z)$

(b) The variance, var(Z)

(c) The conditional probability, $\Pr(Z > z + x \mid Z > x)$. Describe the result you get by words.

5.12 Let N_1 and N_2 be independent Poisson random variables with $E(N_1) = \lambda_1$ and $E(N_2) = \lambda_2$.

(a) Determine the distribution of $N_1 + N_2$.

(b) Determine the conditional distribution of N_1, given that $N_1 + N_2 = n$.

5.13 Let T_1 and T_2 be independent and gamma distributed with parameters (k_1, λ) and (k_2, λ), respectively. Show that $T_1 + T_2$ has a gamma distribution with parameters $(k_1 + k_2, \lambda)$. Explain why we sometimes say that the gamma distribution is "closed under addition."

5.14 A component with time-to-failure T has failure rate function

$$z(t) = kt \qquad \text{for } t > 0 \text{ and } k > 0.$$

(a) Determine the probability that the component survives 200 hours, when $k = 2.0 \times 10^{-6}\ \text{h}^{-1}$.

(b) Determine the MTTF of the component when $k = 2.0 \times 10^{-6}\ \text{h}^{-1}$.

(c) Determine the probability that a component which is functioning after 200 hours, is still functioning after 400 hours, when $k = 2.0 \times 10^{-6}\ \text{h}^{-1}$.

(d) Does this distribution belong to any of the distribution classes described in Chapter ?

(e) Find the mode and the median of this distribution.

5.15 A component with time-to-failure T has failure rate function

$$z(t) = \lambda_0 + \alpha t \qquad \text{for } t > 0, \lambda_0 > 0, \text{ and } \alpha > 0.$$

(a) Determine the survivor function $R(t)$ of the component.

(b) Determine the MTTF of the component.

(c) Find the probability that the component will survive 2 MTTF when we know that it was functioning at MTTF.

(d) Give a physical interpretation of this model.

5.16 A component with time-to-failure T has failure rate function

$$z(t) = \frac{t}{1+t} \qquad \text{for } t > 0.$$

(a) Make a sketch of the failure rate function.

(b) Determine the corresponding probability density function $f(t)$.
(c) Determine the MTTF of the component.
(d) Does this distribution belong to any of the distribution classes described in Chapter ?

5.17 The failure rate function of an item is $z(t) = t^{-\frac{1}{2}}$. Derive:
(a) The probability density function, $f(t)$,
(b) The survivor function, $R(t)$,
(c) The mean time-to-failure, MTTF, and
(d) The variance of the time-to-failure, T, var(T).

5.18 The time-to-failure T of a component is assumed to be uniformly distributed over (a, b), i.e. $T \sim$ unif(a, b). The probability density is thus

$$f(t) = \frac{1}{b - a} \qquad \text{for } a < t < b.$$

Derive the corresponding survivor function $R(t)$ and failure rate function $z(t)$. Draw a sketch of $z(t)$.

5.19 The time-to-failure T of a component has probability density $f(t)$ as shown in Figure 5.33.
(a) Determine c such that $f(t)$ is a valid probability density.
(b) Derive the corresponding survivor function $R(t)$.
(c) Derive the corresponding failure rate function $z(t)$ and make a sketch of $z(t)$.

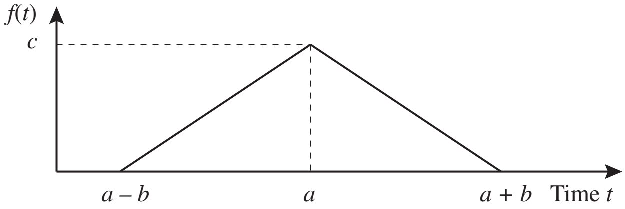


Figure 5.33 Probability density (Problem 5.19).

5.20 let T be the time-to-failure of an item. Assume that we know that MTTF = 10 000 hours and the standard deviation, SD = 2500 hours.
(a) Assume that T has a Weibull distribution with parameters θ and α and determine θ and α.
(b) Assume that T has a lognormal distribution with parameters ν and τ and determine the parameters.

(c) Determine the MRL(t) at time $t = 9000$ hours for both distributions and make comments to the difference between the two MRL-values obtained.

5.21 Let T have a Weibull distribution with shape parameter α and scale parameter λ. Show that the variable $(\lambda T)^{\alpha}$ has an exponential distribution with rate 1.

5.22 The time-to-failure T of an item is assumed to have a Weibull distribution with scale parameter λ and shape parameter α. Show that the rth moment of T is

$$E(T^r) = \frac{1}{\lambda^r}\,\Gamma\left(\frac{r}{\alpha}+1\right).$$

5.23 The time-to-failure T of an item is assumed to have a Weibull distribution with scale parameter $\lambda = 5.0 \times 10^{-5}\ \text{h}^{-1}$ and shape parameter $\alpha = 1.5$. Compute MTTF and var(T).

5.24 Let T have a three parameter Weibull distribution (α, λ, ξ) with probability density

$$f(t) = \frac{d}{dt}F(t) = \alpha\lambda[\lambda(t-\xi)]^{\alpha-1}\, e^{-[\lambda(t-\xi)]^{\alpha}} \qquad \text{for } t > \xi.$$

(a) Show that the density is unimodal if $\alpha > 1$. Also show that the density decreases monotonically with t if $\alpha < 1$.

(b) Show that the failure rate function is $\alpha\lambda[\lambda(t-\xi)]^{\alpha-1}$ for $t > \xi$, and hence is increasing, constant, and decreasing with t, Respectively, as $\alpha > 1, \alpha = 1$, and $\alpha < 1$.

5.25 Let T be Weibull distributed with parameter λ and α. Show that $Y = \log T$ has a Type I asymptotic distribution of the smallest extreme. Find the mode and the scale parameter of this distribution.

5.26 Assume the time-to-failure T to be lognormally distributed such that $Y = \log T$ is $\mathcal{N}(\nu, \tau^2)$. Show that

$$E(T) = e^{\nu+\tau^2/2},$$
$$\text{var}(T) = e^{2\nu}(e^{2\tau^2} - e^{\tau^2}),$$

and that the variance may be written as

$$\text{var}(T) = [E(T)]^2(e^{\tau^2} - 1).$$

5.27 Let $z(t)$ be the failure rate function of the lognormal distribution. Show that $z(0) = 0$, that $z(t)$ increases to a maximum, and then decreases with $z(t) \to 0$ as $t \to \infty$.

5.28 Let T be lognormally distributed with parameters ν and τ^2. Show that $1/T$ is lognormally distributed with parameters $-\nu$ and τ^2.

5.29 Show that the median t_{med} of the lognormal distribution is e^{ν}. Compute k such that $\Pr(t_{\text{med}}/k \le T \le kt_{\text{med}}) = 0.90$.

5.30 Reconsider the item in Example 5.1 with survivor function

$$R(t) = \frac{1}{(0.2t+1)^2} \qquad \text{for } t \ge 0,$$

where the time t is measured in months.

(a) Find the mean residual lifetime (MRL) of the item at age $t = 3$ months.

(b) Make a sketch of MRL(t) as a function of the age t.

5.31 Consider an item with survivor function $R(t)$. Show that the MTTF of the item can be written as

$$\text{MTTF} = \int_0^t R(u)\,du + R(t)\,\text{MRL}(t).$$

Explain the meaning of this formula.

5.32 Derive the mean value of the negative binomially distributed variable Y in (5.133). Show and justify all the steps used to derive $E(Y)$.

5.33 Let $N(t)$ be an HPP with rate $\lambda > 0$. Assume that $n \ge 1$ events have been observed during a specified time interval of length t.

(a) Find the conditional distribution $\Pr(N(t^*) = k \mid N(t) = n)$ for $k = 0, 1, \dots, n$ and $0 < t^* < t$.

(b) Determine the mean and the variance of this distribution.

5.34 The time-to-failure, T, has survivor function $R(t)$. Show that if $E(T^r) < \infty$, then

$$E(T^r) = \int_0^\infty rt^{r-1}R(t)\,dt \qquad \text{for } r = 1, 2, \dots .$$

5.35 Consider an item with time-to-failure T and failure rate function $z(t)$. Show that

$$\Pr(T > t_2 \mid T > t_1) = e^{-\int_{t_1}^{t_2} z(u)\,du} \qquad \text{for } t_2 > t_1.$$

5.36 Consider a component with time-to-failure T, with increasing failure rate (IFR) distribution, and MTTF $= \mu$. Show that

$$R(t) \geq e^{-t/\mu} \quad \text{for } 0 < t < \mu.$$

5.37 Derive the Laplace transform of the survivor function $R(t)$ of the exponential distribution with failure rate λ and use the Laplace transform to determine the MTTF of this distribution.

5.38 Let $F(t)$ be the distribution of the time-to-failure T. Assume $F(t)$ to be strictly increasing. Show that

(a) $F(T)$ is uniformly distributed over $(0, 1)$.

(b) if $U \sim \text{unif}(0, 1)$ random variable, then $F^{-1}(U)$ has distribution F, where $F^{-1}(y)$ is that value of x such that $F(x) = y$.

5.39 Prove that

$$\int_0^{t_0} z(t)\, dt \to \infty \quad \text{when } t_0 \to \infty.$$

5.40 Consider a structure of n independent components with failure rates $\lambda_1, \lambda_2, \dots, \lambda_n$, respectively. Show that the probability that component i fails first is

$$\frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j}.$$

5.41 A component may fail due to two different causes, excessive stresses and aging. A large number of this type of components have been tested. It has been shown that the time to failure T_1 caused by excessive stresses is exponentially distributed with density function

$$f_1(t) = \lambda_1 e^{-\lambda_1 t} \quad \text{for } t \geq 0,$$

whereas the time-to-failure T_2 caused by aging has density function

$$f_2(t) = \frac{1}{\Gamma(k)}\, \lambda_2 (\lambda_2 t)^{k-1} e^{-\lambda_2 t} \quad \text{for } t \geq 0.$$

(a) Describe the rationale behind using

$$f(t) = p f_1(t) + (1 - p) f_2(t) \quad \text{for } t \geq 0,$$

as the probability density function for the time-to-failure T of the component.

(b) Explain the meaning of p in this model.

(c) Let $p = 0.1$, $\lambda_1 = \lambda_2$, and $k = 5$, and determine the failure rate function $z(t)$ corresponding to T. Calculate $z(t)$ for some selected values of t, e.g. $t = 0, \frac{1}{2}, 1, 2, \dots$, and make a sketch of $z(t)$.

5.42 A component may fail due to two different causes, A and B. It has been shown that the time-to-failure T_A caused by A is exponentially distributed with density function

$$f_A(t) = \lambda_A e^{-\lambda_A t} \qquad \text{for } t \geq 0,$$

whereas the time-to-failure T_B caused by B has density function

$$f_B(t) = \lambda_B e^{-\lambda_B t} \qquad \text{for } t \geq 0.$$

(a) Describe the rationale behind using

$$f(t) = pf_A(t) + (1 - p)f_B(t) \qquad \text{for } t \geq 0,$$

as the probability density function for the time-to-failure T of the component.

(b) Explain the meaning of p in this model.

(c) Show that a component with probability density $f(t)$ has a decreasing failure rate (DFR) function.

5.43 Let T_1 and T_2 be independent times-to-failure with failure rate functions $z_1(t)$ and $z_2(t)$, respectively. Show that

$$\Pr(T_1 < T_2 \mid \min\{T_1, T_2\} = t) = \frac{z_1(t)}{z_1(t) + z_2(t)}.$$

5.44 Assume that Z_r has a negative binomial distribution with probability mass function given by (5.131) for specified values of p and r. When $r = 1$, we write $Z_r = Z_1$.

(a) Find $E(Z_r)$ and $\text{var}(Z_r)$.

(b) Verify that $E(Z_r) = rE(Z_1)$ and $\text{var}(Z_r) = r\,\text{var}(Z_1)$ and explain why this is a realistic result.

5.45 Show that

(a) If $X_1, X_2, \dots, X_r$ are independent variables with geometric distribution with parameter p, then $Z_r = \sum_{i=1}^{r} X_i$ has negative binomial distribution with parameters (p, r).

(b) If $Z_1, Z_2, \dots, Z_n$ are independent variables and that Z_i has a negative binomial distribution with parameters (p, r_i) for $i = 1, 2, \dots, n$, then $Z = \sum_{i=1}^{n} Z_{r_i}$ has a negative binomial distribution with parameters $(p, \sum_{i=1}^{n} r_i)$.

5.46 Let X be a random variable with uniform distribution, $X \sim \text{unif}(0, 1)$. Show that the random variable $T = \frac{1}{\lambda} \log(1 - X)$ has distribution $\exp(\lambda)$.

References

Ascher, H. and Feingold, H. (1984). *Repairable Systems Reliability; Modeling, Inference, Misconceptions, and Their Causes*. New York: Marcel Dekker.

Barlow, R.E. and Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing, Probability Models*. New York: Holt, Rinehart, and Winston.

Birnbaum, Z.W. and Saunders, S.C. (1969). A new family of life distributions. *Journal of Applied Probability* 6 (2): 319–327.

Chhikara, R.S. and Folks, J.L. (1989). *The Inverse Gaussian Distribution: Theory, Methodology, and Applications*. New York: Marcel Dekker.

Cramér, C.H. (1946). *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press.

Gertsbakh, I. (1989). *Statistical Reliability Theory*. New York: Marcel Dekker.

Gumbel, E.J. (1958). *Statistics of Extremes*. New York: Columbia University Press.

Johnson, N.L. and Kotz, S. (1970). *Distributions in Statistics. Continuous Univariate Distributions*, vol 1–2. Boston, MA: Hougton Mifflin.

Kapur, K.C. and Lamberson, L.R. (1977). *Reliability in Engineering Design*. Hoboken, NJ: Wiley.

Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*. Hoboken, NJ: Wiley.

Lloyd, D.K. and Lipow, M. (1962). *Reliability: Management, Methods, and Mathematics*. Englewood Cliffs, NJ: Prentice-Hall.

Mann, N.R., Schafer, R.E., and Singpurwalla, N.D. (1974). *Methods for Statistical Analysis of Reliability and Lifetime Data*. Hoboken, NJ: Wiley.

McCool, J.I. (2012). *Using the Weibull Distribution; Reliability, Modeling, and Inference*. Hoboken, NJ: Wiley.

Murthy, D.N.P., Xie, M., and Jiang, R. (2003). *Weibull Models*. Hoboken, NJ: Wiley.

NUREG-75/014 (1975). Reactor Safety: An Assessment of Accident Risk in U.S. Commercial Nuclear Power Plants. *Report NUREG-75/014*. Washington, DC: U.S. Nuclear Regulatory Commission.

O’Connor, A.N., Modarres, M., and Mosleh, A. (2016). *Probability Distributions Used in Reliability Engineering, Center for Risk and Reliability*. College Park, MD: University of Maryland.

Rinne, H. (2014). *The Hazard Rate: Theory and Inference*, *Monograph*. Giessen: Justus-Liebig-Universität.

Sweet, A.L. (1990). On the hazard rate of the lognormal distribution. *IEEE Transactions on Reliability* 39: 325–328.

6

System Reliability Analysis

6.1 Introduction

Chapter 4 deals with the structural relationships between a system and its components and shows how a *deterministic* model of the structure can be established, using a *reliability block diagram* (RBD) or a fault tree. Whether or not a given component will be in a failed state after t time units, can usually not be predicted with certainty. Rather, when studying the occurrence of such failures, one looks for statistical regularity. Hence, it seems reasonable to interpret the state variables of the n components at time t as *random variables*. We denote the random state variables by $X_1(t), X_2(t), \dots, X_n(t)$. The *state vector* and the structure function are denoted by $\boldsymbol{X}(t) = [X_1(t), X_2(t), \dots, X_n(t)]$ and $\phi[\boldsymbol{X}(t)]$, respectively. The following probabilities are of interest:

$$\Pr(X_i(t) = 1) = p_i(t), \qquad \text{for } i = 1, 2, \dots, n \tag{6.1}$$

$$\Pr(\phi[\boldsymbol{X}(t)] = 1) = p_{\text{S}}(t). \tag{6.2}$$

Here, $p_i(t)$ is called the *reliability* of component i and $p_S(t)$ the *system reliability* at time t.

This chapter is delimited to the study of systems where failures of individual components can be interpreted as *independent* events. This implies that the state variables at time t, $X_1(t), X_2(t), \dots, X_n(t)$ are stochastically independent. Unfortunately, independence is often assumed just to "simplify" the analysis, but may sometimes be unrealistic. This problem is discussed in more detail in Chapter 8.

In the first part of this chapter, we consider *nonrepairable* components and systems that are discarded the first time they fail. In that case (6.1) and (6.2) correspond to the *survivor function* of component i and of the system, respectively.

A *repairable* system is a system where at least one of its components is *repaired* or *replaced* upon failure. Repairable components and systems that are considered until the first failure only are treated as nonrepairable. The main reliability metrics

System Reliability Theory: Models, Statistical Methods, and Applications, Third Edition.
Marvin Rausand, Anne Barros, and Arnljot Høyland.

Companion Website: www.wiley.com/go/SystemReliabilityTheory3e

for repairable systems are introduced in Section 6.5, and some simple approaches to reliability analysis of repairable systems are outlined in subsequent sections. A more thorough treatment of repairable (or maintained) systems is provided in Chapter 9. Preventive maintenance is dealt with in Chapter 12.

6.1.1 Assumptions

Throughout this chapter, the following assumptions apply:

(1) All the structures studied are coherent (coherent systems are introduced in Section 4.7.)
(2) Each item (component, subsystem, and system) has two possible states, 1 or 0. Depending on the system and the type of analysis, these states are referred to as functioning or failed, up or down, and true or false.
(3) The system is put into operation at time $t = 0$ with all components in a functioning state.
(4) The operating context is unchanged during the time period considered.
(5) All components are independent, both with respect to failures and repairs.
(6) No preventive maintenance is carried out. The only maintenance action considered is repair of a failure that has occurred. After the repair, the component is considered to be as-good-as-new.
(7) Failure and repair data for the components (or basic events) are known with sufficient accuracy.
(8) Systems are sometimes referred to as structures, and vice versa.

6.2 System Reliability

Because the state variables $X_i(t)$ for $i = 1, 2, \dots, n$ are binary, then

$$\begin{aligned} E[X_i(t)] &= 0 \cdot \Pr(X_i(t) = 0) + 1 \cdot \Pr(X_i(t) = 1) \\ &= p_i(t) \qquad \text{for } \ i = 1, 2, \dots, n. \end{aligned} \tag{6.3}$$

This applies for both nonrepairable and repairable systems. Similarly, the system reliability at time t is

$$p_S(t) = E(\phi[\boldsymbol{X}(t)]\ ). \tag{6.4}$$

It can be shown (see Problem 6.1) that when the components are independent, the system reliability, $p_S(t)$, is a function of the $p_i(t)$'s only. Hence, $p_S(t)$ may be written

$$p_S(t) = h[p_1(t), p_2(t), \dots, p_n(t)] = h[\boldsymbol{p}(t)]. \tag{6.5}$$

Unless stated otherwise, we use $h(\cdot)$ to express system reliability in situations *where the components are independent*. Now, let us determine the reliability of some simple structures.

6.2.1 Reliability of Series Structures

The structure function of a series structure of order n is from (4.4)

$$\phi[\boldsymbol{X}(t)] = \prod_{i=1}^{n} X_i(t).$$

Because $X_1(t), X_2(t), \dots, X_n(t)$ are independent, the system reliability is

$$h[\boldsymbol{p}(t)] = E(\phi[\mathbf{X}(t)]) = E\left(\prod_{i=1}^{n} X_i(t)\right) = \prod_{i=1}^{n} E[X_i(t)] = \prod_{i=1}^{n} p_i(t). \tag{6.6}$$

Observe that

$$h[\boldsymbol{p}(t)] \le \min_i \{p_i(t)\}.$$

In other words, a series structure is *at most* as reliable as the *least* reliable component.

Example 6.1 (Series structure)
Consider a series structure of three independent components. At a specified point of time t, the component reliabilities are $p_1 = 0.95$, $p_2 = 0.97$, and $p_3 = 0, 94$. The system reliability at time t is from (6.6)

$$p_S = h(\boldsymbol{p}) = p_1 p_2 p_3 = 0.95 \cdot 0.97 \cdot 0.94 \approx 0.866. \qquad \square$$

If all the components have the same reliability $p(t)$, then the system reliability of a series structure of order n is

$$p_S(t) = p(t)^n.$$

If, for example $n = 10$ and $p(t) = 0.950$, then

$$p_S(t) = 0.950^{10} \approx 0.599.$$

The system reliability of a series structure is low already when $n = 10$, even when the component reliability is 0.950.

The reliability $h[\boldsymbol{p}(t))]$ of a series structure may also be determined by a more direct approach, without using the structure function. Let $E_i(t)$ be the event that component i is functioning at time t. The probability of this event is $\Pr[E_i(t)] = p_i(t)$. Because a series structure is functioning if, and only if, all its components are functioning, and because the components are independent, the reliability of the series structure is

$$\begin{aligned} h[\boldsymbol{p}(t)] &= \Pr[E_1(t) \cap E_2(t) \cap \cdots \cap E_n(t)] \\ &= \Pr[E_1(t)] \Pr[E_2(t)] \cdots \Pr[E_n(t)] = \prod_{i=1}^{n} p_i(t), \end{aligned}$$

which is the same result we got in (6.6) by using the structure function.

6.2.2 Reliability of Parallel Structures

The structure function of a parallel structure of order n is from (4.5)

$$\phi[\boldsymbol{X}(t)] = \coprod_{i=1}^{n} X_i(t) = 1 - \prod_{i=1}^{n}[1 - X_i(t)].$$

Hence,

$$h[\boldsymbol{p}(t)] = E(\phi[\boldsymbol{X}(t)]) = 1 - \prod_{i=1}^{n}(1 - E[X_i(t)]) = 1 - \prod_{i=1}^{n}[1 - p_i(t)]. \tag{6.7}$$

This expression may alternatively be written as follows:

$$h[\boldsymbol{p}(t)] = \coprod_{i=1}^{n} p_i(t).$$

Observe that

$$h[\boldsymbol{p}(t)] \geq \max_i\{p_i(t)\}.$$

Example 6.2 (Parallel structure)
Consider a parallel structure of three independent components. At a specified time t, the component reliabilities are $p_1 = 0.95$, $p_2 = 0.97$, and $p_3 = 0, 94$. The system reliability at time t is from (6.7)

$$p_S = h(\boldsymbol{p}) = 1 - (1 - p_1)(1 - p_2)(1 - p_3) = 1 - 0.05 \cdot 0.03 \cdot 0.06 \approx 0.99991.$$

□

If all the components have the same reliability $p(t)$, then the system reliability at time t of a parallel structure of order n is

$$p_S(t) = 1 - [1 - p(t)]^n.$$

As for the series structure, the reliability $h[\boldsymbol{p}(t)]$ of a parallel structure may be determined by a more direct approach, without using the structure function. Let $E_i^*(t)$ be the event that component i is in a failed state at time t. The probability of this event is $\Pr[E_i^*(t)] = 1 - p_i(t)$. Because a parallel structure is in a failed state if, and only if, all its components are in a failed state, and because the components are independent, we have that

$$\begin{aligned} 1 - h[\boldsymbol{p}(t)] &= \Pr[E_1^*(t) \cap E_2^*(t) \cap \cdots \cap E_n^*(t)] \\ &= \Pr[E_1^*(t)]\Pr[(E_2^*(t)] \cdots \Pr[E_n^*(t)] = \prod_{i=1}^{n}[1 - p_i(t)] \end{aligned}$$

and, therefore, in accordance with (6.7)

$$h[\boldsymbol{p}(t)] = 1 - \prod_{i=1}^{n}[1 - p_i(t)].$$

This direct approach is feasible for series and parallel structures, but is cumbersome for more complicated structures, in which case, the approach by using structure functions is much more suitable.

6.2.3 Reliability of *k*oo*n* Structures

The structure function of a $k$oo$n$:G structure has a structure function (see Eq. (4.7))

$$\phi[\boldsymbol{X}(t)] = \begin{cases} 1 & \text{for} \quad \sum_{i=1}^{n} X_i(t) \geq k \\ 0 & \text{for} \quad \sum_{i=1}^{n} X_i(t) < k \end{cases}. \tag{6.8}$$

To simplify the notation, we omit the explicit reference to G ("good") and write $k$oo$n$ instead of $k$oo$n$:G. In cases where a $k$oo$n$:F structure is studied, we always include the F ("failed").

Consider a $k$oo$n$ structure, where all the n components have identical reliabilities $p_i(t) = p(t)$ for $i = 1, 2, \dots, n$. Because it is assumed that failures of individual components are independent events, then at a given time t, $Y(t) = \sum_{i=1}^{n} X_i(t)$ is binomially distributed $[n, p(t)]$

$$\Pr(Y(t) = y) = \binom{n}{y} p(t)^y [1 - p(t)]^{n-y} \qquad \text{for } y = 0, 1, \dots, n.$$

The reliability of a $k$oo$n$ structure of components with identical reliabilities is hence

$$p_S(t) = \Pr(Y(t) \geq k) = \sum_{y=k}^{n} \binom{n}{y} p(t)^y [1 - p(t)]^{n-y}. \tag{6.9}$$

Example 6.3 (2oo3 structure)

The 2oo3 structure is shown in Figure 2.14, and the structure function is from (4.8)

$$\phi[\boldsymbol{X}(t)] = X_1(t)X_2(t) + X_1(t)X_3(t) + X_2(t)X_3(t) - 2X_1(t)X_2(t)X_3(t).$$

When the three components are independent, the reliability of the 2oo3 structure is

$$p_S(t) = p_1(t)p_2(t) + p_1(t)p_3(t) + p_2(t)p_3(t) - 2p_1(t)p_2(t)p_3(t).$$

When all the three components have the same reliability, $p_i(t) = p(t)$ for $i = 1, 2, 3$, then

$$p_S(t) = 3p(t)^2 - 2p(t)^3.$$

In this example, the structure function is used to find the reliability $p_S(t)$ of the 2oo3 structure. Observe that the same result is obtained by using (6.9). □

Finally, let us see how the system reliability of a more complicated structure can be determined.

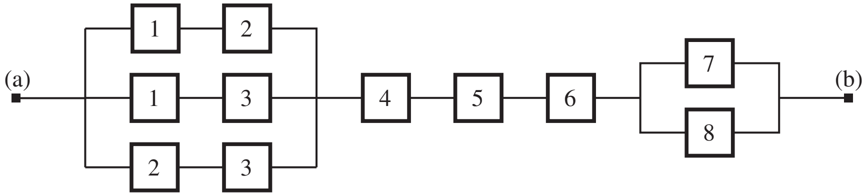


Figure 6.1 RBD of a simplified automatic alarm system for gas leakage.

Example 6.4 (Alarm system for gas leakage)
Figure 6.1 shows an RBD of a simplified automatic alarm system for gas leakage. In the case of gas leakage, “connection” is established between a and b so that at least one of the alarm bells (7 and 8) will start ringing. The system has three independent gas detectors (1, 2, and 3) that are connected to a 2oo3 voting unit (4); that is, at least two detectors must indicate gas leakage before an alarm is raised. Component 5 is a power supply, and component 6 is a relay.

Consider the system at a given time t. To simplify the notation, we omit the explicit reference to the time t. The structure function of the system is

$$\phi(\boldsymbol{X}) = (X_1X_2 + X_1X_3 + X_2X_3 - 2X_1X_2X_3)\ (X_4X_5X_6)\ (X_7 + X_8 - X_7X_8).$$

If the component reliability at time t of component i is denoted by $p_i,\ i = 1, 2, \dots, 8$, and $X_1, X_2, \dots, X_8$ are independent, then the system reliability at time t_0 is

$$p_S = (p_1p_2 + p_1p_3 + p_2p_3 - 2p_1p_2p_3)\ p_4p_5p_6\ (p_7 + p_8 - p_7p_8). \qquad \square$$

6.2.4 Pivotal Decomposition

By pivotal (or Shannon) decomposition, the structure function $\phi[\boldsymbol{X}(t)]$ at time t may be written as (see Eq. (4.21))

$$\begin{aligned}\phi[\boldsymbol{X}(t)] &= X_i(t)\phi[1_i, \boldsymbol{X}(t)] + [1 - X_i(t)]\phi[0_i, \boldsymbol{X}(t)]\\ &= X_i(t)(\phi[1_i, \boldsymbol{X}(t)] - \phi[0_i, \boldsymbol{X}(t)]) + \phi[0_i, \boldsymbol{X}(t)]\end{aligned}$$

When the components are independent, the system reliability becomes

$$h[\boldsymbol{p}(t)] = p_i(t)E(\phi[(1_i, \boldsymbol{X}(t)]) + [1 - p_i(t)]E(\phi[0_i, \boldsymbol{X}(t)]).$$

Let $h[1_i, \boldsymbol{p}(t)] = E(\phi[1_i, \boldsymbol{X}(t)])$ and $h[0_i, \boldsymbol{p}(t)] = E(\phi[0_i, \boldsymbol{X}(t)])$, which implies that

$$\begin{aligned}h[\boldsymbol{p}(t)] &= p_i(t)h[1_i, \boldsymbol{p}(t)] + [1 - p_i(t)]h[0_i, \boldsymbol{p}(t)]\\ &= p_i(t)(h[1_i, \boldsymbol{p}(t)] - h[0_i, \boldsymbol{p}(t)]) + h[0_i, \boldsymbol{p}(t)].\end{aligned} \tag{6.10}$$

Observe that the system reliability $h[\boldsymbol{p}(t)]$ is a linear function of $p_i(t)$ when all the other component reliabilities are kept constant. Also observe that (6.10) follows directly from the *law of total probability* in basic probability theory (see box):

Law of Total Probability

Let $\mathcal{S}$ be the sample space of an experiment and let $C_1, C_2, \dots, C_n$ be a partition of $\mathcal{S}$ such that $\mathcal{S} = \bigcup_{i=1}^{n} C_i$ and $C_i \cap C_j = \emptyset$ for all $i \neq j$. Let A be an event in $\mathcal{S}$. The probability of A is

$$\Pr(A) = \Pr(A \cap \mathcal{S}) = \Pr\left(A \cap \bigcup_{i=1}^{n} C_i\right) = \Pr\left(\bigcup_{i=1}^{n} A \cap C_i\right) = \sum_{i=1}^{n} \Pr(A \cap C_i).$$

The last equality follows because $C_i, C_2, \dots, C_n$ are mutually exclusive, and therefore, $(A \cap C_1), (A \cap C_2), \dots, (A \cap C_n)$ are also mutually exclusive. We now use the definition of conditional probability $\Pr(A \cap C_i) = \Pr(A \mid C_i)\Pr(C_i)$ to arrive at the *law of total probability*

$$\Pr(A) = \sum_{i=1}^{n} \Pr(A \mid C_i)\ \Pr(C_i). \tag{6.11}$$

6.2.5 Critical Component

Component i is said to be *critical* for a (coherent) system if the rest of the components are in such states that the system is functioning when component i is functioning and fails when component i fails. This means that the rest of the system has state $[\cdot_i, \boldsymbol{X}(t)]$ such that $\phi[1_i, \boldsymbol{X}(t)] = 1$ and $\phi[0_i, \boldsymbol{X}(t)] = 0$. Because the system is coherent and has binary states, component i is critical when

$$\phi[1_i, \boldsymbol{X}(t)] - \phi[0_i, \boldsymbol{X}(t)] = 1.$$

Because $\phi[1_i, \boldsymbol{X}(t)] - \phi[0_i, \boldsymbol{X}(t)]$ can take only the values 0 and 1, the probability that the system comes in such a state that component i is critical is

$$\begin{aligned}\Pr(\text{Component } i \text{ is critical}) &= \Pr(\phi[1_i, \boldsymbol{X}(t)] - \phi[0_i, \boldsymbol{X}(t)] = 1)\\ &= E(\phi[1_i, \boldsymbol{X}(t)] - \phi[0_i, \boldsymbol{X}(t)])\\ &= h[1_i, \boldsymbol{p}(t)] - h[0_i, \boldsymbol{p}(t)].\end{aligned} \tag{6.12}$$

Component i is said to *cause system failure* if component i is critical and then fails. Critical components are discussed further in Chapter 7.

6.3 Nonrepairable Systems

This section deals solely with nonrepairable systems. As explained in Section 6.1, the component reliability and the *survivor function* coincide for nonrepairable components:

$$p_i(t) = R_i(t) \qquad \text{for } i = 1, 2, \ldots, n.$$

6.3.1 Nonrepairable Series Structures

According to (6.6) the survivor function of a nonrepairable series structure consisting of independent components, is

$$R_S(t) = \prod_{i=1}^{n} R_i(t). \tag{6.13}$$

Furthermore, according to (5.11)

$$R_i(t) = e^{-\int_0^t z_i(u)\, du}, \tag{6.14}$$

where $z_i(t)$ is the failure rate function of component i at time t.

Inserting (6.14) into (6.13) yields

$$R_S(t) = \prod_{i=1}^{n} e^{-\int_0^t z_i(u)\, du} = e^{-\int_0^t \sum_{i=1}^n z_i(u)\, du} = e^{-\int_0^t z_S(u)\, du}.$$

The failure rate function $z_S(t)$ of a series structure (of independent components) is, hence, equal to the sum of the failure rate functions of the individual components:

$$z_S(t) = \sum_{i=1}^{n} z_i(t). \tag{6.15}$$

The mean time-to-failure (MTTF) of this series structure is

$$\text{MTTF}_S = \int_0^\infty R_S(t)\, dt = \int_0^\infty e^{-\int_0^t \sum_{i=1}^n z_i(u)\, du}\, dt. \tag{6.16}$$

Example 6.5 (Series structure with constant failure rates)

Consider a series structure on n (independent) components with constant failure rates λ_i, for $i = 1, 2, \ldots, n$. The survivor function of the series structure is

$$R_S(t) = e^{-(\sum_{i=1}^n \lambda_i)\, t}. \tag{6.17}$$

The failure rate of the series structure is constant and equal to

$$\lambda_S = \sum_{i=1}^{n} \lambda_i \tag{6.18}$$

and the mean to failure is

$$\mathrm{MTTF}_S = \int_0^\infty R_S(t)\,dt = \frac{1}{\sum_{i=1}^n \lambda_i}. \tag{6.19}$$

When all the failure rates are equal, $\lambda_i = \lambda$ for $i = 1, 2, \ldots, n$, the failure rate of the series structure is $\lambda_S = n\lambda$, and the MTTF of the series structure is MTTF $= 1/(n\lambda)$. □

Example 6.6 (Series structure with Weibull-distributed times-to-failure) Consider a series structure with n independent components. The time-to-failure of component i has a Weibull distribution with common shape parameter α and scale parameter θ_i, for $i = 1, 2, \ldots, n$. The survivor function of the series structure is from (6.16)

$$R_S(t) = \prod_{i=1}^n e^{-\left(\frac{t}{\theta_i}\right)^\alpha} = e^{-\left[\left(\sum_{i=1}^n \left(\frac{1}{\theta_i}\right)^\alpha\right)^{1/\alpha} t\right]^\alpha}.$$

Introducing $\theta_0 = \left(\sum_{i=1}^n \theta_i^{-\alpha}\right)^{-1/\alpha}$, the survivor function $R_S(t)$ can be written as

$$R_S(t) = e^{-\left(\frac{t}{\theta_0}\right)^\alpha}. \tag{6.20}$$

The time-to-failure of the series structure is therefore Weibull distributed with shape parameter α and scale parameter $\theta_0 = \left(\sum_{i=1}^n \theta_i^{-\alpha}\right)^{-1/\alpha}$. □

Example 6.7 (Bathtub curve obtained by three Weibull distributions) Consider a series structure of $n = 3$ independent components. Component 1 has a decreasing failure rate, for example a Weibull distributed time-to-failure with shape parameter $\alpha < 1$. Component 2 has a constant failure rate, whereas component 3 has an increasing failure rate, for example a Weibull distributed time-to-failure with shape parameter $\alpha > 2$. The failure rates of the three components are illustrated in Figure 6.2. The failure rate function of the series structure is from (6.15) the sum of the three individual failure rate functions, and is illustrated by the fully drawn line in Figure 6.2. The failure rate function of the series structure is seen to have a bathtub shape. A bathtub-shaped failure rate of a component may therefore be obtained by replacing the component by three independent and virtual components in series; one with decreasing failure rate function, one with constant failure rate, and one with increasing failure rate function. □

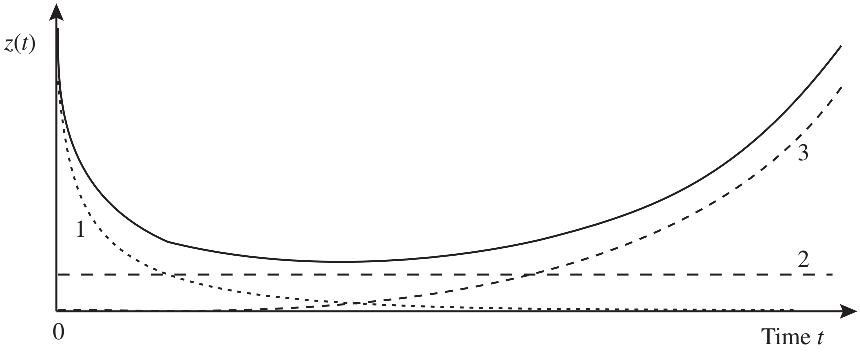


Figure 6.2 The failure rate function of a series structure of three independent components, where component 1 has decreasing failure rate, component 2 has constant failure rate, and component 3 has increasing failure rate.

6.3.2 Nonrepairable Parallel Structures

From (6.7), the survivor function of a nonrepairable parallel structure of independent components is

$$R_S(t) = 1 - \prod_{i=1}^{n} [1 - R_i(t)]. \tag{6.21}$$

To determine the survivor function for general time-to-failure distributions is complicated, and we therefore suffice with assuming that all components have constant failure rates. When all the components have constant failure rates $z_i(t) = \lambda_i$, for $i = 1, 2, \dots, n$, then

$$R_S(t) = 1 - \prod_{i=1}^{n} (1 - e^{-\lambda_i t}). \tag{6.22}$$

Parallel Structure of Identical Components

Consider a parallel structure of n independent components of the same type with constant failure rate λ. The survivor function of the parallel structure is

$$R_S(t) = 1 - (1 - e^{-\lambda t})^n. \tag{6.23}$$

The parallel structure may be illustrated by the *transition diagram* in Figure 6.3.[1] In the first state (i.e. circle), all the n components are functioning. When the first of these fails, with rate $n\lambda$, the structure moves to the second state with $n - 1$ components functioning. After some time, one of these $n - 1$ components fails, with rate $(n - 1)\lambda$, and the structure moves to the next state, with $n - 2$ components

[1] Transition diagrams are discussed in detail in Chapter 11.

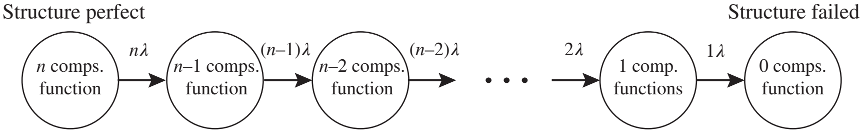


Figure 6.3 Transition diagram for a parallel structure of n independent and identical components with failure rate λ.

functioning, and so on, until all the n components have failed and the structure fails. The mean time to the first transition is $1/n\lambda$, the mean time to the second transition is $1/(n-1)\lambda$, the mean time to the third is $1/(n-2)\lambda$, and so on. The mean time to structure failure is hence

$$\begin{aligned}\text{MTTF}_S &= \frac{1}{n\lambda} + \frac{1}{(n-1)\lambda} + \cdots + \frac{1}{2\lambda} + \frac{1}{\lambda} \\ &= \frac{1}{\lambda}\left(1 + \frac{1}{2} + \cdots + \frac{1}{n-1} + \frac{1}{n}\right) = \frac{1}{\lambda}\sum_{x=1}^{n}\frac{1}{x} \end{aligned} \tag{6.24}$$

Remark 6.1 (An alternative derivation)
Equation (6.24) is formally derived in Section 6.3.5 for a $k$oo$n$ structure of n identical and independent components with failure rate λ. The MTTF$_S$ of a parallel structure is listed in the first row of Table 6.2 for some selected values of n. □

Example 6.8 (Parallel structure of two identical components)
Consider a parallel structure of two independent and identical components with failure rate λ. The survivor function is

$$R_S(t) = 2e^{-\lambda t} - e^{-2\lambda t}. \tag{6.25}$$

The probability density function of the time-to-failure of the parallel structure is

$$f_S(t) = -R_S'(t) = 2\lambda e^{-\lambda t} - 2\lambda e^{-2\lambda t}.$$

The mode of the distribution is the value of t that maximizes $f_S(t)$

$$t_{\text{mode}} = \frac{\ln 2}{\lambda}.$$

The median life of the parallel structure is

$$t_{\text{med}} = R_S^{-1}(0.5) \approx \frac{1.228}{\lambda}.$$

The MTTF is

$$\text{MTTF}_S = \int_0^\infty R_S(t)\,dt = \frac{3}{2\lambda}. \tag{6.26}$$

Observe that the MTTF of a parallel structure of two independent components is 50% longer than for a single component.

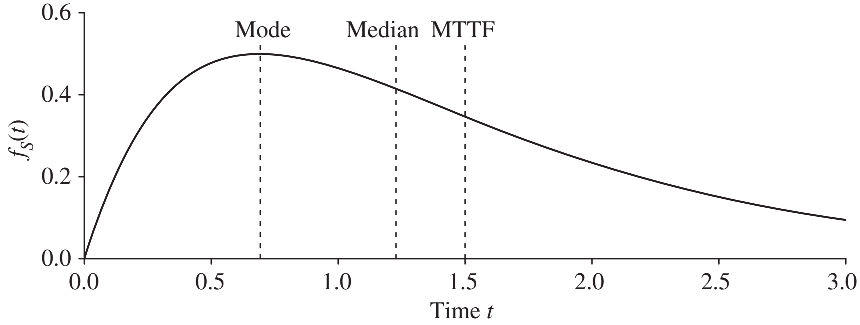


Figure 6.4 The probability density function of a parallel structure with two independent and identical components with failure rate $\lambda = 1$, together with its mode, median, and MTTF.

The probability density $f_S(t)$ of the parallel structure, together with its mode, median, and MTTF$_S$ are illustrated in Figure 6.4. The mean residual lifetime of the parallel structure at age t is

$$\text{MRL}_S(t) = \frac{1}{R_S(t)} \int_t^\infty R_S(x)\, dx = \frac{1}{2\lambda}\, \frac{4 - e^{-\lambda t}}{2 - e^{-\lambda t}}.$$

Observe that $\lim_{t\to\infty} \text{MRL}_S(t) = 1/\lambda$. Because the two components are nonrepairable, and one of them will fail first, we will sooner or later be left with only one component. When one of the components has failed, the mean residual lifetime of the structure is equal to the mean residual lifetime of the remaining component. Because the failure rate is constant, the mean residual lifetime of the remaining component is equal to its MTTF, MTTF = $1/\lambda$. □

Example 6.9 reveals that the time-to-failure T_S of a parallel structure is *not* exponentially distributed, even if all components have exponentially distributed times-to-failure.

Example 6.9 (Parallel structure of two different components)
Consider a parallel structure of two nonrepairable components with constant failure rates λ_1 and λ_2, respectively.

The survivor function of the structure is

$$\begin{aligned} R_S(t) &= 1 - (1 - e^{-\lambda_1 t})(1 - e^{-\lambda_2 t}) \\ &= e^{-\lambda_1 t} + e^{-\lambda_2 t} - e^{-(\lambda_1+\lambda_2)t}. \end{aligned} \tag{6.27}$$

The MTTF of the parallel structure is

$$\text{MTTF}_S = \int_0^\infty R_S(t)\, dt = \frac{1}{\lambda_1} + \frac{1}{\lambda_2} - \frac{1}{\lambda_1 + \lambda_2}. \tag{6.28}$$

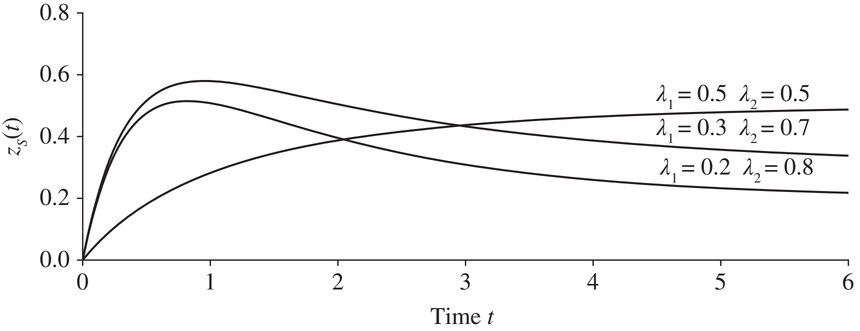


Figure 6.5 The failure rate for a parallel structure of two independent components for selected values of λ_1 and λ_2 $(\lambda_1 + \lambda_2 = 1)$.

The corresponding failure rate function is

$$z_S(t) = -\frac{R_S'(t)}{R_S(t)}.$$

Hence,

$$z_S(t) = \frac{\lambda_1 e^{-\lambda_1 t} + \lambda_2 e^{-\lambda_2 t} - (\lambda_1 + \lambda_2) e^{-(\lambda_1+\lambda_2)t}}{e^{-\lambda_1 t} + e^{-\lambda_2 t} - e^{-(\lambda_1+\lambda_2)t}}. \tag{6.29}$$

Figure 6.5 shows $z_S(t)$ for selected combinations of λ_1 and λ_2, such that $\lambda_1 + \lambda_2 = 1$. Observe that when $\lambda_1 \neq \lambda_2$, the failure rate function $z_S(t)$ increases up to a maximum at a time t_0, and then decreases for $t \geq t_0$ down to min $\{\lambda_1, \lambda_2\}$. □

Example 6.10 (Parallel structure of Weibull-distributed components)
Consider a parallel structure of two independent and identical components with Weibull (α, θ) life distribution. The survivor function of the structure is

$$R_S(t) = 2e^{-\left(\frac{t}{\theta}\right)^\alpha} - e^{-2\left(\frac{t}{\theta}\right)^\alpha}.$$

The last term in this expression may be written as

$$e^{-2\left(\frac{t}{\theta}\right)^\alpha} = e^{-\left(\frac{t}{2^{-1/\alpha}\theta}\right)^\alpha},$$

which is the survivor function of a Weibull (α, θ_1) distribution, where the scale parameter $\theta_1 = 2^{-1/\alpha}\theta$.

The MTTF of the parallel structure can now be determined as

$$\text{MTTF}_S = 2\theta\, \Gamma\left(1 + \frac{1}{\alpha}\right) - \theta_1 \Gamma\left(1 + \frac{1}{\alpha}\right) = (2\theta - \theta_1)\, \Gamma\left(1 + \frac{1}{\alpha}\right),$$

and the failure rate function of the parallel structure is next obtained from

$$z_S(t) = \frac{f_S(t)}{R_S(t)} = \frac{-R_S'(t)}{R_S(t)}.$$

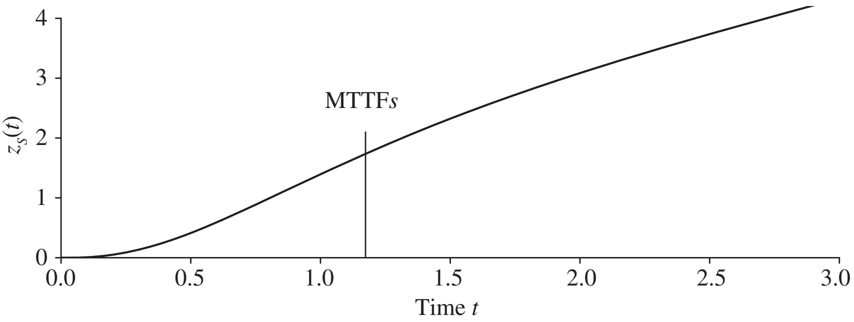


Figure 6.6 Failure rate function for a parallel structure of two independent and identical components that are Weibull distributed with $\alpha = 1.8$ and $\theta = 1$.

The failure rate function for the parallel structure is illustrated in Figure 6.6 for $\alpha = 1.8$ and $\theta = 1$. The corresponding R script is

```
t  <- seq(0, 3, length=300)  # time axis
a  <- 1.8   # the Weibull shape parameter
th   <- 1  # the Weibull scale parameter
th1   <- 2^(-1/a)*th # the transformed scale parameter
m  <- (2*th-th1)*gamma(1+1/a) # the MTTF
x  <- 2*dweibull(t,a,th, log=FALSE) -
dweibull(t,a,th1, log=FALSE)
y  <- 1+ pweibull(t,a,th1, log=FALSE) -
2*pweibull(t,a,th, log=FALSE)
z  <-x/y
plot(t, z, type="l")
segments(m,0,m,2.1)
text(m,2.6, expression(MTTF[S]))
```

□

6.3.3 Nonrepairable 2oo3 Structures

The survivor function of a 2oo3 structure of independent components can, by using Example 6.2, be written as

$$R_S(t) = R_1(t)R_2(t) + R_1(t)R_3(t) + R_2(t)R_3(t) - 2R_1(t)R_2(t)R_3(t).$$

In the special case, where all the three components have the common constant failure rate λ, then

$$R_S(t) = 3\, e^{-2\lambda t} - 2e^{-3\lambda t}. \tag{6.30}$$

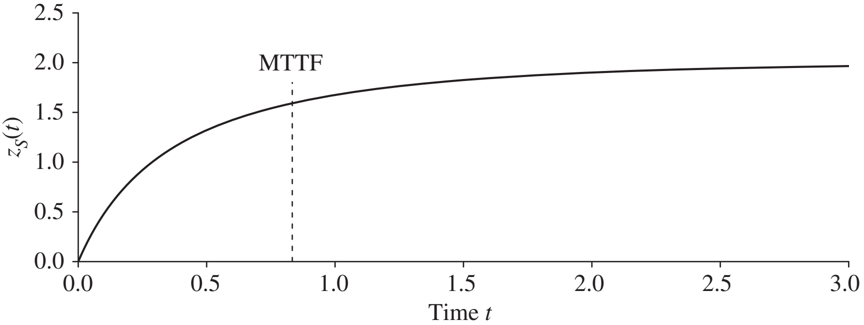


Figure 6.7 The failure rate function $z_S(t)$ for a 2oo3 structure of independent and identical components with failure rate $\lambda = 1$. The MTTF_S of the structure is indicated.

The failure rate function of this 2oo3 structure is

$$z_S(t) = \frac{-R_S'(t)}{R_S(t)} = \frac{6\lambda(e^{-2\lambda t} - e^{-3\lambda t})}{3e^{-2\lambda t} - 2e^{-3\lambda t}}. \tag{6.31}$$

The failure rate function $z_S(t)$ is shown in Figure 6.7.

Observe that $\lim_{t\to\infty} z_C(t) = 2\lambda$ (see Problem 6.9). The MTTF of this 2oo3 structure is

$$\text{MTTF}_S = \int_0^\infty R_S(t)\,dt = \frac{3}{2\lambda} - \frac{2}{3\lambda} = \frac{5}{6}\,\frac{1}{\lambda}. \tag{6.32}$$

Observe that the MTTF of a 2oo3 structure is shorter than the MTTF of a single component.

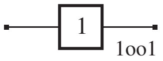
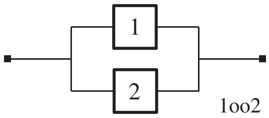
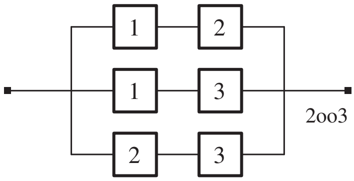
6.3.4 A Brief Comparison

Let us compare the three simple structures:

(1) A single component;
(2) A parallel structure of two identical components;
(3) A 2oo3 structure of identical components.

All the components are assumed to be independent with a common constant failure rate λ. A brief comparison of the three structures is presented in Table 6.1. Observe that a single component has a higher MTTF than the 2oo3 structure. The survivor functions of the three simple structures are compared in Figure 6.8. The introduction of a 2oo3 structure instead of a single component, hence, reduces the MTTF by about 16%, but the 2oo3 structure has a significantly higher survival probability in the interval $(0, t]$ for $t < \ln 2/\lambda$.

Table 6.1 A brief comparison of the structures (1), (2), and (3).

	Survivor function	**Mean time-to-failure**
System	$R_S(t)$	**MTTF**
	$e^{-\lambda t}$	$\frac{1}{\lambda}$
	$2e^{-\lambda t} - e^{-2\lambda t}$	$\frac{3}{2}\frac{1}{\lambda}$
	$3e^{-2\lambda t} - 2e^{-3\lambda t}$	$\frac{5}{6}\frac{1}{\lambda}$

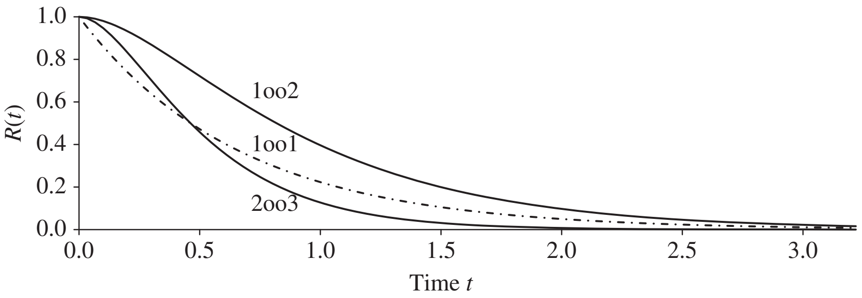


Figure 6.8 The survivor functions of the three structures in Table 6.1 ($\lambda = 5$).

6.3.5 Nonrepairable *k*oo*n* Structures

Assume that we have a *k*oo*n* structure of n identical and independent components with constant failure rate λ. The survivor function of the *k*oo*n* structure is from (6.8)

$$R_S(t) = \sum_{x=k}^{n} \binom{n}{x} e^{-\lambda t x}(1 - e^{-\lambda t})^{n-x}. \tag{6.33}$$

Table 6.2 MTTF of some *k*oo*n* structures of identical and independent components with constant failure rate λ.

k\n	1	2	3	4	5
1	$\frac{1}{\lambda}$	$\frac{3}{2\lambda}$	$\frac{11}{6\lambda}$	$\frac{25}{12\lambda}$	$\frac{137}{60\lambda}$
2	—	$\frac{1}{2\lambda}$	$\frac{5}{6\lambda}$	$\frac{13}{12\lambda}$	$\frac{77}{60\lambda}$
3	—	—	$\frac{1}{3\lambda}$	$\frac{7}{12\lambda}$	$\frac{47}{60\lambda}$
4	—	—	—	$\frac{1}{4\lambda}$	$\frac{9}{20\lambda}$
5	—	—	—	—	$\frac{1}{5\lambda}$

The mean time-to-failure is

$$\text{MTTF}_S = \int_0^\infty R_S(t)\,dt = \sum_{x=k}^{n} \binom{n}{x} \int_0^\infty e^{-\lambda t x}(1 - e^{-\lambda t})^{n-x}\,dt. \tag{6.34}$$

By introducing $v = e^{-\lambda t}$, we obtain by using the beta function

$$\begin{aligned} \text{MTTF}_S &= \sum_{x=k}^{n} \binom{n}{x} \frac{1}{\lambda} \int_0^1 v^{x-1}(1-v)^{n-x}\,dv \\ &= \sum_{x=k}^{n} \binom{n}{x} \frac{1}{\lambda} \frac{\Gamma(x)\Gamma(n-x+1)}{\Gamma(n+1)} \\ &= \frac{1}{\lambda} \sum_{x=k}^{n} \binom{n}{x} \frac{(x-1)!(n-x)!}{n!} = \frac{1}{\lambda} \sum_{x=k}^{n} \frac{1}{x}. \end{aligned} \tag{6.35}$$

The MTTF of some simple *k*oo*n* structures, computed by (6.35), are listed in Table 6.2. Observe that a 1oo*n* structure is a parallel structure, whereas a *n*oo*n* structure is a series structure.

6.4 Standby Redundancy

In some structures, single items (components, subsystems) may be of much greater importance for the system's ability to function than others. If, for example a single item is operating in series with the rest of the system, failure of this item leads to system failure. Two ways of ensuring higher system reliability in such situations are to (i) use items with very high reliability in these critical places in the system, or (ii) introduce *redundancy* in these places (i.e. introduce one or more reserve items). The type of redundancy obtained by replacing the important item with two or more items operating in parallel is called *active redundancy*. These items then share the load right from the start until one of them fails.

Reserve items may be kept in standby in such a way that the first of them is activated when the ordinary item fails, the second is activated when the first reserve item fails, and so on. If the reserve items carry no load and are not subject to deterioration in the waiting period before activation (and therefore cannot fail in this period), the redundancy is called *passive*. In the waiting period, such an item is said to be in *cold* standby. If the standby items carry a weak load or deteriorate in the waiting period (and therefore might fail in this period), the redundancy is called *partly loaded*. In the following sections, we illustrate these types of redundancy by considering some simple examples.

6.4.1 Passive Redundancy, Perfect Switching, No Repairs

Consider the standby system in Figure 6.9. The system functions in the following way: Item 1 is put into operation at time $t = 0$. When it fails, item 2 is activated. When it fails, item 3 is activated, and so forth. The item that is in operation is called the *active* item, whereas the items that are standing by ready to take over are called *standby* or *passive* items. When item n fails, the system fails.

We assume that the switch S functions perfectly and that items cannot fail when they are passive. Let T_i denote the time-to-failure of item i, for $i = 1, 2, \dots, n$. The time-to-failure, T_S, of the whole standby system is then

$$T_S = \sum_{i=1}^{n} T_i.$$

The mean time to system failure, MTTF_S, is obviously

$$\text{MTTF}_S = \sum_{i=1}^{n} \text{MTTF}_i,$$

where MTTF_i is the mean time-to-failure of item i, for $i = 1, 2, \dots, n$.

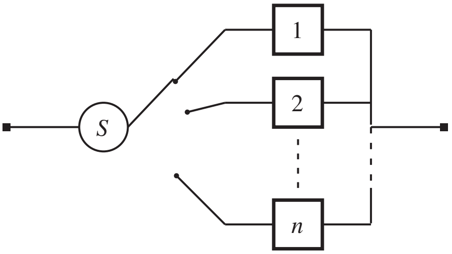


Figure 6.9 Standby system with n items.

The exact distribution of the time-to-failure T_S can only be determined in some very special cases. Such a special case occurs when $T_1, T_2, \dots, T_n$ are independent and exponentially distributed with failure rate λ. According to (5.142), T_S is gamma distributed with parameters n and λ. The survivor function of the system is then

$$R_S(t) = \sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!} e^{-\lambda t}. \tag{6.36}$$

If we have only one standby item, such that $n = 2$, the survivor function is

$$R_S(t) = e^{-\lambda t} + \frac{\lambda t}{1!} e^{-\lambda t} = (1 + \lambda t)\, e^{-\lambda t}. \tag{6.37}$$

If we have two standby items (i.e. $n = 3$), the survivor function is

$$R_S(t) = e^{-\lambda t} + \frac{\lambda t}{1!} e^{-\lambda t} + \frac{(\lambda t)^2}{2!} e^{-\lambda t} = \left(1 + \lambda t + \frac{(\lambda t)^2}{2}\right) e^{-\lambda t}. \tag{6.38}$$

If we are unable to determine the exact distribution of T_S, we have to be content with an approximate expression for the distribution. Assume, for example that the time-to-failure $T_1, T_2, \dots, T_n$ are independent and identically distributed with MTTF μ and variance σ^2. According to *central limit theorem* (see box), when $n \to \infty$, T_S is asymptotically normally distributed with mean $n\mu$ and variance $n\sigma^2$.

Central Limit Theorem

Let $X_1, X_2, \dots, X_n$ be a sequence of independent and identically distributed random variables with mean value $E(X_i) = \mu$ and variance $\text{var}(X_i) = \sigma^2 < \infty$, for $i = 1, 2, \dots, n$, and consider the sum $\sum_{i=1}^n X_i$. We know that $E\left(\sum_{i=1}^n X_i\right) = n\mu$ and $\text{var}\left(\sum_{i=1}^n X_i\right) = n\sigma^2$. The central limit theorem says that the sum $\sum_{i=1}^n X_i$ converges in distribution to a normal distribution when $n \to \infty$, such that

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1). \tag{6.39}$$

This means that, when n is large

$$\Pr\left(\sum_{i=1}^n X_i \le x\right) = \Pr\left(\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \le \frac{x - n\mu}{\sigma\sqrt{n}}\right) = \Phi\left(\frac{x - n\mu}{\sigma\sqrt{n}}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution $\mathcal{N}(0,1)$.

Using the central limit theory, the survivor function of the system is approximately

$$\begin{aligned}R_S(t) &= \Pr\left(\sum_{i=1}^{n} T_i > t\right) = 1 - \Pr\left(\sum_{i=1}^{n} T_i \le t\right)\\ &= 1 - \Pr\left(\frac{\sum_{i=1}^{n} T_i - n\mu}{\sigma\sqrt{n}} \le \frac{t - n\mu}{\sigma\sqrt{n}}\right) \approx \Phi\left(\frac{n\mu - t}{\sigma\sqrt{n}}\right).\end{aligned}$$

6.4.2 Cold Standby, Imperfect Switch, No Repairs

Here, we restrict ourselves to considering the simplest case with $n = 2$ items. Figure 6.10 shows a standby system with an active item (item 1) and an item in *cold* standby (item 2). The active item is under surveillance by a switch, which activates the standby item when the active item fails.

Furthermore, assume that the active item has constant failure rate λ_1. When the active item fails, the switch activates the standby item. The probability that this switching is successful is $1 - p$. The failure rate of item 2 in standby position is assumed to be negligible. When the standby item is activated, its failure rate is λ_2. The three items operate independently. No repairs are carried out. In addition, assume that the only way in which the switch S can fail is by not activating the standby item when the active item fails. In many practical applications, the switching is performed by a human operator. The probability p of unsuccessful activation of the standby item often includes the probability of not being able to start the standby item.

The system is able to survive the interval $(0, t]$ in two *disjoint* ways.

(1) Item 1 does *not* fail in $(0, t]$ (i.e. $T_1 > t$)
(2) Item 1 fails in a time interval $(\tau, \tau + d\tau]$, where $0 < \tau < t$. The switch S is able to activate item 2. Item 2 is activated at time τ and does not fail in $(\tau, t]$.

Let T_S denote the time-to-system failure. Events 1 and 2 are clearly disjoint. Hence, the survivor function of the system $R_S(t) = \Pr(T_S > t)$ is the sum of the probability of the two events.

The probability of event 1 is

$$\Pr(T_1 > t) = e^{-\lambda_1 t}.$$

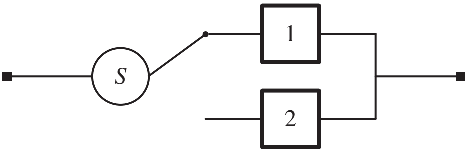


Figure 6.10 Standby system with 2 items.

Next, consider event 2: Item 1 fails in $(\tau, \tau + d\tau]$ with probability $f_1(\tau)\, d\tau = \lambda_1 e^{-\lambda_1 \tau}\, d\tau$. The switch S is able to activate item 2 with probability $(1-p)$.

Item 2 does *not* fail in $(\tau, t]$ with probability $e^{-\lambda_2(t-\tau)}$. Because item 1 may fail at any point of time τ in $(0, t]$, the survivor function of the system is when $\lambda_1 \neq \lambda_2$

$$\begin{aligned} R_S(t) &= e^{-\lambda_1 t} + \int_0^t (1-p)\, e^{-\lambda_2(t-\tau)} \lambda_1 e^{-\lambda_1 \tau}\, d\tau \\ &= e^{-\lambda_1 t} + (1-p)\lambda_1 e^{-\lambda_2 t} \int_0^t e^{-(\lambda_1 - \lambda_2)\tau}\, d\tau \\ &= e^{-\lambda_1 t} + \frac{(1-p)\lambda_1}{\lambda_1 - \lambda_2} e^{-\lambda_2 t} - \frac{(1-p)\lambda_1}{\lambda_1 - \lambda_2} e^{-\lambda_1 t}. \end{aligned} \tag{6.40}$$

When $\lambda_1 = \lambda_2 = \lambda$, we get

$$\begin{aligned} R_S(t) &= e^{-\lambda t} + \int_0^t (1-p) e^{-\lambda(t-\tau)} \lambda e^{-\lambda \tau}\, d\tau \\ &= e^{-\lambda t} + (1-p)\lambda e^{-\lambda t} \int_0^t d\tau \\ &= e^{-\lambda t} + (1-p)\lambda t e^{-\lambda t}. \end{aligned} \tag{6.41}$$

The MTTF_S for the system is

$$\begin{aligned} \text{MTTF}_S &= \int_0^\infty R_S(t)\, dt = \frac{1}{\lambda_1} + \frac{(1-p)\lambda_1}{\lambda_1 - \lambda_2} \left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1} \right) \\ &= \frac{1}{\lambda_1} + (1-p)\frac{1}{\lambda_2}. \end{aligned} \tag{6.42}$$

This result applies for all values of λ_1 and λ_2.

Example 6.11 (Standby pump)

Consider the standby system in Figure 6.10 with two identical pumps, each with constant failure rate $\lambda = 10^{-3}$ failures/hour. The probability p that the switch S fails to activate (switch over and start) the standby pump has been estimated to 1.5% (i.e. $p = 0.015$).

The survivor function of the pump system at time $t = 1000$ hours is from (6.38)

$$R_S(1000) = 0.7302.$$

The mean time to system failure is from (6.42)

$$\text{MTTF}_S = \frac{1}{\lambda}[1 + (1-p)] = 1985 \text{ hours}.$$

□

6.4.3 Partly Loaded Redundancy, Imperfect Switch, No Repairs

Consider the same standby system as the one in Figure 6.10, but change the assumptions such that item 2 carries a certain load before it is activated. Let λ_0

denote the failure rate of item 2 while in partly loaded standby. The system is able to survive the interval $(0, t]$ in two disjoint ways.

(1) Item 1 does *not* fail in $(0, t]$ (i.e. $T_1 > t$)
(2) Item 1 fails in a time interval $(\tau, \tau + d\tau)$, where $0 < \tau < t$. The switch S is able to activate item 2. Item 2 does not fail in $(0, \tau]$, is activated at time τ and does not fail in $(\tau, t]$.

Let T_S denote the time-to-system failure. The survivor function of the system, $R_S(t) = \Pr(T_S > t)$, is the sum of the probabilities for the two disjoint events.

Consider event (2): Item 1 fails in $(\tau, \tau + d\tau]$ with probability $f_1(\tau)\, d\tau = \lambda_1 e^{-\lambda_1 \tau}\, d\tau$. The switch S is able to activate item 2 with probability $1 - p$. Item 2 does not fail in $(0, \tau]$ in partly loaded standby with probability $e^{-\lambda_0 \tau}$, and item 2 does not fail in $(\tau, t]$ in active state with probability $e^{-\lambda_2(t-\tau)}$.

Because item 1 may fail at any point of time t in $(0, \tau]$, the survivor function of the system becomes

$$\begin{aligned} R_S(t) &= e^{-\lambda_1 t} + \int_0^t (1-p)e^{-\lambda_0 \tau} e^{-\lambda_2(t-\tau)} \lambda_1 e^{-\lambda_1 \tau}\, d\tau \\ &= e^{-\lambda_1 t} + \frac{(1-p)\lambda_1}{\lambda_0 + \lambda_1 - \lambda_2}(e^{-\lambda_2 t} - e^{-(\lambda_0+\lambda_1)t}), \end{aligned} \tag{6.43}$$

where we have assumed that $(\lambda_1 + \lambda_0 - \lambda_2) \neq 0$.

When $(\lambda_1 + \lambda_0 - \lambda_2) = 0$, the survivor function becomes

$$R_S(t) = e^{-\lambda_1 t} + (1-p)\lambda_1 t e^{-\lambda_2 t}. \tag{6.44}$$

The mean time to system failure is

$$\begin{aligned} \text{MTTF}_S &= \frac{1}{\lambda_1} + \frac{(1-p)\lambda_1}{\lambda_1 + \lambda_0 - \lambda_2}\left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1 + \lambda_0}\right) \\ &= \frac{1}{\lambda_1} + (1-p)\frac{\lambda_1}{\lambda_2(\lambda_1 + \lambda_0)}. \end{aligned} \tag{6.45}$$

This result applies for all values of λ_0, λ_1, and λ_2. In this section, we have tacitly made certain assumptions about independence. These assumptions are not discussed thoroughly here.

The concept of redundancy is discussed further in Chapter 11, where Markov models are used to study repairable as well as nonrepairable standby systems.

6.5 Single Repairable Items

This section introduces simple aspects of assessing the reliability of a single *repairable item*, that is, an item that is repaired when failure occurs. Other types of maintenance are not carried out. More advanced repair and maintenance strategies are treated in Chapter 9.

6.5.1 Availability

The main reliability metric for a repairable item is the availability of the item, introduced in Chapter 1 as

Definition 6.1 (Availability)
The availability $A(t)$ at time t of a repairable item is the probability that the item is functioning at time t:

$$A(t) = \Pr(\text{The item is in a functioning state at } t) = \Pr(X(t) = 1). \tag{6.46}$$

□

The availability $A(t)$ is also called the point – or time-dependent availability. Observe that when the item is not repaired, then $A(t) = R(t)$, the survivor function.

Definition 6.2 (Unavailability)
The unavailability $\overline{A}(t)[= 1 - A(t)]$ at time t of a repairable item is the probability that the item is not in a functioning state, but in a failed state, at time t:

$$\overline{A}(t) = \Pr(\text{The item is in a failed state at } t) = \Pr(X(t) = 0). \tag{6.47}$$

□

Sometimes, we are interested in the interval or mission availability in the time interval (t_1, t_2), defined by

Definition 6.3 (Interval availability)
The (average) interval or mission availability $A_{\text{avg}}(t_1, t_2)$ in the time interval (t_1, t_2) is defined as

$$A_{\text{avg}}(t_1, t_2) = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} A(t)\, dt. \tag{6.48}$$

□

$A_{\text{avg}}(t_1, t_2)$ is just the average value of the point availability $A(t)$ over a specified interval (t_1, t_2).

In some applications, we are interested in the interval or mission availability from startup, that is in an interval $(0, \tau)$. This is defined as

$$A_{\text{avg}}(0, \tau) = \frac{1}{\tau} \int_0^{\tau} A(t)\, dt. \tag{6.49}$$

The average availability $[A_{\text{avg}}(t_1, t_2)$ or $A_{\text{avg}}(0, \tau)]$ may be interpreted as the mean proportion of time in the interval where the item is able to function.

When $\tau \to \infty$, the average interval availability (6.49) approaches a limit called the long-run average availability of the item.

Definition 6.4 (Average availability)
The (long-run) average availability of an item is

$$A_{\text{avg}} = \lim_{\tau \to \infty} A_{\text{avg}}(\tau) = \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^{\tau} A(t)\, dt. \tag{6.50}$$

□

The (long-run) average availability may be interpreted as the average proportion of a long period of time where the item is able to function. The availability depends both on the number of failures that may occur and how quickly faults can be rectified (i.e. the maintainability and the maintenance support).

The (long-run) average unavailability $\overline{A}_{\text{avg}} = 1 - A_{\text{avg}}$ is in some application areas (e.g. electro-power generation) called the *forced outage rate.*

Example 6.12 (Average availability)
Consider an item that is assumed to run on a continuous basis. The item has an average availability of 0.95. During a period of one year (i.e. 8760 hours), we then expect the item to be functioning during $8760 \cdot 0.95 = 8322$ hours and not to be functioning during $8760 \cdot 0.05 = 438$ hours. Observe that the average availability does not tell anything about how many times the item will fail during this time interval. □

In many cases, the point availability $A(t)$ approaches a limit A when $t \to \infty$. The limit A is called the limiting availability of the item.

Definition 6.5 (Limiting availability)
The limiting availability is

$$A = \lim_{t \to \infty} A(t) \tag{6.51}$$

when the limit exists. □

The limiting availability is sometimes called the *steady-state availability*. When the limiting availability exists, it is equal to the long-run average availability, that is $A_{\text{avg}} = A$.

6.5.2 Average Availability with Perfect Repair

Consider an item that is put into operation and is functioning at time $t = 0$. Whenever the item fails, it is replaced with a new item of the same type or repaired to an as-good-as-new condition. We then get a sequence of times-to-failure or *uptimes* $T_1, T_2, \ldots$ for the item. We assume that $T_1, T_2, \ldots$ are independent and identically distributed, with distribution function $F_T(t) = \Pr(T_i \leq t)$ for $i = 1, 2, \ldots$ and mean uptime MUT.

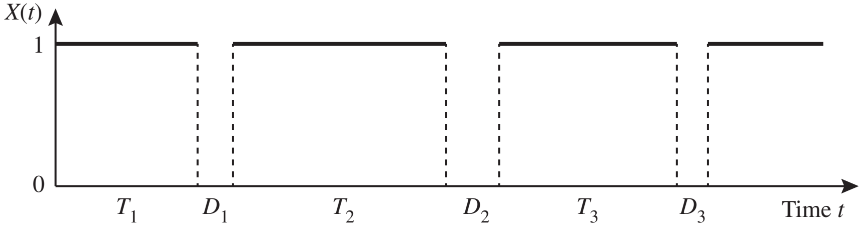


Figure 6.11 States of a repairable item.

We further assume that the downtimes $D_1, D_2, \dots$ are independent and identically distributed with distribution function $F_D(t) = \Pr(D_i \le t)$, for $i = 1, 2, \dots$ and mean downtime (MDT). Finally, we assume that all T_is and D_is are independent. This means, for example that the repair time is not influenced by the length of the uptime. The state variable $X(t)$ of the item is illustrated in Figure 6.11.

Suppose that we have observed an item until repair n is completed. Then we have observed the uptimes $T_1, T_2, \dots, T_n$ and the downtimes $D_1, D_2, \dots, D_n$. According to the *law of large numbers*, then under relatively general assumptions, with probability one

$$\frac{1}{n}\sum_{i=1}^{n} T_i \to E(T) = \text{MUT} \quad \text{when } n \to \infty$$

$$\frac{1}{n}\sum_{i=1}^{n} D_i \to E(D) = \text{MDT} \quad \text{when } n \to \infty.$$

The proportion of time in which the item has been functioning is

$$\frac{\sum_{i=1}^{n} T_i}{\sum_{i=1}^{n} T_i + \sum_{i=1}^{n} D_i} = \frac{(1/n)\sum_{i=1}^{n} T_i}{(1/n)\sum_{i=1}^{n} T_i + (1/n)\sum_{i=1}^{n} D_i}. \tag{6.52}$$

By the law of large numbers, the right-hand side of (6.52) tends to

$$\frac{E(T)}{E(T) + E(D)} = \frac{\text{MUT}}{\text{MUT} + \text{MDT}} \quad \text{as } n \to \infty,$$

which is the average proportion of time where the item has been functioning, when we consider a long period of time. We have therefore found the long-run average availability of the item.

$$A_{\text{avg}} = \frac{\text{MUT}}{\text{MUT} + \text{MDT}}. \tag{6.53}$$

The corresponding average unavailability is

$$\overline{A}_{\text{avg}} = \frac{\text{MDT}}{\text{MUT} + \text{MDT}}, \tag{6.54}$$

which is the average proportion of time where the item is not functioning when we consider a long period of time.

Example 6.13 (Average availability with perfect repair)
A machine with MTTF = 1000 hours has MDT = 5 hours. The assumption of perfect repair implies that MTTF = MUT (see Remark 6.2). This means that the machine has average availability

$$A_{\text{avg}} = \frac{\text{MUT}}{\text{MUT} + \text{MDT}} = \frac{1000}{1000 + 5} = 0.995.$$

On the average, this machine will function 99.5% of the time. The average unavailability is thus 0.5%, which corresponds to approximately 44 hours of downtime per year, when the machine is supposed to run continuously. □

Remark 6.2 (MTTF versus MUT)
The MTTF of an item is defined as the MTTF for an item that is in a fully operating state at time $t = 0$. For repairable items, the item is not always repaired to an as-good-as-new condition when it fails, even if we have assumed that each component fulfills this requirement. Consider, for example an item that is a parallel structure of three components. The item fails if and only if all the three components fail. This means that item failure involves three component failures. For some items, it is important to get the item into an operating state again as soon as possible, the item may therefore be started up again with one or two components functioning and one in failed state. It is therefore clear that the mean uptime, MUT may be different from the MTTF. □

6.5.3 Availability of a Single Item with Constant Failure and Repair Rates

Consider a repairable item where the uptimes are independent and exponentially distributed with failure rate λ. The downtimes are assumed to be independent and exponentially distributed with parameter μ. All repairs are assumed to be perfect. The mean downtime is

$$\text{MDT} = \frac{1}{\mu}.$$

The parameter μ is called the *repair rate*. Chapter 11 shows that the *availability* of the item at time t, with perfect repair, is

$$A(t) = \frac{\mu}{\lambda + \mu} + \frac{\lambda}{\lambda + \mu}\, e^{-(\lambda+\mu)t}. \tag{6.55}$$

The availability $A(t)$ is shown in Figure 6.12. For these uptime and downtime distributions, the availability $A(t)$ is seen to approach a constant A when $t \to \infty$.

$$A = \lim_{t\to\infty} A(t) = \frac{\mu}{\lambda + \mu} = \frac{1/\lambda}{1/\lambda + 1/\mu} = \frac{\text{MUT}}{\text{MUT} + \text{MDT}} \tag{6.56}$$

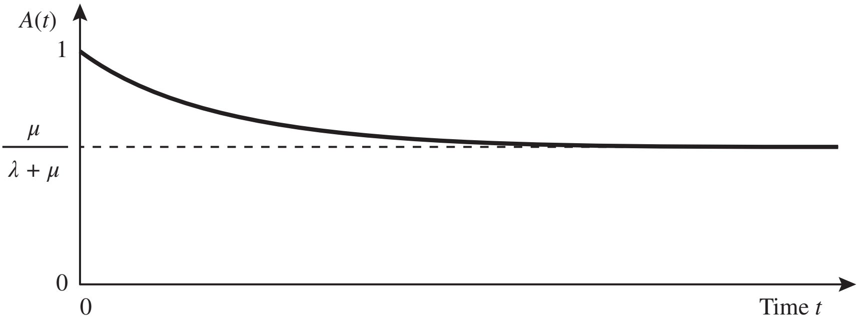


Figure 6.12 The availability $A(t)$ of an item with failure rate λ and repair rate μ.

A is the limiting availability and is, in this case, equal to the average availability of the item. When the item is not repaired, that is when $\mu = 0$, the availability $A(t)$ is seen to be equal to the survivor function $R(t)$.

In most cases, MDT $\ll$ MUT, and the average unavailability of the item may therefore be approximated as

$$\overline{A}_{\text{av}} = \frac{\text{MDT}}{\text{MUT} + \text{MDT}} = \frac{\lambda\ \text{MDT}}{1 + \lambda\ \text{MDT}} \approx \lambda\ \text{MDT}. \tag{6.57}$$

This approximation is often used in hand calculations.

When planning supplies of spare parts, it is of interest to know how many failures that may be expected in a given time interval. Let $W(t)$ denote the mean number of repairs carried out in the time interval $(0, t)$. Obviously, $W(t)$ depends on the distributions of the uptimes and the downtimes. It is often difficult to find an exact expression for $W(t)$ (see Chapter 10). When t is relatively large, the following approximation may be used:

$$W(t) \approx \frac{t}{\text{MTTF} + \text{MDT}}. \tag{6.58}$$

6.5.4 Operational Availability

The *operational availability* A_{OP} of an item is defined as the mean proportion of a mission period the item is able to perform its intended functions. To determine A_{OP}, we have to specify the mission period and estimate the mean total planned downtime and the mean total unplanned downtime in the mission period. The *operational unavailability* $\overline{A}_{\text{OP}} = 1 - A_{\text{OP}}$ may be determined from

$$\overline{A}_{\text{OP}} = \frac{\text{Mean total planned downtime} + \text{Mean total unplanned downtime}}{\text{Mission period}}.$$

When using the concepts availability and operational availability, we only consider two states: a functioning state and a failed state. The output from a production system may vary a lot, and the availability is therefore not a fully adequate measure of the system's performance. Several alternative metrics have been proposed. In ISO 20815 (2018), production assurance is measured as the operational performance of an oil/gas production system. Production availability is a term used to describe the capability of a production system of meeting demands for deliveries or performance. ISO 20815 was developed for the oil and gas industry, but most of the concepts may be used also in other industries. Production availability assessment is further discussed by Kawauchi and Rausand (2002).

6.5.5 Production Availability

Several metrics have been proposed for operational performance. Among these are the following: *Deliverability* is defined by ISO 20815 as the ratio between the actual deliveries and the planned/agreed deliveries over a specified period of time, when the effect of compensating elements such as substitution from other producers and downstream buffer storage are included.

$$\text{Deliverability} = \frac{\text{Actual deliveries}}{\text{Planned or agreed deliveries}}.$$

The deliverability is a metric for the system's ability to meet demands agreed with a customer. Failures and other problems in the production system may be compensated using products from a storage, or by purchasing products from other suppliers. The North Sea operators supply gas to Europe through subsea pipelines. The deliverability is measured at the interface between the subsea pipeline and the national gas network (e.g. in Germany). A relatively short downtime of a production unit does not have any effect on the outlet of the pipeline due to the large volume of gas and the high pressure in the pipeline. A longer downtime may be compensated by increasing the gas production from other production units, connected to the same pipeline.

The *on-stream availability*, OA, is defined as the mean proportion of time, in a specified time period, in which the production (delivery) is greater than zero. In this case, 1 − OA denotes the mean proportion of time the system is not producing at all.

The *100% production availability*, A_{100}, in a time interval (t_1, t_2) is defined as the mean proportion of the time in this interval the system is producing with full production (time is measured in hours).

$$A_{100} = \frac{\text{No. of hours in the interval } (t_1, t_2) \text{ with full production}}{t_2 - t_1}.$$

With A_{100}, we are only concerned with full production. We do not distinguish between 90% production and no production.

We may also define the production availability at a reduced capacity, for example, 80%

$$A_{80} = \frac{\text{No. of hours in } (t_1, t_2)\text{the system is producing with} \geq 80\% \text{ capacity}}{t_2 - t_1}.$$

6.5.6 Punctuality

A service is said to be *punctual* if it is initiated and/or completed “on time.” In the transport sector, punctuality is a commonly used reliability metric. The definition of being punctual varies between transport means and between countries. In civil aviation, punctuality may refer to the time leaving and arriving at the terminal, but may also refer to the time the wheels get off or on the runway. The criterion for on-time departure or arrival is specified as the number of minutes after the scheduled departure or arrival. In some countries, the accepted deviation for civil aviation is 15 minutes, but may be both shorter and longer. The punctuality in civil aviation is defined as

$$\text{Punctuality} = \frac{\text{No. of flights on time}}{\text{Total no. of scheduled flights}}.$$

The punctuality is usually presented as a percentage. The same definition of punctuality is used for railways, ferries, buses, and so on.

6.5.7 Failure Rate of Repairable Items

Assume that the item considered is functioning when it is put into operation at time $t = 0$, such that $X(0) = 1$.

Failure Rate Function

The failure rate function of a nonrepairable item was defined in Chapter 5 as

$$z(t) = \lim_{\Delta t \to 0} \frac{\Pr(t < T \leq t + \Delta t \mid T > t)}{\Delta t} = \frac{f(t)}{R(t)},$$

where $f(t)$ is the probability density function for the time-to-failure T and $R(t) = \Pr(T > t)$ is the survivor function. When Δt is small, we may use the approximation

$$\Pr(t < T \leq t + \Delta t \mid T > t) \approx z(t)\Delta t.$$

Because the item considered is *nonrepairable*, the events $T > t$ and $X(t) = 1$ give exactly the same information. The same applies for the two events $(t < T \leq t + \Delta t)$ and (Failure in $(t, t + \Delta t]$). When the item is known to be nonrepairable, the definition of the failure rate function may, alternatively, be expressed as

$$z(t) = \lim_{t \to \infty} \frac{\Pr(\text{Failure in } (t, t + \Delta t] \mid X(t) = 1)}{\Delta t}. \tag{6.59}$$

ROCOF

Another “failure rate” is the *rate of occurrence of failures*, ROCOF, that was briefly mentioned in Chapter 5. To define ROCOF, we start with the variable $N(t)$ = number of failures that occur in the time interval $(0, t]$ and its mean value $W(t) = E[N(t)]$.

The ROCOF at time t is defined as

$$w(t) = \lim_{\Delta t \to 0} \frac{E[N(t+\Delta t) - N(t)]}{\Delta t} = \lim_{\Delta t \to 0} \frac{W(t+\Delta t) - W(t)}{\Delta t} = \frac{d}{dt} W(t).$$

Because the time interval $(t, t+\Delta t]$ is very short, at most one failure can occur in this interval. The mean number of failures in this interval is therefore close to 1 times the probability of a failure in $(t, t+\Delta t]$. ROCOF can therefore be written as

$$w(t) = \lim_{\Delta t \to 0} \frac{\Pr(\text{Failure in } (t, t+\Delta t])}{\Delta t}. \tag{6.60}$$

When Δt is small, we may use the approximation

$$\Pr(\text{Failure in } (t, t+\Delta t]) \approx w(t)\Delta t. \tag{6.61}$$

The mean number of failures in a specific time interval $(t_1, t_2]$ is

$$W(t_1, t_2) = \int_{t_1}^{t_2} w(t)\, dt = W(t_2) - W(t_1). \tag{6.62}$$

ROCOF is more thoroughly discussed in Chapter 10.

Approximation Formula for ROCOF

Consider a repairable item that is always repaired to an as-good-as-new state (i.e. perfect repair). This will create a sequence of uptimes (U) and downtimes (D). Assume that the item is observed until failure n has been repaired. We then have the two sequences of observed uptimes and downtimes $u_1, u_2, \ldots, u_n$ and $d_1, d_2, \ldots, d_n$. During the observation period $\sum_{i=1}^{n}(u_i + d_i)$, n failures have occurred. The ROCOF of this process can therefore be determined as

$$w = \frac{n}{\sum_{i=1}^{n}(u_i + d_i)} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} u_i + \frac{1}{n}\sum_{i=1}^{n} d_i} \xrightarrow[n\to\infty]{} \frac{1}{\text{MUT} + \text{MDT}}.$$

For a future time t, we may therefore approximate the ROCOF by

$$w(t) \approx \frac{1}{\text{MUT} + \text{MDT}}. \tag{6.63}$$

This is an intuitive result because, on the average, we expect a failure every MUT + MDT time units.

Example 6.14 (Constant failure and repair rates)

Consider a single repairable item with constant failure rate λ and constant repair rate μ. The item is perfectly repaired upon failure. The mean time between failures, MTBF, of the item is MUT + MDT $= 1/\lambda + 1/\mu = (\lambda + \mu)/\lambda\mu$. After some time, the ROCOF becomes

$$w = \frac{1}{\text{MUT} + \text{MDT}} = \frac{\lambda\mu}{\lambda + \mu}. \tag{6.64}$$

If the item is put into operation and is functioning at time $t = 0$, the ROCOF is slightly different just after $t = 0$ but will soon become close to the value in (6.64). It is usually claimed to be sufficiently close after three MDTs, that is for $t \geq 3$ MDT. □

Vesely's Failure Rate

A third "failure rate" is the rate $z^V(t)$ defined as

$$z^V(t) = \lim_{\Delta t \to 0} \frac{\Pr(\text{Failure in } (t, t + \Delta t] \mid X(t) = 1)}{\Delta t}. \tag{6.65}$$

This failure rate for repairable (and nonrepairable) items was proposed by Vesely (1970) and is commonly known as the *Vesely failure rate*. When the item is nonrepairable, $z^{\mathrm{V}}(t)$ is identical with the failure rate function $z(t)$ in (6.59). For a repairable item, $X(t) = 1$ means that the item is functioning at time t, but provides no information about how long time it has been functioning since the previous repair (or startup).

When Δt is small, we may use the approximation

$$\Pr(\text{Failure in } (t, t + \Delta t] \mid X(t) = 1) \approx z^{\mathrm{V}}(t)\Delta t. \tag{6.66}$$

Let $E_t^{\Delta t}$ be the event "Failure in $(t, t + \Delta t]$." When this failure occurs, the state of the item at time t can be either $X(t) = 1$ or $X(t) = 0$. This means that

$$\Pr(E_t^{\Delta t}) = \Pr(E_t^{\Delta t} \cap X(t) = 1) + \Pr(E_t^{\Delta t} \cap X(t) = 0).$$

For the event $E_t^{\Delta t} \cap X(t) = 0$, the item is failed at time t and must be repaired before it can fail. Because Δt is small, two different events cannot take place in the interval $(t, t + \Delta t]$ and consequently $\Pr(E_t^{\Delta t} \cap X(t) = 0) = 0$. This yields

$$\Pr(E_t^{\Delta t}) = \Pr(E_t^{\Delta t} \cap X(t) = 1) = \Pr(E_t^{\Delta t} \mid X(t) = 1) \Pr(X(t) = 1)$$

and hence,

$$\Pr(E_t^{\Delta t} \mid X(t) = 1) = \frac{\Pr(E_t^{\Delta t})}{\Pr(X(t) = 1)}.$$

Because $\Pr(X(t) = 1) = A(t)$ is the availability of the item at time t, dividing by Δt on both sides and taking the limits, we obtain

$$z^V(t) = \frac{w(t)}{A(t)}. \tag{6.67}$$

For a nonrepairable item, $w(t) = f(t)$, the probability density function, and $A(t) = R(t)$, the survivor function, which shows that $z^V(t) = z(t)$ for nonrepairable components.

Example 6.15 (Constant failure and repair rates – cont.)
Reconsider the repairable item with constant failure rate λ and constant repair rate μ in Example 6.14. The Vesely failure rate for this item is from (6.67) equal to

$$z^V(t) = \frac{w(t)}{A(t)} = \frac{\lambda\mu/(\lambda+\mu)}{\mu/(\lambda+\mu)} = \lambda,$$

which is an obvious result. An item with constant failure rate λ is as-good-as-new as long as it is functioning. A nonrepairable item that is functioning at time t, such that $X(t) = 1$, hence has exactly the same properties as an item that has survived up to time t, such that $T > t$. □

6.6 Availability of Repairable Systems

A repairable system is a system of n components where at least one of the n components is repaired upon failure. Consider a repairable system with structure function $\phi[\boldsymbol{X}(t)]$. Because we have assumed that the state variables $X_1(t), X_2(t), \dots, X_n(t)$ are independent random variables, the system availability, $A_S(t)$ can be determined by the procedure described in Section 6.2:

$$A_S(t) = E(\phi[\boldsymbol{X}(t)]) = h[\boldsymbol{A}(t)], \tag{6.68}$$

where $\boldsymbol{A}(t)$ is the vector of the component availabilities $A_1(t), A_2(t), \dots, A_n(t)$. When we are only interested in the average availability, we skip the reference to time t and write $A_S = h(\boldsymbol{A})$. The system can be regarded as an item with mean uptime MUT_S and mean downtime MDT_S. The system average availability can then be written

$$A_S = h(\boldsymbol{A}) = \frac{\text{Mean uptime}}{\text{Total time}} = \frac{\text{MUT}_S}{\text{MUT}_S + \text{MDT}\ {}_{\cdot S}}. \tag{6.69}$$

This approach is illustrated in Example 6.16.

Example 6.16 (System availability calculation)
Consider the repairable system of three independent components in Figure 6.13. We are interested in the average availability only, and therefore skip the reference to time t. The structure function of the system is

$$\phi(\boldsymbol{X}) = X_1[X_2 + X_3 - X_2X_3]. \tag{6.70}$$

The MUTs and MDTs of the three components are listed below, together with the average component availabilities, calculated by

$$A_{\ avg,i} = \frac{\text{MUT}_i}{\text{MUT}_i + \text{MDT}_i} \quad \text{for } i = 1, 2, 3. \tag{6.71}$$

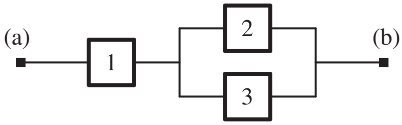


Figure 6.13 RBD for Example 6.16.

To simplify the notation, we omit the reference to average (avg), and write A_i instead of $A_{\text{avg},i}$.

i	MUT_i (hours)	MDT_i (hours)	A_i
1	1000	20	0.980
2	500	5	0.990
3	500	12	0.977

The average availability of the system is

$$A_S = A_1(A_2 + A_3 - A_2A_3) \approx 0.980.$$

The average system unavailability is thus $\overline{A}_S \approx 0.002$, which corresponds to approximately 174 hours of downtime per year when the system is supposed to operate on a continuous basis, that is, 8760 hours per year. □

The approach in Example 6.16 is based on the assumption that the system components fail and are repaired independently. This means that when a component is down for repair, all the other components continue to operate as if nothing had happened. This assumption is often not realistic, but despite of this, the approach is often used in practical analyses because it is easy to use. Most of the computer programs for FTA apply this simple approach to repairable systems.

6.6.1 The MUT and MDT of Repairable Systems

Consider a coherent and repairable system of n independent components with constant failure and repair rates (λ_i, μ_i) for $i = 1, 2, \dots, n$.

Component i is from (6.12) critical with probability

$$\Pr(\text{Component } i \text{ is critical}) = h(1_i, \boldsymbol{A}) - h(0_i, \boldsymbol{A}),$$

where $\boldsymbol{A}$ is the vector of component availabilities.

Component i is said to *cause system failure* when component i is critical and then fails. The frequency of system failures caused by component i, written as $w_S^{(i)}$ is equal to the frequency of failures of component i multiplied by the probability that component i is critical. By using (6.64), the frequency of system failures caused by component i is

$$w_S^{(i)} = \frac{\lambda_i \mu_i}{\lambda_i + \mu_i}[h(1_i, \boldsymbol{A}) - h(0_i, \boldsymbol{A})] \quad \text{for } i = 1, 2, \ldots, n. \tag{6.72}$$

The total frequency of system failures, the system's ROCOF w_S, is obtained by adding the contributions from the n components

$$w_S = \sum_{i=1}^{n} \frac{\lambda_i \mu_i}{\lambda_i + \mu_i}[h(1_i, \boldsymbol{A}) - h(0_i, \boldsymbol{A})]. \tag{6.73}$$

The system ROCOF can from (6.64) be written as

$$w_S = \frac{1}{\text{MUT}_S + \text{MDT}_S}. \tag{6.74}$$

Combining (6.64) and (6.74) yields

$$\text{MUT}_S = \frac{A_S}{w_S} \tag{6.75}$$

$$\text{MDT}_S = \frac{[1 - A_S]\ \text{MUT}_S}{A_S}, \tag{6.76}$$

where w_S is given in (6.74).

Remark 6.3 (Birnbaum's metric of importance)
In Chapter 7, Birnbaum's metric of component importance $I^{\text{B}}(i)$ is defined as

$$I^{\text{B}}(i) = h(1_i, \boldsymbol{A}) - h(0_i, \boldsymbol{A}).$$

This means that the frequency of system failures caused by component i can be written

$$w_S^{(i)} = \frac{\lambda_i \mu_i}{\lambda_i + \mu_i} I^{\text{B}}(i), \tag{6.77}$$

and that the frequency of system failures (i.e. the system ROCOF), w_S, can be written

$$w_S = \sum_{i=1}^{n} \frac{\lambda_i \mu_i}{\lambda_i + \mu_i} I^{\text{B}}(i). \tag{6.78}$$

□

Example 6.17 (Repairable series systems)
Consider a repairable series system with n independent components, and independent and perfect repairs. Component i has constant failure rate λ_i and constant repair rate μ_i, for $i = 1, 2, \ldots, n$. The average availability of component i is

$$A_i = \frac{\text{MUT}_i}{\text{MUT}_i + \text{MDT}_i} = \frac{\mu_i}{\lambda_i + \mu_i} \quad \text{for } i = 1, 2, \ldots, n.$$

The average *availability* of the series structure is

$$A_S = \prod_{i=1}^{n} \frac{\mu_i}{\lambda_i + \mu_i}.$$

The frequency of system failures caused by component i is from (6.77)

$$\begin{aligned} w_S^{(i)} &= \frac{1}{\text{MUT}_i + \text{MDT}_i}[h(1_i, \boldsymbol{A}) - h(0_i, \boldsymbol{A})] = \frac{1}{\text{MUT}_i + \text{MDT}_i} \prod_{j \neq i} A_j \\ &= \frac{\lambda_i \mu_i}{\lambda_i + \mu_i} \prod_{j \neq i} \frac{\mu_j}{\lambda_j + \mu_j} = \lambda_i \prod_{j=1}^{n} \frac{\mu_j}{\lambda_j + \mu_j} = A_S \lambda_i, \end{aligned} \tag{6.79}$$

which is an obvious result for a series system. For component i to cause system failure, the system must function (with probability A_S) and then component i must fail (with rate λ_i).

The frequency of system failures is

$$w_S = A_S \sum_{i=1}^{n} \lambda_i. \tag{6.80}$$

The mean system uptime, MUT_S is from (6.80)

$$\text{MUT}_S = \frac{A_S}{w_S} = \frac{1}{\sum_{i=1}^{n} \lambda_i}, \tag{6.81}$$

which is an obvious result for a series system. The system is functioning only when all its components are functioning and will remain in this state until the first component failure occurs with rate $\sum_{i=1}^{n} \lambda_i$.

The mean system downtime, MDT_S is from (6.81)

$$\text{MDT}_S = \frac{(1 - A_S)\,\text{MUT}_S}{A_S} = \frac{1 - A_S}{A_S} \frac{1}{\sum_{i=1}^{n} \lambda_i}. \tag{6.82}$$

□

A Numerical Example

Consider a series structure of $n = 4$ independent components with the following mean uptimes and mean downtimes:

i	**MUT$_i$ (hours)**	**MDT$_i$ (hours)**
1	1000	20
2	500	5
3	600	12
4	1200	30

The system availability is

$$A_S = \prod_{i=1}^{4} A_i \approx 0.9284.$$

By using the above formulas, we obtain

i	A_i	$w_S^{(i)}$	Percent
1	0.980	0.000928	18.2%
2	0.990	0.001857	36.4%
3	0.977	0.001547	30.3%
4	0.976	0.000774	36.4%
S	0.928	0.005106	100%

The percent column gives the fraction (in percent) of the system failures that are caused by component i for $i = 1, 2, 3, 4$. The mean system uptime and downtime are

$$\text{MUT}_S = 181.8\,\text{hours}$$
$$\text{MDT}_S = 14.0\,\text{hours}$$

Example 6.18 (Repairable parallel systems)
Consider a repairable parallel system with n independent components and independent and perfect repairs. Component i has constant failure rate λ_i and constant repair rate μ_i, for $i = 1, 2, \dots, n$. The average unavailability of component i is

$$\overline{A}_i = \frac{\text{MDT}_i}{\text{MUT}_i + \text{MDT}_i} = \frac{\lambda_i}{\lambda_i + \mu_i} \qquad \text{for } i = 1, 2, \dots, n.$$

The average unavailability of the parallel structure is therefore

$$\overline{A}_S = \prod_{i=1}^{n} \overline{A}_i = \prod_{i=1}^{n} \frac{\lambda_i}{\lambda_i + \mu_i} \tag{6.83}$$

and the system availability is

$$A_S = 1 - \prod_{i=1}^{n} \overline{A}_i.$$

The frequency of system failures caused by component i is

$$w_S^{(i)} = \frac{1}{\text{MUT}_i + \text{MDT}_i}[h(1_i, \boldsymbol{A}) - h(0_i, \boldsymbol{A})].$$

A parallel structure is functioning when at least one of its components is functioning and only fails when all components fail. Therefore,

$$h(1_i, \boldsymbol{A}) = 1$$
$$h(0_i, \boldsymbol{A}) = 1 - \prod_{j \neq i} \overline{A}_j$$

and

$$\begin{aligned} w_S^{(i)} &= \frac{1}{\text{MUT}_i + \text{MDT}_i} \prod_{j \neq i} \overline{A}_j = \frac{\lambda_i \mu_i}{\lambda_i + \mu_i} \prod_{j \neq i} \frac{\lambda_j}{\lambda_j + \mu_j} \\ &= \mu_i \prod_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \mu_j} = \overline{A}_S \mu_i. \end{aligned} \tag{6.84}$$

The frequency of system failures is

$$w_S = \overline{A}_S \sum_{i=1}^{n} \mu_i. \tag{6.85}$$

The mean system uptime, MUT_S is

$$\text{MUT}_S = \frac{1 - \overline{A}_S}{w_S} = \frac{1 - \overline{A}_S}{\overline{A}_S} \frac{1}{\sum_{i=1}^{n} \mu_i}. \tag{6.86}$$

The mean system downtime, MDT_S is

$$\text{MDT}_S = \frac{\overline{A}_S \, \text{MUT}_S}{1 - \overline{A}_S} = \frac{1}{\sum_{i=1}^{n} \mu_i}. \tag{6.87}$$

□

A Numerical Example

Consider a parallel structure of $n = 4$ independent components with the same mean uptimes and mean downtimes as in Example 6.18. The system unavailability is

$$\overline{A}_S = 9.284 \times 10^{-8}.$$

A repairable parallel system of four independent components is generally very reliable and will very seldom fail, even when the components have rather high failure rates.

By using the above formulas, we obtain

i	$\bar{A}_i$	$w_S^{(i)}$	Percent
1	0.019 61	4.64×10^{-9}	13.6%
2	0.009 90	1.86×10^{-8}	54.5%
3	0.019 61	7.74×10^{-9}	22.7%
4	0.024 39	3.09×10^{-9}	9.1%
S	9.284×10^{-8}	3.40×10^{-8}	100%

The mean system uptime and downtime are

$$\text{MUT}_S \approx 29\ 375\ 000\,\text{hours} \approx 3353\,\text{years}$$

$$\text{MDT}_S = 14.0\,\text{hours}$$

Remark 6.4 (Assumptions and limitations)
The approach outlined in Section 6.6.1 is based on several assumptions that may be questioned. The assumption that all component uptimes and downtimes are independent implies that failed components are repaired online, that is, when the other components are functioning as normal and that the repair actions do not influence the functioning components. Another consequence is that there is no shortage of repair resources. When a component fails, there is always a repair team available to carry out the repair.

The formulas in Section 6.6.1 are correct for independent components for a specific point in time, but they are not fully correct for the average availability, that is, over a long interval in time. Section 6.3.2 shows that the failure rate of a parallel structure is not constant even if all components have constant failure rates. A similar effect is also the case for repairable parallel structures (but is not shown here). The frequency w_S of system failures is hence not constant as assumed in the calculations above. □

6.6.2 Computation Based on Minimal Cut Sets

Consider a repairable system of n independent components. When all the minimal cut sets $C_1, C_2, \dots, C_k$ of the structure are determined, the structure can be represented as a series structure of the k minimal cut parallel structures (MCPSs). The system reliability properties can therefore be determined from the results on series and parallel structures in Examples 6.17 and 6.18. The approach is illustrated in Example 6.19

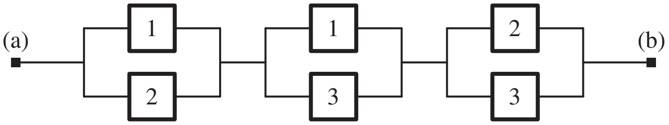


Figure 6.14 RBD for Example 6.20, drawn as a series structure of its three MCPSs.

Example 6.19 (Repairable 2oo3 structure)

The repairable 2oo3 structure in Figure 6.14 has the following three minimal cut sets:

$$C_1 = \{1,2\}, \quad C_2 = \{1,3\}, \quad C_3 = \{2,3\}.$$

Assume that the three components are identical with failure rate λ and repair rate μ such that the three MCPSs have the same probabilistic properties. Consider one specific MCPS. Using the results from Example 6.18, the average unavailability of a component is

$$\overline{A} = \frac{\text{MDT}}{\text{MUT} + \text{MDT}} = \frac{\lambda}{\lambda + \mu}.$$

The average unavailability of an MCPS is

$$\overline{A}_{\text{MCPS}} = \overline{A}^2 = \left(\frac{\lambda}{\lambda + \mu}\right)^2.$$

The frequency of failures of an MCPS caused by a specific component i is from (6.84)

$$w_{\text{MCPS}}^{(i)} = \overline{A}_{\text{MCPS}}\mu = \frac{\lambda^2 \mu}{(\lambda + \mu)^2}.$$

Because there are two components in each MCPS, the frequency of MCPS-failures is

$$w_{\text{MCPS}} = \frac{2\lambda^2 \mu}{(\lambda + \mu)^2}.$$

The mean MCPS uptime, MUT_{MCPS} is

$$\text{MUT}_{\text{MCPS}} = \frac{1 - \overline{A}_{\text{MCPS}}}{w_{\text{MCPS}}} = \frac{1}{\lambda} + \frac{\mu}{2\lambda^2}.$$

The mean MCPS downtime, MDT_{MCPS} is

$$\text{MDT}_{\text{MCPS}} = \frac{1}{2\mu}.$$

Consider the three MCPSs as components in a series structure and use the results from Example 6.17 to find the system availability

$$A_S = (1 - \overline{A}_{\text{MCPS}})^3.$$

The frequency of system failures caused by a specific MCPS j is from (6.84)

$$w_S^{(j)} = A_S \frac{2\lambda^2\mu}{(\lambda+\mu)^2}.$$

The frequency of system failures

$$w_S = 3A_S \frac{2\lambda^2\mu}{(\lambda+\mu)^2} = A_S \frac{6\lambda^2\mu}{(\lambda+\mu)^2}.$$

The mean system up-time, MUT_S, is

$$\text{MUT}_S = \frac{1}{3w_{\text{MCPS}}} = \frac{(\lambda+\mu)^2}{6\lambda^2\mu}.$$

The mean system downtime, MDT_S, is

$$\text{MDT}_S = \frac{1-A_S}{A_S}\frac{1}{3w_{\text{MCPS}}}.$$

□

A Numerical Example

A repairable 2oo3 system has independent and identical components with failure rate $\lambda = 7.2 \times 10^{-5}$ per hour and mean repair time MDT = 24 hours. The repair rate is then $\mu = 1/\text{MDT}$. Using the equations above, yields

- The system unavailability is $\overline{A}_S = 2.976 \times 10^{-6}$.
- The frequency of failures of a specific MCPS is $w_{\text{MCPS}} = 2.480 \times 10^{-7}$ per hour (this corresponds to one system failure per 153 years).
- The mean MCPS uptime is $\text{MUT}_{\text{MCPS}} = 4.033 \times 10^6$ hours.
- The mean MCPS downtime is $\text{MDT}_{\text{MCPS}} = 12$ hours.
- The frequency of system failures is $w_S = 7.439 \times 10^{-7}$ per hour.
- The mean system uptime is $\text{MUT}_S = 1.344 \times 10^6$ hours.
- The mean system downtime is $\text{MDT}_S = 12$ hours.

Remark 6.5 (Not fully correct result)

The system results obtained in Example 6.19 are not fully correct. The MCPSs are not independent because each component is a member of two MCPSs. The same problem arises for all systems having overlapping MCPSs, but in many cases, the results are approximately correct. □

6.6.3 Uptimes and Downtimes for Reparable Systems

In general, the uptimes and downtimes of a repairable system are not identically distributed, but depends on the repair strategy and the completeness of the individual repairs. This is illustrated for a parallel structure of three independent and identical components in Example 6.20.

Example 6.20 (Parallel structure of three components)
Consider a repairable parallel structure of three independent and identical components with constant failure rate λ. There are three independent repair teams. When the structure has failed (i.e. all three components have failed), the teams start repairing one component each. The repair time T_r for each component has constant repair rate μ. The repair strategy is as follows:

(1) The structure is put into operation again after a specified downtime of t_r hours if at least one component is repaired.
(2) If all three components are repaired before time t_r, the structure is put into operation again as soon as all repairs are finished.
(3) If none of the components are repaired at time t_r, the repair continues until the first repair is finished. Then the structure is put into operation with only one component functioning.

The probability that a repair is finished before time t_r is $p_r = 1 - e^{-\mu t_r}$. Let N denote the number of components that have been repaired before time t_r. N takes the values 0, 1, 2, 3 with probabilities:

$$\begin{aligned}
\Pr(N = 0) &= (1 - p_r)^3, \\
\Pr(N = 1) &= \binom{3}{1}(1 - p_r)^2 p_r, \\
\Pr(N = 2) &= \binom{3}{2}(1 - p_r) p_r^2, \\
\Pr(N = 3) &= p_r^3.
\end{aligned}$$

The mean downtime, MDT, for the four possible outcomes is

$$\begin{aligned}
\text{MDT}_0 &= t_r + \frac{1}{3\mu}, \\
\text{MDT}_1 &= t_r, \\
\text{MDT}_2 &= t_r, \\
\text{MDT}_3 &= t_r - \frac{1}{(1 - e^{-\mu t_r})^3} \\
&\quad \left(t_r - \frac{3}{\mu}(1 - e^{-\mu t_r}) + \frac{3}{2\mu}(1 - e^{-2\mu t_r}) - \frac{1}{3\mu}(1 - e^{3\mu t_r}) \right).
\end{aligned}$$

The mean uptimes following the three possible outcomes of the repair are from Table 6.2:

$$\begin{aligned}
\text{MUT}_0 &= \text{MTTF}_{1\text{oo}1} = \frac{1}{\lambda} \\
\text{MUT}_1 &= \text{MTTF}_{1\text{oo}1} = \frac{1}{\lambda} \\
\text{MUT}_2 &= \text{MTTF}_{1\text{oo}2} = \frac{3}{2\lambda}
\end{aligned}$$

$$\text{MUT}_3 = \text{MTTF}_{1\text{oo}3} = \frac{11}{6\lambda}$$

When $N = 1$ and $N = 2$, the downtime is t_r. $N = 0$ means that none of the three repair teams have finished by time t_r. Because of the memoryless property of the exponential distribution, we can as well start a new repair process at time t_r, and the mean time until the first component is repaired is $1/3\lambda$. For $N = 3$, we know that all the three repairs have finished no later than t_r. The probability that the repair time T_3 of all three components is longer than t is equal to the conditional "survival" probability

$$R(t \mid t_r) = \Pr(T_3 > t \mid T_3 \leq t_r).$$

Solving the integral $\int_0^{t_r} R(t \mid t_r)\, dt$ yields MDT_3. The average availability for a single cycle (one downtime and one uptime), A, is MUT/(MUT+MDT) □

A Numerical Example

Let $t_r = 8$ hours and $\mu = 0.10$ (hours)$^{-1}$, such that MTTR for each component is 10 hours. Further, let $\lambda = 0.001$ (hours)$^{-1}$, such that the MTTF of a single component is 1000 hours. With these input values, the following results are obtained:

$N = n$	$\Pr(N = n)$	MDT_n	MUT_n	A_n
0	0.0907	11.33	1000	0.9888
1	0.3335	8.00	1000	0.9921
2	0.4088	8.00	1500	0.9947
3	0.1670	5.48	1833	0.9970
Average	—	7.88	1344	0.9937

The total mean downtime in the cycle is calculated as $\text{MDT}_{\text{av}} = \sum_{n=0}^{3} \text{MDT}_n \Pr(N = n)$. Similar for the mean uptime and the availability.

The *mean time-to-first-failure* (i.e. to the first occurrence of the system failure), MTTFF, is always greater or equal to the mean uptimes. This is because all components are assumed to be in a functioning state at time $t = 0$, but this is not always the case after a component has been restored.

6.7 Quantitative Fault Tree Analysis

Fault tree construction and its qualitative aspects are introduced in Section 4.3, whereas the present section deals with quantitative analysis of fault trees. Quantitative FTA can be approximative or "exact." This section is delimited to

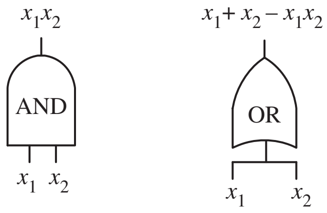


Figure 6.15 State variables for fault tree AND and OR gates.

approximative approaches, because these are feasible for hand calculation and because they, for most purposes, give sufficiently accurate results.

Because any static fault tree (i.e. with only AND and OR gates) can be converted to an RBD, the fault tree can be analyzed based on structure functions in the same way as for RBDs. The Boolean functions for AND and OR gates are illustrated in Figure 6.15 and are the same as for RBDs. The algebra that was developed for RBDs can therefore be used to obtain a structure function for a fault tree. This topic is not pursued any further in this book, because practical fault trees tend to be so large that this approach is not tractable.

A number of guidelines and handbooks on FTA are available on the Internet. Two of the most comprehensive references are NUREG-0492 (1981) and NASA (2002).

6.7.1 Terminology and Symbols

The main symbols used in quantitative FTA are

$q_i(t)$ The probability that basic event B_i occurs (i.e. is present) at time t, i.e. $q_i(t) = \Pr[B_i(t)]$. This probability may be interpreted as the unavailability of the corresponding component/item.

$Q_0(t)$ The probability that the TOP event occurs (i.e. is present) at time t. $Q_0(t)$ is called the TOP event probability and may be interpreted as the unavailability of the system.

$\check{Q}_j(t)$ The probability that *minimal cut parallel structure j* (MCPS$_j$) of the fault tree is failed E$_j$ at time t, i.e. $\check{Q}_j(t) = \Pr[E_j(t)]$. An MCPS fails if and only if all the basic events in the minimal cut set occurs (i.e. are present) at the same time.

6.7.2 Delimitations and Assumptions

This section is delimited to *static* FTA with only AND, OR, and voting gates. The following assumptions apply:

(1) All basic events are binary. They are either present or not present.
(2) All basic events are statistically independent.
(3) No basic events are present (i.e. no components are failed) at time $t = 0$.

(4) The fault tree logic is coherent with respect to both failure and repair.
(5) Transitions between the (binary) states of basic events occur instantaneously. It is not possible for a basic event to be in an intermediate state, even for a very short time.
(6) When items are repaired, they are always brought back to an as-good-as-new state.
(7) All repairs are performed on-line and do not influence the performance of other components.

Remark 6.6 (Basic events are states)
The statement "basic event *i* occurs at time *t*" may be misleading. The basic events and the TOP event in a fault tree are in reality *states*, and not events. When we say that a basic event (or the TOP event) occurs at time *t*, we mean that the corresponding *state* is present at time *t*. □

In the same way as for RBDs, it can be shown that when the basic events are (statistically) independent, $Q_0(t)$ is a function $g(\cdot)$ of the $q_i(t)$s only, for $i = 1, 2, \dots, n$, where n is the number of different basic events in the fault tree. Hence, $Q_0(t)$ may be written

$$Q_0(t) = g[q_1(t), q_2(t), \dots, q_n(t)] = g[\boldsymbol{q}(t)]. \tag{6.88}$$

6.7.3 Fault Trees with a Single AND-Gate

Consider the fault tree in Figure 6.16a with a single AND-gate. The TOP event occurs if and only if all the basic events $B_1, B_2, \dots, B_n$ occur simultaneously.

By the same type of Boolean reasoning as used for RBDs, the TOP event probability may be written

$$Q_0(t) = q_1(t)q_2(t) \cdots\ q_n(t) = \prod_{i=1}^{n} q_i(t). \tag{6.89}$$

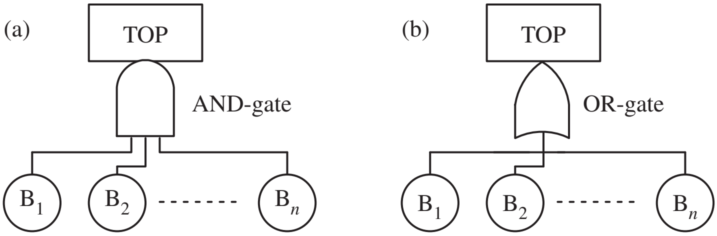


Figure 6.16 Fault trees with single AND-gate and single OR-gate.

$Q_0(t)$ may also be determined directly by set theory arguments (which are also Boolean): Let $B_i(t)$ denote that basic event B_i occurs at time t for $i = 1, 2, \dots, n$. Then

$$\begin{aligned} Q_0(t) &= \Pr[B_1(t) \cap B_2(t) \cap \cdots \cap B_n(t)] \\ &= \Pr[B_1(t)] \Pr[B_2(t)] \ \cdots \Pr[B_n(t)] \\ &= q_1(t) q_2(t) \cdots \ q_n(t) = \prod_{i=1}^{n} q_i(t) \end{aligned}$$

6.7.4 Fault Tree with a Single OR-Gate

Consider the fault tree in Figure 6.16b. The TOP event of this fault tree occurs if at least one of the independent basic events $B_1, B_2, \dots, B_n$ occurs.

The same type of Boolean reasoning as used for RBDs gives the TOP event probability

$$Q_0(t) = 1 - \prod_{i=1}^{n} [1 - q_i(t)]. \tag{6.90}$$

As for the AND gate, $Q_0(t)$ can be determined directly in the following way: Let $B_i^*(t)$ denote that basic event B_i does *not* occur at time t. Then

$$\Pr[B_i^*(t)] = 1 - \Pr(B_i(t)) = 1 - q_i(t), \qquad \text{for } i = 1, 2, \dots, n.$$

$$\begin{aligned} Q_0(t) &= \Pr[B_1(t) \cup B_2(t) \cup \cdots \cup B_n(t)] \\ &= 1 - \Pr[B_1^*(t) \cap B_2^*(t) \cap \cdots \cap B_n^*(t)] \\ &= 1 - \Pr[B_1^*(t)] \Pr[B_2^*(t)] \ \cdots \Pr[B_n^*(t)] \\ &= 1 - \prod_{i=1}^{n} [1 - q_i(t)] \end{aligned}$$

6.7.5 The Upper Bound Approximation Formula for $Q_0(t)$

To determine the TOP event probability by means of the structure function may in many cases be both time-consuming and cumbersome. Hence, there may be a need for approximation formulas.

Consider a system (fault tree) with k minimal cut sets $C_1, C_2, \dots, C_k$. This system may be represented as a series structure of the k MCPSs, as illustrated by the RBD in Figure 6.17.

The TOP event occurs if at least one of the k MCPSs fails. An MCPS fails if each and all the basic events in the minimal cut set occur simultaneously. Observe that the same input event may enter in many different cut sets.

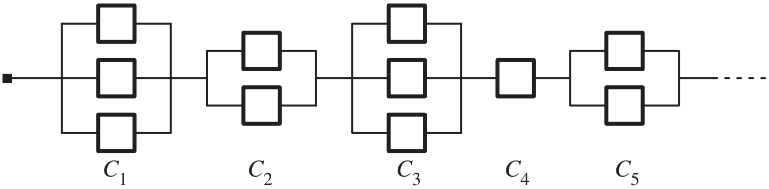


Figure 6.17 A structure represented as a series structure of the minimal cut parallel structures (MCPSs).

The probability that minimal cut parallel structure j fails at time t, when its basic events are independent, is

$$\check{Q}_j(t) = \prod_{i \in C_j} q_i(t). \tag{6.91}$$

If all the k minimal cut parallel structure were independent, the TOP event probability would be

$$Q_0(t) = \coprod_{j=1}^{k} \check{Q}_j\,(t) = 1 - \prod_{j=1}^{k}\,[1 - \check{Q}_j(t)]. \tag{6.92}$$

Because the same basic event may occur in several minimal cut sets, the MCPSs can obviously be positively dependent, but it may be shown (e.g. see Barlow and Proschan 1975) that

$$Q_0(t) \leq 1 - \prod_{j=1}^{k} [1 - \check{Q}_j(t)]. \tag{6.93}$$

Hence, the right-hand side of (6.93) may be used as an upper (conservative) bound for the TOP event probability.

When all the $q_i(t)$'s are very small, it may be shown that with good approximation

$$Q_0(t) \approx 1 - \prod_{j=1}^{k} [1 - \check{Q}_j(t)]. \tag{6.94}$$

This approximation is called the *upper bound approximation*, and it is used in many computer programs for FTA, but the approximation (6.94) must be used with care when at least one of the $q_i(t)$s is of order 10^{-2} or larger.

Assume that all the $\check{Q}_j(t)$s are so small that we can disregard their products. In this case, (6.94) may be approximated by

$$Q_0(t) \approx 1 - \prod_{j=1}^{k}\,[1 - \check{Q}_j(t)] \approx \sum_{j=1}^{k} \check{Q}_i(t). \tag{6.95}$$

It is straightforward to verify that the last approximation is more conservative than the first one:

$$Q_0(t) \le 1 - \prod_{j=1}^{k} [1 - \check{Q}_j(t)] \le \sum_{j=1}^{k} \check{Q}_i(t). \tag{6.96}$$

Remark 6.7 (Rare event approximation)
The approximation used in (6.95) is called the *rare event approximation*. In its most simple form, it says that if we have two events A and B, then

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) \approx \Pr(A) + \Pr(B).$$

The approximation reduces the result to its first-order terms and is generally an adequate approximation when $\Pr(A \cap B)$ is small, that is, when it is deemed unlikely that two events occur at the same time. □

6.7.6 The Inclusion–Exclusion Principle

The TOP event probability can also be determined by the *inclusion–exclusion principle*. The same approach can be used to determine the system *reliability*.

A fault tree with n different and independent basic events has k minimal cut sets $C_1, C_2, \dots, C_k$. Let E_j denote the event that the MCPS j fails at time t. To simplify the notation, we skip the reference to the time t in the formulas.

Because the TOP event occurs as soon as one of its MCPSs fails, the TOP event probability may be expressed by

$$Q_0 = \Pr\left(\bigcup_{j=1}^{k} E_j\right). \tag{6.97}$$

In general, the individual events $E_j, j = 1, 2, \dots, k$ are not disjoint. Hence, the probability $\Pr(\bigcup_{j=1}^{k} E_j)$ is determined by using the *general addition theorem* in probability theory.

$$\begin{aligned} Q_0 &= \sum_{j=1}^{k} \Pr(E_j) - \sum_{i<j} \Pr(E_i \cap E_j) + \cdots \\ &\quad + (-1)^{j+1} \Pr(E_1 \cap E_2 \cap \cdots \cap E_k) \end{aligned} \tag{6.98}$$

By introducing

$$W_1 = \sum_{j=1}^{k} \Pr(E_j)$$

$$W_2 = \sum_{i<j} \Pr(E_i \cap E_j)$$

$$\vdots$$

$$W_k = \Pr(E_1 \cap E_2 \cap \cdots \cap E_k)$$

Eq. (6.98) may be written

$$\begin{aligned} Q_0 &= W_1 - W_2 + W_3 - \cdots + (-1)^{k+1} W_k \\ &= \sum_{j=1}^{k} (-1)^{j+1} W_j \end{aligned} \tag{6.99}$$

Example 6.21 (Bridge structure)
Consider the bridge structure illustrated by the RBD in Figure 6.18. The minimal cut sets of the bridge structure are

$$C_1 = \{1, 2\}, \quad C_2 = \{4, 5\}, \quad C_3 = \{1, 3, 5\}, \quad C_4 = \{2, 3, 4\}.$$

Based on these minimal cut sets, a fault tree for the bridge structure may be established. As before, let B_i denote the basic event that component i is failed, for $i = 1, 2, 3, 4, 5$.

According to (6.99), the TOP event probability Q_0 of the fault tree for bridge structure is

$$Q_0 = W_1 - W_2 + W_3 - W_4,$$

where

$$\begin{aligned} W_1 &= \sum_{j=1}^{4} \Pr(E_j) \\ &= \Pr(B_1 \cap B_2) + \Pr(B_4 \cap B_5) + \Pr(B_1 \cap B_3 \cap B_5) + \Pr(B_2 \cap B_3 \cap B_4) \\ &= q_1 q_2 + q_4 q_5 + q_1 q_3 q_5 + q_2 q_3 q_4 \end{aligned}$$

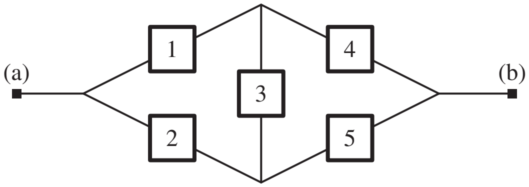


Figure 6.18 RBD for the bridge structure.

$$\begin{aligned}
W_2 = \sum_{i<j} \Pr(E_i \cap E_j) &= \Pr(E_1 \cap E_2) + \Pr(E_1 \cap E_3) + \Pr(E_1 \cap E_4) \\
&\quad + \Pr(E_2 \cap E_3) + \Pr(E_2 \cap E_4) + \Pr(E_3 \cap E_4) \\
&= \Pr(B_1 \cap B_2 \cap B_4 \cap B_5) \\
&\quad + \Pr(B_1 \cap B_2 \cap B_1 \cap B_3 \cap B_5) \\
&\quad + \Pr(B_1 \cap B_2 \cap B_2 \cap B_3 \cap B_4) \\
&\quad + \Pr(B_4 \cap B_5 \cap B_1 \cap B_3 \cap B_5) \\
&\quad + \Pr(B_4 \cap B_5 \cap B_2 \cap B_3 \cap B_4) \\
&\quad + \Pr(B_1 \cap B_3 \cap B_5 \cap B_2 \cap B_3 \cap B_4) \\
&= q_1q_2q_4q_5 + q_1q_2q_3q_5 + q_1q_2q_3q_4 + q_1q_3q_4q_5 \\
&\quad + q_2q_3q_4q_5 + q_1q_2q_3q_4q_5.
\end{aligned}$$

Similarly,

$$W_3 = 4q_1q_2q_3q_4q_5,$$

and

$$W_4 = q_1q_2q_3q_4q_5.$$

Hence, the TOP event probability – and the system unavailability – is

$$\begin{aligned}
Q_0 &= W_1 - W_2 + W_3 - W_4 \\
&= q_1q_2 + q_4q_5 + q_1q_3q_5 + q_2q_3q_4 - q_1q_2q_4q_5 - q_1q_2q_3q_5 \\
&\quad - q_1q_3q_4q_5 - q_2q_3q_4q_5 + 2q_1q_2q_3q_4q_5
\end{aligned}$$

□

Example 6.21 shows that, when using the general addition theorem (6.99) we have to calculate the probability of a large number of terms that later cancel each other. An alternative approach is proposed by Satyanarayana and Prabhakar (1978). The idea behind their method is, with the help of graph theoretical arguments, to leave out the cancelling terms at an early stage without having to calculate them.

Calculating the exact value of a system's unreliability Q_0 by means of (6.98) may be cumbersome and time-consuming, even when the system is relatively simple. In such cases, one may sometimes be content with an approximative value for the TOP event probability.

Approximation Formulas by the Inclusion–Exclusion Principle

One way of determining approximate values of the TOP event probability (or the system unavailability) Q_0 utilizes the following result based on inclusion–exclusion:

$$\begin{aligned} Q_0 &\leq W_1 \\ W_1 - W_2 &\leq Q_0 \\ Q_0 &\leq W_1 - W_2 + W_3 \\ &\vdots \end{aligned} \tag{6.100}$$

It can be shown that

$$(-1)^{j-1} Q_0 \leq (-1)^{j-1} \sum_{v=1}^{j} (-1)^{v-1} W_v \qquad \text{for } j = 1, 2, \dots, k. \tag{6.101}$$

Equation (6.100) may give the impression that the differences between the consecutive upper and lower bounds are monotonically decreasing, but this is not true in general.

In practice, (6.100) is used the following way: Successively, we determine upper and lower bounds for Q_0, proceeding downwards in (6.100) until we obtain bounds that are sufficiently close.

Example 6.22 (Bridge structure – cont.)
Reconsider the bridge structure in Example 6.21 and assume that all the basic event probabilities q_i are equal to 0.05. Introducing these q_is in the expression for the W_is in Example 6.20, yields

$$\begin{aligned} W_1 &= 5250 \times 10^{-6} \\ W_2 &= 3156 \times 10^{-6} \\ W_3 &= 1.25 \times 10^{-6} \\ W_4 &= 0.31 \times 10^{-6} \end{aligned}$$

From (6.100) we get:

$$\begin{aligned} Q_0 \leq & W_1 \approx 5250 \times 10^{-6} = 0.5250\% \\ Q_0 \geq & W_1 - W_2 \approx 5218.4 \times 10^{-6} = 0.5218\% \end{aligned}$$

From the first two inequalities of (6.100), we hence know that

$$0.5218\% \leq Q_0 \leq 0.5250\%.$$

For many applications, this precision may be sufficient. If not, we proceed and calculate the next inequality:

$$Q_0 \leq W_1 - W_2 + W_3 \approx 5219.69 \times 10^{-6} = 0.5220\%.$$

Now we know that Q_0 is bounded by

$$0.5218\% \leq Q_0 \leq 0.5220\%.$$

The exact value is

$$Q_0 = W_1 - W_2 + W_3 - W_4 = 5219.38 \times 10^{-6} \approx 0.5219\%.$$

By comparison, the upper bound obtained by (6.93) is equal to

$$1 - \prod_{j=1}^{k}(1-\check{Q}_j) = 0.005\ 242\ 49 \approx 0.5242\%.$$

□

6.7.7 ROCOF of a Minimal Cut Parallel Structure

Consider an MCPS of two independent and repairable components (basic events), 1 and 2. Let $\check{Q}^{(1)}(t)$ denote the event that the MCPS fails at time t because of a failure of component 1. For this event to happen, component 2 must be down at time t, because if component 2 were functioning, the parallel structure would not fail when component 1 failed. The ROCOF of MCPS failures caused by component 1 is therefore

$$w^{\ (1)}(t) = w_1(t)\ q_2(t),$$

where $w_1(t)$ is the ROCOF of component 1. The ROCOF of MCPS failures caused by component 2 is determined by the same arguments. The total ROCOF of the MCPS is therefore

$$\check{w}(t) = \check{w}^{\ (1)}(t) + \check{w}^{\ (2)}(t) = w_1(t)q_2(t) + w_2(t)q_1(t).$$

Consider a minimal cut set C_κ of any order ≥ 2. For component i to cause failure of the MCPS, all the other components of the minimal cut set C_κ must be down and the ROCOF for MCPS κ can therefore be calculated as

$$\check{w}_\kappa(t) = \sum_{i\in C_\kappa} w_i(t) \prod_{\ell\in C_\kappa,\ell\neq i} q_\ell(t). \tag{6.102}$$

6.7.8 Frequency of the TOP Event

This section presents simple formulas for the frequency of the TOP event based on the formulas developed in Section 6.6. The frequency of the TOP event is the expected number of occurrences of the TOP event per time unit (e.g. per year).

A coherent system can always be represented as a series structure of its MCPSs. For an MCPS κ to cause system failure (i.e. occurrence of the TOP event), none of the other MCPSs can be failed. The frequency of the TOP event caused by MCPS κ is therefore approximately

$$w_{\text{TOP}}^{(\kappa)} \approx \check{w}_\kappa(t) \prod_{j=1,j\neq\kappa}^{k} (1-\check{Q}_j(t)). \tag{6.103}$$

The approximation is due to the fact that the minimal cuts are generally not independent because the same basic event can be present in several minimal cut sets.

The overall frequency of the occurrence of the TOP event is approximately

$$w_{\mathrm{TOP}}(t) \approx \sum_{\kappa=1}^{k} w_{\mathrm{TOP}}^{(\kappa)} = \sum_{\kappa=1}^{k} \check{w}_{\kappa}(t) \prod_{j=1,j\neq\kappa}^{k} (1 - \check{Q}_j(t)). \tag{6.104}$$

Example 6.23 (Bridge structure)

Reconsider the bridge structure in Example 6.21 and assume we have established a corresponding fault tree based on the minimal cut sets $C_1 = \{1,2\}, C_2 = \{4,5\}, C_3 = \{1,3,5\}$, and $C_4 = \{2,3,4\}$. The five components are independent and repairable. Repair is carried out of individual components (i.e. on-line repair) and always returns the component to an as-good-as-new condition. Failure and repair rates are constant and downtimes are independent of uptimes. The following rates (per hour) are provided:

Component *i*	λ_i	μ_i
1	0.001	0.10
2	0.002	0.08
3	0.005	0.03
4	0.003	0.10
5	0.002	0.12

The basic event probability of component i is

$$q_i = \frac{\mathrm{MDT}_i}{\mathrm{MUT}_i + \mathrm{MDT}_i} = \frac{\lambda_i}{\lambda_i + \mu_i}.$$

The mean time between failures for component i is

$$\mathrm{MUT}_i + \mathrm{MDT}_i = \frac{1}{\lambda_i} + \frac{1}{\mu_i}.$$

The frequency of component i failures is

$$w_i = \frac{1}{\mathrm{MUT}_i + \mathrm{MDT}_i} = \frac{\lambda_i \mu_i}{\lambda_i + \mu_i}. \tag{6.105}$$

With the provided data, we obtain the following values (with time unit hours):

Component i	q_i	MTBF_i
1	0.0099	1010.0
2	0.0244	512.5
3	0.1429	233.3
4	0.0291	343.3
5	0.0164	508.3

□

The last two subsections and also Section 6.6 are strongly influenced by the kinetic tree theory (see box).

Kinetic Tree Theory

Kinetic tree theory (KTT) is an early method for quantitative FTA developed by William E. Vesely in the late 1960s. The approach presented in Section 6.7 has its roots in KTT, but KTT is much more than what we have presented here. Interested readers may consult Vesely's original work in Vesely (1970). The authors of this book wholeheartedly acknowledge Vesely's huge contributions to the development of reliability and risk analysis.

Several authors have pointed at weaknesses of the KTT, but still we consider KTT to be a huge achievement.

6.7.9 Binary Decision Diagrams

A binary decision diagram (BDD) is an alternative to – and an extension of – the approach described above for performing both qualitative and quantitative FTA. A BDD is a directed acyclic graph (DAG) with a single root. The BDD algorithm leads to an exact calculation of the TOP event probability and is not based on minimal cut sets.

The *truth table* was introduced in Section 4.3 as an alternative representation of a fault tree with only AND and OR gates. A truth table can be transferred to a BDD as illustrated in Example 6.24.

Example 6.24 (BDD deduced from a truth table)
Consider a fault tree of two basic events A and B that are connected by an OR gate. This means that the TOP event occurs when either A or B or both of them occur. Let 0 denote that an event does not occur and 1 that it occurs. The truth table for

Truth table

A	*B*	TOP
0	0	0
0	1	0
1	0	0
1	1	1

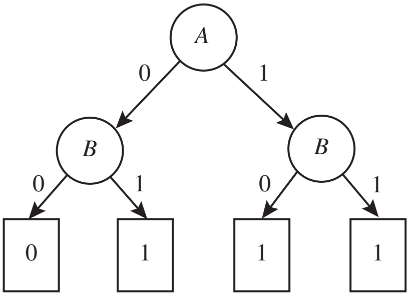


Figure 6.19 BDD deduced from a truth table.

this fault tree is shown in Figure 6.19 together with the associated BDD. Observe that the only case the TOP event does not occur (i.e. state 0) is that the basic events *A* and *B* both have state 0. Also observe that the BDD is established by stepwise pivotal decomposition (i.e. Shannon decomposition). □

Observe that any Boolean function can be represented by a BDD. The order of the nodes in the BDD can be different from the order of the variables in the Boolean expression. In Example 6.24, we started with node A, but could as well start with node B. A BDD satisfies the following conditions:

(1) All leaves are labeled by 0 or 1
(2) All nodes are labeled by a name (letters or numbers) and have exactly two children, a 0-child and a 1-child. It is common to label the edges leading to these children by 0 and 1, respectively. (Some authors indicate the type of edge (arrow) by using two different line-styles.)
(3) The root of the tree (i.e. the top node) does not have any parents.

A BDD is a compact representation and has been used since the 1990s to compute large fault trees. It is actually very well adapted for numerical storage and algorithmic treatments. It is obtained through repeated application of two compression rules:

(1) Sharing of identical subtrees,
(2) Elimination of nodes for which the 0-child and 1-child coincide (i.e. redundant nodes).

These two rules are applied until all subtrees are different and there are no redundant nodes. The obtained BDD is said to be reduced and ordered, and we sometimes use the notion “reduced, ordered binary decision diagram” (ROBDD). The application of the two compression rules require a set of additional concepts and are not treated any further in this book. The numerical evaluation of the BDD is based on repeatedly using binary decompositions (Shannon decompositions) and several algorithms have been developed. Several computer programs for fault tree analysis now use BDDs in the quantitative evaluation of fault trees.

When the fault tree is established by the procedures outlined in Section 4.3, the BDD may be constructed from the fault tree by a bottom-up procedure. Each basic event is associated with single-node BDD with two children. Starting at the bottom of the tree, the BDD is constructed for each basic event, and then combined according to the logic defined by the gate. The BDD for the OR and AND gates are constructed by applying the OR and AND functions to the BDD. The NASA *Fault Tree Handbook* recommends using BDDs as part of FTA and discusses the pros and cons of the BBD approach compared to the minimal cut set approach that we have presented earlier in this chapter (see NASA 2002, p. 78-82).

A high number of slide presentations, lecture notes, and articles on BDD may be found by searching the Internet. Most of these are of high quality. A comprehensive and relevant treatment is given by Andrews and Remenyte (2005) and Xing and Amari (2015).

6.8 Event Tree Analysis

The basic theory and the construction of event trees are introduced in Section 4.4. The current section deals with the quantification of event trees, mainly through an example.

When input data are available for the initiating event and all the relevant safety functions and hazard contributing factors, a quantitative analysis of the event tree may be carried out to give frequencies or probabilities of the resulting consequences.

The occurrences of the initiating event is usually modeled by a homogeneous Poisson process with frequency λ, which is measured as the expected number of occurrences per year (or some other time unit). Homogeneous Poisson processes are further discussed in Chapter 10.

For each safety function, we have to estimate the conditional probability that it will function properly in the relevant context, that is, when the previous events in the event chain have occurred. Some safety functions, such as emergency shutdown (ESD) systems on offshore oil/gas platforms, may be very complicated and will require a detailed reliability analysis.

The (conditional) reliability of a safety function depends on a wide range of environmental and operational factors, such as loads from previous events in the event chain, and the time since the last function test. In many cases, it is difficult to distinguish between “functioning” and “nonfunctioning.” A fire pump may, for example, start, but stop prematurely before the fire is extinguished.

The reliability assessment of a safety function may in most cases be performed by an FTA or an analysis based on RBD. If the analysis is computerized, a link may be established between the reliability assessment and the appropriate node in the event tree to facilitate automatic updating of the outcome frequencies and sensitivity analyzes. It may, for example, be relevant to study the effect on the outcome frequencies by changing the testing interval of a safety valve. Graphically, the link may be visualized by a transfer symbol on one of the output branches from the node.

The probabilities of the various hazard contributing factors (events/states) that enter into the event tree must be estimated for the relevant contexts. Some of these factors may be independent of the previous events in the event chain, whereas others are not.

It is important to observe that most of the probabilities in the event tree are *conditional* probabilities. The probability that the sprinkler system in Figure 4.11 will function is not equal to a probability that is estimated based on tests in the normal operating context. We have to take into account that the sprinkler system may have been damaged during the dust explosion and the first phase of the fire (i.e. before it is activated).

Consider the event tree in Figure 4.11. Let λ_A denote the frequency of the initiating event A, "explosion." In this example, λ_A is assumed to be equal to 10^{-2} per year, which means that an explosion on the average will occur once every 100 years. Let B denote the event "start of a fire," and let $\Pr(B) = 0.8$ be the conditional probability of this event when a dust explosion has already occurred. A more correct notation would be $\Pr(B \mid A)$ to make clear that event B is considered when event A has already occurred.

In the same way, let C denote the event that the sprinkler system does not function, following the dust explosion and the outbreak of a fire. The conditional probability of C is assumed to be $\Pr(C) = 0.01$

The fire alarm will not be activated (event D) with probability $\Pr(D) = 0.001$. This example assumes that this probability is the same whether the sprinkler system is functioning or not, but in most cases, the probability of this event would depend on the outcome of the previous event.

Let B^*, C^*, and D^* denote the negation (nonoccurrence) of the events B, C, and D respectively. We know that $\Pr(B^*)$ is equal to $1 - \Pr(B)$, and so on.

The frequencies (per year) of the end consequences may now be calculated as follows:

(1) Uncontrolled fire with no alarm:

$$\lambda_4 = \lambda_A \Pr(B) \Pr(C) \Pr(D) = 10^{-2} \cdot 0.8 \cdot 0.01 \cdot 0.001 \approx 8.0 \times 10^{-8}.$$

(2) Uncontrolled fire with alarm:

$$\lambda_3 = \lambda_A \Pr(B) \Pr(C^*) \Pr(D) = 10^{-2} \cdot 0.8 \cdot 0.01 \cdot 0.999 \approx 8.0 \times 10^{-5}.$$

(3) Controlled fire with no alarm:

$$\lambda_2 = \lambda_A \Pr(B) \Pr(C^*) \Pr(D) = 10^{-2} \cdot 0.8 \cdot 0.99 \cdot 0.001 \approx 7.9 \times 10^{-6}.$$

(4) Controlled fire with alarm:

$$\lambda_1 = \lambda_A \Pr(B) \Pr(C^*) \Pr(D^*) = 10^{-2} \cdot 0.8 \cdot 0.99 \cdot 0.999 \approx 7.9 \times 10^{-3}.$$

(5) No fire:

$$\lambda_5 = \lambda_A \Pr(B^*) = 10^{-2} \cdot 0.2 \approx 2.0 \times 10^{-3}.$$

It is seen that the frequency of a specific outcome (consequence) simply is obtained by multiplying the frequency of the initiating event by the probabilities along the event sequence leading to the outcome in question.

If we assume that occurrences of the initiating event may be described by a homogeneous Poisson process, and that all the probabilities of the safety functions and hazard contributing factors are constant and independent of time, then the occurrences of each outcome will also follow a homogeneous Poisson process.

As for FTA, event trees may also be converted to and analyzed by BDDs (e.g. see Andrews and Dunnett 2000).

6.9 Bayesian Networks

A qualitative introduction to Bayesian networks (BN) is given in Section 4.8. The current section gives a brief introduction to probabilistic evaluation of BNs. We tacitly assume that all nodes represent a state of an item and that each item has only two states 1 (= functioning) or 0 (= failed). It is also assumed that each node corresponds to a random variable with the same symbol as the node. A random variable A is said to *represent* node A. Observe that this is similar to an RBD, where the (random) state variable X_i represents component i.

Consider the simple BN in Figure 6.20. Node (and variable) A is said to influence node (and variable) B. The random variable A represent a root node A. Node A is a *parent* of node B, and node B is a *child* of node A. In a BN set-up, the distribution of A is given as a table, such as Table 6.3, where probability values are entered for illustration.

In a BN, the Bayesian interpretation of probability as "degree of belief" is adopted. The value of $\Pr(A = 1) = p_A$ is determined based on our knowledge about A. This interpretation of probability is further discussed in Chapter 15. A BN

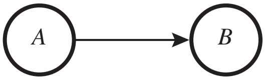


Figure 6.20 Simple BN with two nodes.

Table 6.3 Prior probability of root node *A*.

A	**Pr(*A*)**
0	0.1
1	0.9

provides a structured, graphical representation of the probabilistic relationships between the random variables (nodes).

6.9.1 Influence and Cause

The arc from node A to node B in Figure 6.20 means that node B is directly influenced by node A. This influence is sometimes referred to causal influence, even though statisticians are generally reluctant to use the word *causal*. For A to be a cause of B, the following three conditions need to be fulfilled:

(1) There is correlation between A and B.
(2) There is a temporal asymmetry (precedence) – one is occurring before the other.
(3) There is no hidden variable explaining the correlation.

Very often, we observe that components A and B are correlated, but where a thorough analysis shows that they are both influenced by a common cause, which may not be easy to identify. Correlation does not always imply causation. For a thorough discussion, see Pearl (2009).

6.9.2 Independence Assumptions

Recall that two random variables A and B are *independent* if

$$\begin{aligned} &\Pr(A = a \mid B = b) = \Pr(A = a) \quad \text{and} \\ &\Pr(B = b \mid A = a) = \Pr(B = b) \quad \text{for all } a \text{ and } b \end{aligned}$$

When the conditional probabilities $\Pr(A = a \mid B = b) \neq \Pr(A = a)$ or $\Pr(B = b \mid A = a) \neq \Pr(B = b)$, the two variables are *dependent*. Dependence is discussed further in Chapter 8.

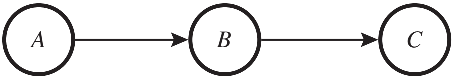


Figure 6.21 Linear BN with three nodes.

The BN approach is based on the assumption that the state of a node, say X in is influenced only by the states of its parents. This means that the state of X is independent of the states of all other nondescendant nodes of X, given that the states of the parents of X are known. We say that each node of a BN fulfills the *local Markov property* (see Chapter 11). In Figure 6.21, this assumption implies that when the state of node B is known, node C is independent of node A. The joint distribution of the variables A, B, and C, hence, can be written as

$$\begin{aligned}&\Pr(A = a \cap B = b \cap C = c)\\ &\qquad= \Pr(C = c \mid B = b)\Pr(B = b \mid A = a)\Pr(A = a),\end{aligned} \tag{6.106}$$

where a, b, and c are given values in $\{0, 1\}$.

6.9.3 Conditional Probability Table

A BN describes how a node *directly influences* other nodes. The nodes (and variables) of a BN are generally not independent, and we have to make use of conditional probabilities.Consider the simple BN in Figure 6.21, where the root node A has a direct influence on B. This influence is specified as a *conditional probability table* (CPT) as shown in Table 6.4. Again, probability values are included as illustration.

When component A is failed ($A = 0$), the CPT in Table 6.4 says that component B will fail with probability 0.7 and function with probability 0.3. Observe that for a given state of A, the conditional probabilities of B must add up to 1.

If we have observed that component B failed, such that $B = 0$, we may ask what is the probability that failure of B was caused (i.e. influenced) by a failure of component A. This is written as $\Pr(A = 0 \mid B = 0)$ and may be determined by *Bayes'*

Table 6.4 Conditional probability table for two nodes.

A	B	$\Pr(B \mid A)$
0	0	0.7
0	1	0.3
1	0	0.1
1	1	0.9

formula.

$$\Pr(A = 0 \mid B = 0) = \frac{\Pr(B = 0 \mid A = 0)\Pr(A = 0)}{\Pr(B = 0)}, \tag{6.107}$$

where $\Pr(B = 0)$ is determined by the law of total probability

$$\Pr(B = 0) = \Pr(B = 0 \mid A = 0)\Pr(A = 0) + \Pr(B = 0 \mid A = 1)\Pr(A = 1).$$

The values in Tables 6.3 and 6.4 yields

$$\Pr(A = 0 \mid B = 0) = \frac{0.7 \cdot 0.1}{0.7 \cdot 0.1 + 0.1 \cdot 0.9} \approx 0.44.$$

6.9.4 Conditional Independence

As pointed out above, many variables in a BN are not independent. To study all types of dependencies is overwhelming for BNs with a high number of nodes and a limited type of dependence is therefore assumed. The analysis of BNs is therefore delimited to random variables (nodes) that are *conditionally independent* (see box).

Conditional Independence

The variables A and B are said to be *conditionally independent* given C if for all a, b, and a given value c

$$\Pr(A = a \cap B = b \mid C = c) = \Pr(A = a \mid C = c)\Pr(B = b \mid C = c).$$

The rule indicated in Eq. (6.106) is general and may be expressed as follows: Consider a BN with nodes $X_1, X_2, \dots, X_n$. The joint distribution of $X_1, X_2, \dots, X_n$ is

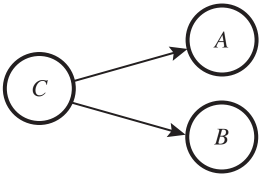
$$\Pr(X_1 = x_1 \ \cap X_2 = x_2 \cap \dots \cap X_n = x_n) \\ = \prod_{i=1}^{n} \Pr(X_i = x_i \mid \text{States of the parents of } X_i). \tag{6.108}$$

In the BN in Figure 6.22, C has a direct influence on both A and B. Because they are influenced by the same variable C, they are obviously dependent, but when the state of C is known, we assume that A and B are conditionally independent (see box).

If A and B are conditionally independent given C, then (for all a, b, and a given value c)

$$\begin{aligned}\Pr(A = a \mid B = b \cap C = c) &= \frac{\Pr(A = a \cap B = b \mid C = c)}{\Pr(B = b \mid C = c)} \\ &= \frac{\Pr(A = a \mid C = c)\Pr(B = b \mid C = c)}{\Pr(B = b \mid C = c)} \\ &= \Pr(A = a \mid C = c),\end{aligned}$$

Figure 6.22 BN with Three Nodes.



which means that the information that $B = b$ has no influence on $\Pr(A = a \mid C = c)$.

With this assumption, each node in the BN is conditionally independent of all its nondescendants given the value of its parents. Observe that a node X is not independent of its descendants when the states of its parents are known. Also observe that nodes that are not connected (i.e. there are no arc from one of the nodes to the other) are conditionally independent.

Example 6.25 (System of two pumps)
The BN of two identical pumps, A and B, with power supply C may be illustrated by Figure 6.22.

When the power supply is functioning (i.e. $C = 1$), we assume that the pumps are functioning independent of each other. If one pump fails, it has no influence on the functioning of the other pump. This means that when $C = 1$, the two pumps are conditionally independent.

$$\Pr(A = 1 \cap B = 1 \mid C = 1) = \Pr(A = 1 \mid C = 1)\Pr(B = 1 \mid C = 1).$$

We want to check whether or not conditional independence implies that the two pumps are independent.

The probability that the power supply is functioning is $\Pr(C = 1) = 0.95$. When the power supply is functioning, the probability that the pumps are functioning are $\Pr(A = 1 \mid C = 1) = 0.90$ and $\Pr(B = 1 \mid C = 1) = 0.90$. When the power supply is failed, the pumps cannot function. The law of total probability gives

$$\begin{aligned}\Pr(A = 1) = \Pr(A = 1 \cap B = 1 \mid C = 1)\Pr(C = 1)\\ + \Pr(A = 1 \cap B = 1 \mid C = 0)\Pr(C = 0)\\ = 0.9 \cdot 0.9 \cdot 0.95 + 0 \approx 0.77.\end{aligned}$$

In the same way, $\Pr(B = 1) \approx 0.77$. Further, by the law of total probability

$$\begin{aligned}\Pr(A = 1 \cap B = 1) = \Pr(A = 1 \cap B = 1 \mid C = 1)\Pr(C = 1)\\ + \Pr(A = 1 \cap B = 1 \mid C = 0)\Pr(C = 0)\\ = 0.9 \cdot 0.9 \cdot 0.95 + 0 \approx 0.77.\end{aligned}$$

We have found that $\Pr(A = 1) \Pr(B = 1) \approx 0.77 \cdot 0.77 \approx 0.59$, and $\Pr(A = 1 \cap B = 1) \approx 0.77$, such that $\Pr(A = 1 \cap B = 1) \neq \Pr(A = 1) \Pr(B = 1)$, and we conclude that the two pumps are *not* independent.

This conclusion may be justified intuitively. If we observe that pump A is not functioning, this can mean that pump A has failed, or that the power supply is not functioning. The information that $A = 0$, hence, increases the conditional probability that pump B is not functioning, which means that A and B are not independent. □

6.9.5 Inference and Learning

When we get information related to a BN in the form of expert judgments and observed data, we may use this information to make inference. Inference is further discussed in Chapters 14 and 15. Inference comprises computation of conditional probabilities, parameter estimation, and determination of posterior distributions.

Learning a BN based on data means to acquire knowledge about (i) the structure of the graphical model and (ii) the conditional probability distributions. The last option means to update a prior belief on the basis of the evidence (i.e. the data available). This is done by Bayes' formula, which is thoroughly discussed in Chapter 15. Learning the BN structure involves to learn causal relationships and to verify the correctness and the consistency of the structure. Learning may also involve to identify the most likely explanation of an item failure. It is further possible to determine the effect of an intervention into the system (e.g. to repair or modify a component).

Exact inference and learning is feasible only in small- to medium-sized BNs. For larger BNs, we have to suffice with approximative approaches usually based on Monte Carlo simulation, which are much faster and give pretty good results.

6.9.6 BN and Fault Tree Analysis

Because this chapter is delimited to systems of *independent* components, many important features fall outside the scope of the chapter. Dependent components are dealt with in Chapter 8.

Consider a system S of two independent components A and B. A BN for the system is shown in Figure 6.23. To compare BN analysis and fault tree analysis, let $A = 1$ denote that basic event A occurs, $B = 1$ that basic event B occurs, and $S = 1$ that the TOP event occurs. We want to find the TOP event probability Q_0 when the basic event probabilities q_A and q_B are specified.

The information about the basic events A and B is given in Table 6.5, where the probabilities are included for illustration. Observe that both A and B are root nodes.

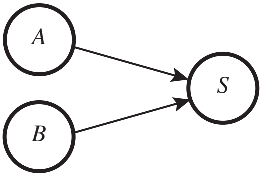


Figure 6.23 BN for a simple system of two components.

Table 6.5 Prior probability of the root nodes *A* and *B*.

A	*B*	Pr(*A*)	Pr(*B*)
0	0	0.99	0.96
1	1	0.01	0.04

For an AND-gate (i.e. a parallel structure), the TOP event occurs only when $A = 1$ AND $B = 1$, such that

$$Q_0 = \Pr(A = 1 \cap B = 1) = \Pr(A = 1)\Pr(B = 1) = 0.01 \cdot 0.04 = 4.0 \times 10^{-4}.$$

For an OR-gate (i.e. a series structure), the TOP event occurs if $A = 1$ OR $B = 1$, such that

$$\begin{aligned} Q_0 &= \Pr(A = 1 \cup B = 1) = \Pr(A = 1) + \Pr(B = 1) - \Pr(A = 1 \cap B = 1) \\ &= \Pr(A = 1) + \Pr(B = 1) - \Pr(A = 1)\Pr(B = 1) \\ &= 0.01 + 0.04 - 0.01 \cdot 0.04 = 4.96 \times 10^{-2} \end{aligned}$$

The strength of BNs becomes visible when the basic events are not independent, but influence each other and also when we have basic events that do not have only two states.

To illustrate the calculation procedure, consider the simple BN in Figure 4.38. This BN has three root nodes A, B, and C, each with two states, where state 1 means that the basic event occurs and state 0 that it does not occur. The probabilities of the root nodes are specified as a CPT, such as

A	*B*	*C*	Pr(*A*)	Pr(*B*)	Pr(*C*)
0	0	0	$1 - q_A$	$1 - q_B$	$1 - q_C$
1	1	1	q_A	q_B	q_C

where q_i is the probability that basic event i occurs, for $i = A, B, C$.

Node M is a child node with parents A and B. The truth table for M is

A	***B***	***M***
0	0	0
0	1	1
1	0	1
1	1	1

The probability distribution for M is then

$$\begin{aligned}\Pr(M = 0) &= \Pr[(A = 0) \cap (B = 0)] = (1 - q_A)(1 - q_B)\\ \Pr(M = 1) &= 1 - \Pr(M = 0) = 1 - (1 - q_a)(1 - q_B)\end{aligned}$$

which is the same result we got when using FTA. With $q_M = \Pr(M = 1)$, we can repeat the above arguments to find the probability distribution of T as a child of the parents M and C, and we get the probability

$$\begin{aligned}\Pr(T = 0) &= \Pr[(M = 0) \cap (C = 0)] = (1 - q_M)(1 - q_C)\\ &= (1 - q_A)(1 - q_B)(1 - q_C)\\ \Pr(T = 1) &= 1 - \Pr(T = 0) = 1 - (1 - q_A)(1 - q_B)(1 - q_C)\end{aligned}$$

Quantitative analysis of Bayesian networks is not discussed further in this book. For further information, see for example, Jensen and Nielsen (2007), Kjærulff and Madsen (2008), Scutari and Denis (2015), and Bobbio et al. (2001). Several R packages make it possible to use R for BN analysis, such as `bnlearn`. A number of both free and commercial computer programs for BN analysis are available. Programs may be found by searching the Internet.

6.10 Monte Carlo Simulation

Monte Carlo simulation got its name from the city of Monte Carlo in Monaco and its many casinos and is a computerized mathematical technique that generates random samples that are used to obtain numerical results. To illustrate the approach, consider an RBD with n independent and nonrepairable components. The time-to-failure T_i of component i is assumed to be Weibull-distributed with parameters α_i and θ_i, for $i = 1, 2, \dots, n$. All parameters are assumed to be known. We are interested in determining the distribution $F_S(t)$ of the time-to-failure T_S for the system. For a complicated system, this may be a difficult task.

By Monte Carlo simulation, a computer is used to generate a sequence of times-to-failure $t_1, t_2, \dots, t_n$ from the given distributions, for the n components.

Then, the structure function is used to determine a corresponding time-to-failure t_s for the system.

By repeating this simulation a large number (e.g. 1000) of times, we can use the methods described in Chapter 14 to obtain an empirical survivor function and to fit a continuous distribution to the data. This way, an estimate of $R_S(t)$ is obtained without cumbersome calculations. When having the system survivor function $F_S(t)$, the methods from Chapter 5 can be used to obtain estimates of the system failure rate function $z_S(t)$ and the system mean time-to-failure MTTF_S.

To illustrate how this is done in more detail, we start by describing what a random number is and how to obtain a sequence of random numbers.

6.10.1 Random Number Generation

A sequence of random values from the interval $(0, 1)$ can be obtained by using a *random number generator* on a computer. The sequence of numbers should be generated in such a way that each number has the same probability of assuming any of the values in the interval and be independent of the other numbers in the sequence. Random numbers generation is therefore similar to sampling from a uniform distribution unif(0, 1) with probability density function

$$f_Y(y) = \begin{cases} 1 & \text{for } 0 < y < 1 \\ 0 & \text{otherwise} \end{cases} .$$

The numbers (y) obtained by this procedure on a computer are not truly random because they are determined by an initial value called the *seed*. We therefore say that the numbers generated are *pseudo-random numbers*. A wide range of pseudo-random number generators are available (e.g. as part of spreadsheet programs and programs for statistical analysis, such as R). Most of these pseudo-random number generators are able to generate variables $Y_1, Y_2, \dots$ that are approximately independent with a uniform distribution over $[0, 1]$. In R, the command `runif(n)` generates a sequence of n pseudo-random numbers. Rerunning this command, we get another sequence of n pseudo-random numbers.

In Monte Carlo simulations for reliability analyses, we are mainly interested in generating pseudo-random numbers from a life distribution, such as the exponential or the Weibull distributions and from some downtime/repair distributions, such as the lognormal distribution. In the next paragraph, we illustrate how this is accomplished.

Generation of Random Variables with a Specified Distribution

Let T denote a random variable, not necessarily a time-to-failure, with distribution function $F_T(t)$ that is strictly increasing for all t, such that $F_T^{-1}(y)$ is uniquely

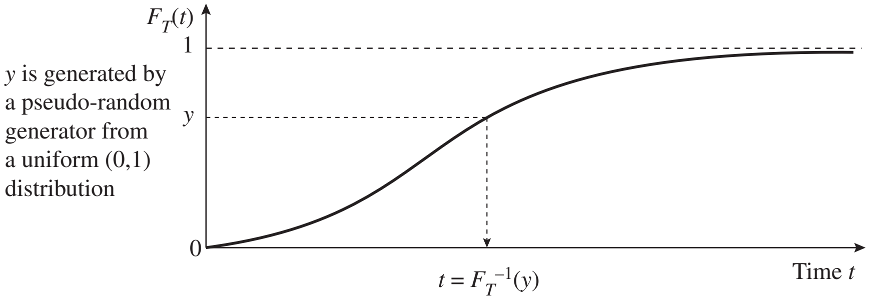


Figure 6.24 Generation of a random variable with distribution $F_T(t)$.

determined for all $y \in (0,1)$. Further let $Y = F_T(T)$. Then the distribution function $F_Y(y)$ of Y is

$$\begin{aligned} F_Y(y) &= \Pr(Y \le y) = \Pr(F_T(T) \le y) \\ &= \Pr[T \le F_T^{-1}(y)] = F_T[F_T^{-1}(y)] = y \qquad \text{for } 0 < y < 1. \end{aligned}$$

Hence, $Y = F_T(T)$ has a uniform distribution over $(0,1)$. This implies that if a random variable Y has a uniform distribution over $(0,1)$, then $T = F_T^{-1}(Y)$ has the distribution function $F_T(t)$.

This result can be used to generate random variables $T_1, T_2, \dots$ with a specified distribution function $F_T(t)$ on a computer. Variables $Y_1, Y_2, \dots$ that are uniformly distributed over $(0,1)$, may be generated by a pseudo-random number generator. The variables $T_i = F_T^{-1}(Y_i)$ for $i = 1, 2, \dots,$, then have distribution function $F_Y(t)$. The generation of random variable is illustrated in Figure 6.24.

In R, a sequence of n pseudo-random numbers from a Weibull distribution with given shape parameter (α) and scale parameter (θ) is obtained by the command `rweibull(n, shape, scale)`.

Simulating the Lifespan of a Repairable Component

Consider a repairable component where each repair brings the component back to an as-good-as-new state. If there is no trend or external influences, the lifespan of the component may be illustrated as in Figure 6.11. If the time-to-failure distribution and the downtime distribution are specified, it is easy to simulate a specified lifespan of, for example $\tau = 20$ years for the component. The simulation of a single lifespan gives the outcome $\{t_1, d_1, t_2, d_2, \dots\}$ until the total lifespan τ is reached, where t_i is an uptime and d_i is a downtime, for $i = 1, 2, \dots$. Based on this dataset, we may calculate the observed availability as $\sum_i t_i / \tau$, the total number of failures and the total downtime. By repeating this procedure a large number of times and by taking averages, we obtain accurate estimates of the availability and the associated metrics. This may seem to be a waste of time for a single component, because we can obtain the exact values by simple hand calculation.

The power of the Monte Carlo simulation approach becomes clear when studying a complicated system with many components and many different uptime and downtime distributions. Many computer programs for system reliability assessment are therefore based on Monte Carlo simulation.

6.10.2 Monte Carlo Next Event Simulation

Monte Carlo's next event simulation is carried out by simulating "typical" lifespans or mission scenarios for a system on a computer. We start with a model of the system, such as an RBD. Random events (i.e. events associated with item failures) are generated in the computer model and scheduled events (e.g. proof testing and servicing) and conditional events (i.e. events initiated by the occurrence of other events) are included to create a simulated mission scenario that is so close to a real lifetime scenario as possible.

Applications may demand different types of input data. In the oil and gas industry, this type of simulation is often used in the design phase to determine the production availability of suggested design options. Production availability may be measured as the number of barrels of oil or the number of cubic meters of gas produced per day and may be compared to the agreed sales volume, which may vary with the time of the year.

For most applications, the following input data to the simulation must be available:

- A description of the system based on flow diagrams, control schematics, and component information.
- Knowledge of component failure modes, failure effects, and failure consequences, usually in the form of an FMECA.
- Component failure and repair data (failure mode specific uptime and downtime distributions and estimates of the required parameters).
- Maintenance strategies. Frequency of preventive maintenance/servicing and duration of each type of maintenance action.
- Resource data (e.g. availability of spare parts and maintenance resources).
- Decision rules – what is to be done when a component failure mode occurs?
- Throughput data and system/component capacities (if relevant).

When a "typical" lifetime scenario has been simulated on the computer, this scenario is treated as a real experiment, and performance measures are calculated. We may, for example, calculate

- The observed availability of the system in the simulated time period (e.g. the observed uptime divided by the length of the simulated period).
- The number of system failures.
- The number of failures for each component.
- The contribution to system unavailability from each component.
- The use of maintenance resources.
- The system throughput (production) as a function of time.
- And several more.

The simulation can be repeated to generate a number of "independent" lifetime scenarios. From these scenarios, we may deduct estimates of the performance metrics of interest.

Single Item with Only One Failure Mode

We illustrate the next event simulation technique by a very simple example, a single repairable item with only one failure mode. A lifetime scenario for the item may be simulated as follows:

(1) The simulation is started at time $t = 0$ (the simulator clock is set to 0 that corresponds to a specified date). The item is assumed to be functioning at time $t = 0$.
(2) The time t_1 to the first failure is generated from the life distribution $F_{T_1}(t)$ that is specified by the analyst. The simulator clock is now set to t_1.
(3) The downtime d_1 is generated from a specified repair time distribution $F_{D_1}(d)$ that is specified by the analyst, and may, for example depend on the season (the date) and the time of the day of the failure. The repair time may, for example be longer for a failure that occurs during the night than for the same failure occurring during ordinary working hours. The simulator clock is now set to $t_1 + d_1$.
(4) The uptime t_2 to the second failure is generated from a the life distribution $F_{T_2}(t)$. The item may not be as-good-as-new after the repair action and the life distribution $F_{T_2}(t)$ may therefore be different from $F_{T_1}(t)$. The simulator clock is set to $t_1 + d_1 + t_2$.
(5) The downtime d_2 is generated from a specified repair time distribution $F_{D_2}(d)$,
(6) and so on.

The simulation is continued until the simulator clock reaches a predefined time, for example 10 years. The computer creates a chronological log file where all events (failures, repairs) and the (simulator clock) time for each event are recorded. From this log file, we are able to calculate the number of failures in the simulated period, the accumulated use of repair resources and utilities, the observed availability, and so on, for this specific life scenario. The observed availability A_1 is, for example calculated as the accumulated time the item has been functioning divided by the length of the simulated period.

The simulation described above is repeated n times (with different *seed* values), and the parameters of interest are calculated for each simulation. Let A_i be the observed availability in simulation i for $i = 1, 2, \dots, n$. The average availability A of the item is then calculated as the sample mean $\sum_{i=1}^{n} A_i/n$. The sample standard deviation may be used as a measure of the uncertainty of A. It is possible to split the simulation period into a number of intervals and calculate the average availability within each interval. The availability may, for example, be reported per year. A variety of approaches to reduce the variation in the estimates are available. For a comprehensive introduction to Monte Carlo simulation (e.g., see Rubinstein and Kroese 2017).

The simulation on a computer can theoretically take into account virtually any aspects and contingencies of an item:

- Seasonal and daily variations
- Variations in loading and output
- Periodic testing and interventions into the item
- Phased mission schemes
- Planned shutdown periods
- Interactions with other components and systems
- Dependencies between functioning times and downtimes

6.10.3 Simulation of Multicomponent Systems

Simulation of a mission scenario for a system with a high number of components requires a lot of input data to the computer. In addition, we have to establish a set of decision rules for the various events and combinations of events. These rules must state which actions should be a consequence of each event. Examples are decision rules related to

- Setting priorities between repair actions of simultaneous failures when we have limited repair resources.
- Switching policies between standby items.
- Deciding to replace or refurbish some additional components of the same subsystem when a component fails.
- Deciding to shut down the whole subsystem after a failure of a component, until repair action of the component is completed.

To obtain estimates of satisfactory accuracy, we have to simulate a rather high number of life histories of the system. The number of replicated simulations depends on how many components the system has and the reliabilities of the various system components. Systems with a high reliability will in general require more replications than systems with low reliability. The simulation time will be especially long when the model involves extremely rare events with extreme consequences. For multicomponent systems, we may need several thousands of replications. The simulation time is often excessive even on a fast computer, and the log file may become very large.

Next event simulation of simple systems can in principle be accomplished by using a spreadsheet program together with a Visual Basic code. Most spreadsheet programs have a random generator and a library of statistical distribution, and it is easy to generate random values from a specific distribution. The values simulated may next be combined according to given rules by using standard spreadsheet operations. Even more options are available by using R.

A number of simulation programs have been developed for availability assessment of specific systems. A list of program vendors may be found on the `book companion site`.

Example 6.26 (Production availability simulation)

Consider a system of two production items as illustrated in Figure 6.25. When both items are functioning, 60% of the system output comes from item 1 and 40% from item 2. The system is started up on a specific date (e.g. 1 January 2020). The times-to-failure are assumed to be independent and Weibull distributed with known parameters (α_i, λ_i), for $i = 1, 2$. The simulation is started by generating two Weibull distributed times-to-failure t_1 and t_2. Let us assume that $t_1 < t_2$. From time t_1, item 1 is out of operation during a random downtime that has a lognormal distribution with known parameters (ν_1, τ_1) that depends on the date at which item 1 failed. The production from item 2 is increased to 60% to partly compensate for the outage of item 1. The time-to-failure of item 2 with 60% production is Weibull-distributed with parameters (α_2, λ_2^1). (A conditional Weibull distribution might be selected.) The next step of the simulation is to generate the repair time d_1 of item 1, and the time-to-failure t_2^1 of item 2 with increased production. Let us assume that $d_1 < t_2^1$. At time d_1, item 1 is put into operation again, with 60% production, and the load on item 2 is reduced to 40%. Time-to-failure distributions are allocated to the two items. Conditional distributions, given the time in operation, may be used. New times-to-failure are generated according to the same procedure as described above. Periodic stops with adjustments, cleaning, and

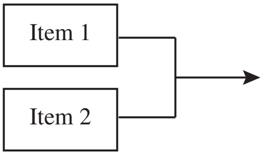


Figure 6.25 System of two production items.

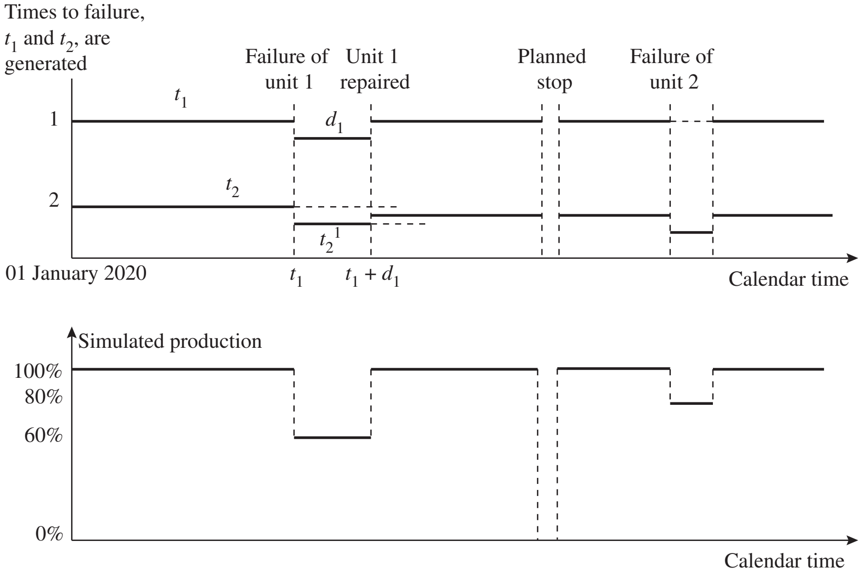


Figure 6.26 Simulation of the performance of the production system in Figure 6.25.

lubrication, may easily be included in the simulation. If item 2 fails, the load on item 1 is increased to 80%. The simulation is illustrated in Figure 6.26 together with the resulting simulated production. Several other metrics may be recorded, such as total item downtime, use of repair resources, and spare parts. The simulation of times-to-failure may further be split into different failure modes. The simulation is repeated a large number of times to give average values. □

6.11 Problems

6.1 Show that when the components are independent, the system reliability $p_S(t)$ may be written by Eq. (6.5), that is, as a function of the component reliabilities, $p_i(t)$ $(i = 1, 2, \dots, n)$, only.

6.2 An old-fashioned string of Christmas tree lights has 10 bulbs connected in series. The 10 identical bulbs are assumed to have independent times-to-failure with constant failure rate λ. Determine λ such that the probability that the string survives three weeks is at least 99%.

6.3 Consider three identical and independent items in parallel. What is the system reliability when each item has a reliability of 98%?

6.4 A system consists of five identical and independent components in parallel. Determine the reliability of the components such that the system reliability is 99%.

6.5 Consider the failure rate function $z_S(t)$ for a 2oo3 structure of independent and identical components with constant failure rate λ. Show that $\lim_{t\to\infty} z_S(t) = 2\lambda$, and give a physical explanation of why this is a realistic limit.

6.6 Consider a coherent structure of n independent components with system survivor function $R_S(t) = h[R_1(t), R_2(t), \dots, R_n(t)]$. Assume that all the n components have life distributions with increasing failure rate (IFR), and that the mean time-to-failure of component i is $\text{MTTF}_\text{i} = \mu_i$, for $i = 1, 2, \dots, n$. Show that

$$R_S(t) \geq h(e^{-t/\mu_1}, e^{-t/\mu_2}, \dots, e^{-t/\mu_n}) \qquad \text{for } 0 < t < \min\{\mu_1, \mu_2, \dots, \mu_n\}.$$

6.7 Consider the RBD in Figure 6.27.

(a) Find the minimal cut sets of the structure.

(b) Determine the availability of the system when the components are independent, nonrepairable, and:

- Component C1 has constant failure rate λ.
- Components Ci, for $i = 2, 3, 4, 5$, have constant probability q_i to be failed.

(c) Calculate the system availability at time $t = 5000$ hours, when $\lambda = 0.01$ failures per hour, and $q_i = 0.1$.

6.8 The mean number of failures per 10^6 hours of an item A is 100 and the mean time-to-first-failure for an item B is 100 days. Let MTTF_A and MTTF_B be the MTTF of items A and B, respectively.

A system S is functioning if and only if at least one item A and one item B in a series structure are functioning.

(a) What is the reliability of S at times $t = \text{MTTF}_A$ and $t = \text{MTTF}_B$? Comment your results.

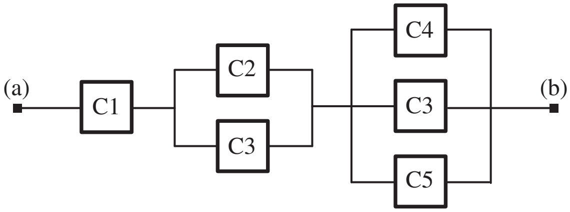
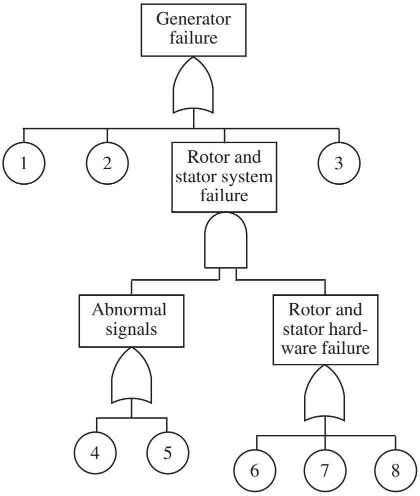


Figure 6.27 RBD for Problem 6.7.

(b) What is the MTTF of S, MTTF_S?
(c) What is the probability that S survives MTTF_S? Comment your result.
(d) To increase the reliability of S, we want to re-design the structure and wonder which one of the two following options would be the best: (i) add a redundant item A or (ii) add a redundant item B. Explain which option is most reliable and determine the survivor function of this design option at time $t = \text{MTTF}_A$ and at time $t = \text{MTTF}_B$.

6.9 You are going to make a quantitative reliability analysis of the generator of an offshore wind turbine. The generator converts mechanical energy to electrical energy, and adapts the output energy from the wind turbine to the grid. A simplified fault tree for this system is shown in Figure 6.28 (all items are considered to be nonrepairable). The basic events and their constant occurrence rates are given in Figure 6.28.
(a) Establish the corresponding RBD.
(b) Determine the structure function.
(c) Find the minimal cut sets.
(d) Determine the generator unreliability at time t =10 000 hours. Do you need an approximation?

6.10 Consider a water storage tank that is supposed to provide enough water in case of fire. A sensor is installed to monitor the water level. It sends a signal to the control unit if the water level is *below* a critical level, and the


Generator failure
1
2
Rotor and stator system failure
3
Abnormal signals
Rotor and stator hardware failure
4
5
6
7
8

Event	Code	Failure rate (failures per hour)
Parameter deviation	1	1×10^{-5}
Wire fault	2	1×10^{-7}
External facilities media leak	3	8×10^{-5}
Abnormal vibration	4	2×10^{-6}
Abnormal instrument reading	5	2×10^{-6}
Fail to syncronize	6	3×10^{-6}
Broken bars	7	2×10^{-7}
Fail to start on demand	8	3×10^{-6}

Figure 6.28 Fault tree for Problem 6.9.

Table 6.6 Table for Problem 6.10.

Item	Symbol	Failure mode	Failure rate (per hour)
Sensor	S	Nondetection of low water level	2×10^{-7}
Control unit	C	Fail to function	1×10^{-8}
Pumps	P_1,P_2	Fail to function	2×10^{-6}
Low power supply	L	Cut of low power supply	1×10^{-3}
High power supply	H	Cut of high power supply	1×10^{-5}

control unit activates the filling function. This function is carried out by two electrical pumps. Each pump is able to provide the required water flow. Only one pump is activated, and if it fails, the second pump is activated. If the second pump also fails, the filling function cannot be fulfilled anymore. The switching is assumed to be perfect (no fail to start). Some valves are involved in the filling function but are considered to be outside the system boundaries in this study. The sensor and the control unit need supply from a low power circuit (12 V), and the pumps need supply from the high power circuit (240 V).

The failure modes in Table 6.6 are considered. The switches for the two pumps are supposed to be perfect. All the failure modes are supposed to be independent.

(a) Carry out a functional analysis of the system.

(b) Establish a RBD.

(c) Find the minimal cut sets.

6.11 Assume that you are to evaluate the reliability of a heat exchanger used on an offshore oil and gas paltform OREDA (2015) gives the following numbers for all failure modes:

- Mean number of failures per 10^6 hours: 96.93
- Standard deviation for the number of failures per 10^6 hours: 35.81

(a) Assume that the global failure rate of the heat exchanger is constant.

  i. Estimate the failure rate by using the mean number of failures.

  ii. Show that the time-to-failure is exponentially distributed and plot the corresponding density function.

  iii. Calculate the MTTF.

  iv. Find the survivor function as a function of time and make a plot.

(b) During the design phase, the designers wonder what would be the benefit of using two heat exchangers instead of one. You are to quantify the extra lifetime obtained by two exchangers, in order to balance it with the equipment cost. We make the same assumptions as for question (a), and both heat exchangers are assumed to have the same constant failure rate:

i. First concept (active redundancy, no repairs): two exchangers are used all the time, such that the whole system may be considered as failed when both of them are in a failed state. Determine the survivor function of the whole system at different times and make a plot.
ii. Second concept (passive redundancy, perfect switching, no repairs): one exchanger is used first, and the other one is only started when the first one is failed. The whole system is considered to be failed when both heat exchangers have failed. Determine the survivor function of the whole system at different times and make a plot. You may use the analytical formulas in Section 6.4.1 or Monte Carlo simulation.
iii. Discuss and compare the survivor function of the two concepts.

(c) What do we mean by the "standard deviation for the number of failures per 10^6 hours?" How can we take this standard deviation into account when answering the previous questions?

6.12 Consider the primary cooling system of a nuclear power plant, comprising a lithium loop that circulates and removes the heat from the deuteron beam. The flow of the lithium is regulated by a flow controller (FC1), which controls a pump (P) that is driven by electrical power (EP). The lithium loop also contains a system (SHP) for maintaining the high purity of the lithium required for avoiding plugging, or corrosion and leakages. In addition, there is a trace heating system (STH) driven by electric power (EP) to maintain the temperature throughout the loop above the melting point of the lithium. In case of shutdown of electric power, an electrical generator (EG) is installed to take over.

(a) Construct a fault tree for the TOP event: "Loss of primary cooling system." Please add extra assumptions if required. The TOP event may occur if the lithium flow in the loop is not sufficient. You are to consider the failures of FC1, P, EP, EG, SHP, STH.

(b) Assume that none of the items is repairable. The failure rates of FC1, P, SHP, STH equal 10^{-5} failures per hour, the failure rate of EP equals 10^{-4} failures per hour, and the failure rate of EG equals 10^{-3} failures per hour.

i. Find the TOP event probability as a function of time by using one of the methods presented in Chapter

ii. Find the value of the survivor function of the system at times 1000 hours and 100 000 hours.

(c) Assume now that EP and EG are repairable with a repair rate equals 10^{-1} failures per hour for both of them. After a repair, the item is considered to be as-good-as-new.

i. Find the TOP event probability as a function of time by using Eq. (6.55) for EP and EG and the structure function.

ii. Find the TOP event probability by Monte Carlo simulation at times 1000 hours and 100 000 hours and compare with the results obtained results in question (b).

References

Andrews, J.D. and Dunnett, S.J. (2000). Event-tree analysis using binary decision diagrams. *IEEE Transactions on Reliability* 49 (2): 230–238.

Andrews, J.D. and Remenyte, R. (2005). Fault tree conversion to binary decision diagrams. *Proceedings of the 23rd International System Safety Conference*, San Diego, CA.

Barlow, R.E. and Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing, Probability Models*. New York: Holt, Rinehart, and Winston.

Bobbio, A., Portinale, L., Minichino, M., and Ciancamerla, E. (2001). Improving the analysis of dependable systems by mapping fault trees into Bayesian networks. *Reliability Engineering and System Safety* 71: 249–260.

ISO 20815 (2018). *Petroleum, petrochemical, and natural gas industries: Production assurance and reliability management*, *International standard*. Geneva: International Organization for Standardization.

Jensen, F.V. and Nielsen, T.D. (2007). *Bayesian Networks and Decision Graphs*, 2e. Berlin: Springer-Verlag.

Kawauchi, Y. and Rausand, M. (2002). A new approach to production regularity assessment in the oil and chemical industries. *Reliability Engineering and System Safety* 75: 379–388.

Kjærulff, U.B. and Madsen, A.L. (2008). *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis*. Berlin: Springer-Verlag.

NASA (2002). Fault Tree Handbook with Aerospace Applications, *Handbook*. Washington, DC: U.S. National Aeronautics and Space Administration.

NUREG-0492 (1981). Fault Tree Handbook, *Handbook NUREG-0492*. Washington, DC: U.S. Nuclear Regulatory Commission.

OREDA (2015). *Offshore and Onshore Reliability Data*, 6e. OREDA Participants, DNV GL, 1322 Høvik, Norway.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*, 2e. Cambridge: Cambridge University Press.

Rubinstein, R.Y. and Kroese, D.P. (2017). *Simulation and the Monte Carlo Method*. Hoboken, NJ: Wiley.

Satyanarayana, A. and Prabhakar, A. (1978). New topological formula and rapid algorithm for reliability analysis. *IEEE Transactions on Reliability* R-27: 82–100.

Scutari, M. and Denis, J.B. (2015). *Bayesian Networks: With Examples in R*. Boca Raton, FL: CRC Press, Taylor & Francis Group.

Vesely, W.E. (1970). A time-dependent methodology for fault tree evaluation. *Nuclear Engineering and Design* 13: 337–360.

Xing, L. and Amari, S.V. (2015). *Binary Decision Diagrams and Extensions for System Reliability Analysis*. Hoboken, NJ: Wiley and Scrivener Publisher.

7

Reliability Importance Metrics

7.1 Introduction

From Chapter 6, it should be obvious that some system components are more important for the system reliability than other components. A component in series with the rest of the system is a minimal cut set of order one and is generally more important than a component that is a member of a minimal cut set of higher order. This chapter defines and examines nine *component importance metrics*. The importance metrics may be used to arrange the components in order of increasing or decreasing importance, but also to classify the components into two or more groups according to some preset criteria. Importance metrics are mainly used for prioritizing components and modules for improvements and for maintenance planning and execution.

Most of the importance metrics are presented both in reliability block diagram (RBD) and fault tree notation. Chapters 4 and 6 demonstrate that fault trees with only OR and AND gates can be transferred to RBDs without losing information, and vice versa. If the fault tree has additional gates, it cannot be easily converted to an RBD, and we may not be able to define the important metrics in the same way. In this chapter, we tacitly assume that there is a one-to-one correspondence between an RBD and a corresponding fault tree.

In risk assessments, the causal analyses are usually based on fault trees, and the importance metrics that are used mainly in risk assessments are therefore presented only in fault tree notation. The importance metrics are then called *basic event importance metrics* or *risk importance metrics*.

System Reliability Theory: Models, Statistical Methods, and Applications, Third Edition.
Marvin Rausand, Anne Barros, and Arnljot Høyland.

Companion Website: www.wiley.com/go/SystemReliabilityTheory3e

7.1.1 Objectives of Reliability Importance Metrics

The main objectives of a reliability importance metric are to

(1) Identify the items that merit the most additional research and development in the design phase to improve overall system reliability at minimum cost or effort.
The reliability may be improved by using a higher-quality item, by introducing redundant items, by reducing the operational and environmental stresses on the item, or by improving the maintainability of the item.
(2) Identify the items that are most likely to cause system failure and therefore should be prioritized for inspection and maintenance.
(3) Identify the item(s) that, most likely, have caused a system failure. This metric may be used as input to the repairman's checklist in situations where it is important to restore the system function as fast as possible.
(4) Identify the components for which we need to obtain high-quality data during a safety or reliability analysis.
A component with low importance will have a very low influence on the system reliability. Spending resources to get very accurate data for such components may thus be a waste of money. A relevant approach is therefore first to calculate the system reliability and a relevant importance metric based on approximate (best guess) input parameters, and then concentrate the data acquisition resources on the most important components.
(5) Determine the increased risk or the reduced system reliability by taking an item out of service (e.g. for maintenance) when the system is running. This is a common application in, for example, nuclear power plants.

7.1.2 Reliability Importance Metrics Considered

Nine reliability importance metrics are defined and discussed in this chapter.

(1) Birnbaum's metric for structural importance
(2) Birnbaum's metric for component importance (and some variants)
(3) The improvement potential metric (and some variants)
(4) The criticality importance metric
(5) Fussell–Vesely's metric
(6) The differential importance metric
(7) Risk achievement worth
(8) Risk reduction worth
(9) Barlow and Proschan's metric for component importance

Many of the reliability importance metrics are developed for safety and reliability assessments for the nuclear industry, but have later been used in many other application areas. An overview of importance metrics used in *probabilistic risk assessments* (PRAs) of nuclear power plants is given by van der Borst and Schoonakker (2001). The reader can refer to Vu et al. (2016) for applications in maintenance optimization.

7.1.3 Assumptions and Notation

The following assumptions apply throughout Chapter 7.

(1) The structure $\phi[\boldsymbol{X}(t)]$ considered is coherent with n components.
(2) All components, subsystems, and the system have only two states: functioning (1) and failed (0).
(3) Components may be repairable or nonrepairable.
(4) All times-to-failure and all repair times have continuous distribution functions.
(5) All components are independent, both with respect to failures and repairs.
(6) The reliability of component i at time t is denoted $p_i(t)$. For a nonrepairable item, $p_i(t)$ is the survivor function $R_i(t)$ and for a repairable item, $p_i(t)$ is the availability $A_i(t)$ of the item.
(7) The system reliability, with respect to a specified system function, at time t is denoted $p_S(t) = h_S[p_1(t), p_2(t), \dots, p_n(t)]$. For a nonrepairable system, $p_S(t)$ is the system survivor function $R_S(t)$, and for a repairable system, $p_S(t)$ is the system availability $A_S(t)$.
(8) The corresponding unreliabilities are denoted

$$p_i^*(t) = 1 - p_i(t)$$
$$p_S^*(t) = 1 - p_S(t).$$

(9) The structure has k minimal cut sets $K_1, K_2, \dots, K_k$ that have been determined and are available.
(10) The rate of the occurrence of failures (ROCOF) of a repairable item i is denoted $w_i(t)$.
The fault trees considered have the same logical structure as the corresponding RBD.
(11) The fault trees have only AND and OR gates.
(12) All basic events relate to the same component failures as mentioned for the system structure.
(13) The following fault tree notation is used:

$q_i(t)$	The probability that basic event E_i occurs at time t
$Q_0(t)$	The probability that the TOP event occurs at time t
$Q_0(t \mid E_i)$	The probability that the TOP event occurs at time t when it is known that basic event occurs at time t

$Q_0(t \mid E_i^*)$ The probability that the TOP event occurs at time t when it is known that basic event does not occur at time t

$\check{Q}_j(t)$ The probability that minimal cut set j has failed at time t

(14) The probabilities of a structure and a fault tree relate as follows:

$$p_i(t) = 1 - q_i(t) = q_i^*(t)$$

$$h[\boldsymbol{p}(t)] = 1 - Q_0(t) = Q_S^*(t).$$

Additional assumptions are provided when needed in the text. A consequence of assumptions 4 and 5 is that failures (basic events) occur at distinct points in time. A system failure (TOP event) always coincide with the failure of a component (occurrence of a basic event), say component i. In this sense, we say that component i has *caused* system failure or that basic event i has *caused* the TOP event to occur.

When discussing *component importance*, the importance is always seen in relation to the specified *system function*. Most systems have many different functions. A component that has high importance with respect to a particular system function does not need to have high importance with respect to other system functions.

Two factors determine the importance of a component in a system:

- The structure of the system and where the component is located in the system,
- The reliability of the component.

Which of these is the most important depends on the importance metric used. Interested readers may consult (Vesely, 1998; NASA, 2011; Kuo and Zhu, 2012; La Rovere et al., 2013) for further information about importance metrics.

Remark 7.1 (An advice to the reader)
This chapter presents alternative definitions to many of the important metrics and also relationships between the various metrics. Some of the derivations may seem rather tedious, but if you stay on, you will get a deeper insight into many additional aspects of reliability analysis. □

7.2 Critical Components

A *critical component* is the basis for several important metrics. We therefore start by defining what a critical component is in both an RBD and a fault tree context.

Definition 7.1 (Critical component / basic event)

(1) Component i is critical for the system if the other $n - 1$ components are in such states that the system is functioning if and only if component i is functioning.

(2) Basic event E_i is critical for the TOP event if the other $n-1$ basic events are in such states that the TOP event occurs if and only if basic event E_i occurs. □

When saying that component i is critical, this is not a statement about component i, but rather a statement about the states of the other $n-1$ components of the system. Component i is critical for the system if the other $n-1$ components have states $(\cdot_i, \boldsymbol{x})$ such that $\phi(1_i, \boldsymbol{x}) = 1$ and $\phi(0_i, \boldsymbol{x}) = 0$, which can be written

$$\phi(1_i, \boldsymbol{x}) - \phi(0_i, \boldsymbol{x}) = 1.$$

A state vector $(\cdot_i, \boldsymbol{x})$ that makes component i critical is called a *critical state vector* for component i. The number of different critical state vectors for component i is

$$\eta_\phi(i) = \sum_{(\cdot_i, \boldsymbol{x})} [\phi(1_i, \boldsymbol{x}) - \phi(0_i, \boldsymbol{x})], \tag{7.1}$$

where the sum is taken over all possible state vectors $(\cdot_i, \boldsymbol{x})$. Because each state x_j can take only two values, the total number of distinct state vectors $(\cdot_i, \boldsymbol{x})$ is 2^{n-1}. When the state at time t is a random variable $\boldsymbol{X}(t) = [X_1(t), X_2(t), \dots, X_n(t)]$, component i is critical when

$$\phi[1_i, \boldsymbol{X}(t)] - \phi[0_i, \boldsymbol{X}(t)] = 1. \tag{7.2}$$

The probability that component i is critical is therefore

$$\Pr(\text{Component } i \text{ is critical at time } t) = \Pr(\phi[1_i, \boldsymbol{X}(t)] - \phi[0_i, \boldsymbol{X}(t)] = 1). \tag{7.3}$$

Example 7.1 (Critical component)

Consider the simple system of three components in Figure 7.1. Component 1 is seen to be critical if components 2 and 3 are in such states that the lower path in the RBD is failed. This is the case if component 2 or component 3 or both of them are failed. The states of components 2 and 3 that make component 1 critical are therefore

$$(\cdot_1, \boldsymbol{x}) = (\cdot, 0, 1) \qquad (\cdot_1, \boldsymbol{x}) = (\cdot, 1, 0) \qquad (\cdot_1, \boldsymbol{x}) = (\cdot, 0, 0).$$

This means that component 1 has $\eta_\phi(1) = 3$ critical state vectors. The critical state vectors are also shown in Table 7.1.

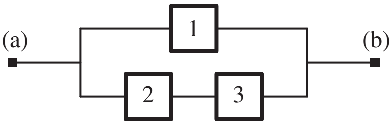


Figure 7.1 Simple system of three components.

Table 7.1 Critical path vectors for component 1.

x_1	x_2	x_3
$\cdot$	0	1
$\cdot$	1	0
$\cdot$	0	0

For component 2 to be critical, component 1 must be failed, and component 3 must be functioning. The states of component 1 and 3 that make component 2 critical are therefore

$$(\cdot_2, \boldsymbol{x}) = (0, \cdot, 1).$$

This means that component 2 has only $\eta_\phi(2) = 1$ critical state vector. The same applies for component 3. □

7.3 Birnbaum's Metric for Structural Importance

Birnbaum (1969) proposes the following metric for the structural importance of component i.

Definition 7.2 (Birnbaum's metric for structural importance)
The proportion of the total number of possible critical state vectors $\eta_\phi(i)$ relative to the total number of possible state vectors, 2^{n-1}

$$I_\phi^B(i) = \frac{\eta_\phi(i)}{2^{n-1}}. \tag{7.4}$$

□

Example 7.2 (Birnbaum's metric for structural importance)
Reconsider the simple system shown by the RBD in Figure 7.1. For component 1, we have

$(\cdot, x_2, x_3)$	$\phi(1, x_2, x_3) - \phi(0, x_2, x_3)$
$(\cdot, 0, 0)$	1
$(\cdot, 0, 1)$	1
$(\cdot, 1, 0)$	1
$(\cdot, 1, 1)$	0

In this case, the total number of critical state vectors for component 1 is 3:

$$\eta_\phi(1) = 3,$$

whereas the total number of possible state vectors is $2^{3-1} = 4$. Birnbaum's metric of structural importance of component 1 is therefore

$$I_\phi^B(1) = \frac{3}{4}.$$

Component 2 has only one critical state vector $(0, \cdot, 1)$ and Birnbaum's metric of structural importance becomes

$$I_\phi^B(2) = \frac{1}{4}.$$

Symmetrical reasoning yields

$$I_\phi^B(3) = \frac{1}{4}.$$

□

7.4 Birnbaum's Metric of Reliability Importance

Birnbaum (1969) proposes a metric of importance that can be defined in three different ways. The first definition is

Definition 7.3 (Birnbaum's metric of reliability importance - 1)
Birnbaum's metric[1] of reliability importance of (1) component i or (2) basic event E_i at time t is

$$(1)\quad I^B(i \mid t) = \frac{\partial h[\boldsymbol{p}(t)]}{\partial p_i(t)} \qquad \text{for } i = 1, 2, \dots, n. \tag{7.5}$$

$$(2)\quad I^B(i \mid t) = \frac{\partial Q_0(t)}{\partial q_i(t)} \qquad \text{for } i = 1, 2, \dots, n. \tag{7.6}$$

□

Birnbaum's metric is obtained by partial differentiation of the system reliability with respect to $p_i(t)$ [or $Q_0(t)$ with respect to $q_i(t)$]. This approach is well known from classical sensitivity analysis. If $I^B(i \mid t)$ is large, a small change in the reliability of component i results in a comparatively large change in the system reliability at time t.

When the reliability importance of component i is determined, all the other $n - 1$ components are assumed to have constant probabilities. When taking the derivative, all the other $n - 1$ probabilities p_j for $j \neq i$ are therefore treated as constants. We illustrate the derivation for a series and a parallel structure.

[1] Named after the Hungarian-American professor Zygmund William Birnbaum (1903–2000).

Assume that the component importance is to be determined at a given (future) time t. To simplify the notation, we suppress the time and write $p_i(t) = p_i$ for $1, 2, \dots, n$.

Example 7.3 (Series structure)
Consider a series structure of n independent components with reliabilities $\boldsymbol{p} = (p_1, p_2, \dots, p_n)$. The system reliability of the series structure is

$$h(\boldsymbol{p}) = \prod_{j=1}^{n} p_j = p_i \prod_{j \neq i} p_j. \tag{7.7}$$

In this context, the system reliability is expressed as p_i times a constant, and the derivative is

$$I^B(i) = \frac{\partial h(\boldsymbol{p})}{\partial p_i} = \prod_{j \neq i} p_j. \tag{7.8}$$

Consider a series structure of $n = 2$ independent components and assume that $p_1 = 0.90$ and $p_2 = 0.70$. Birnbaum's metric of importance of the two components becomes

$$I^B(1) = p_2 = 0.70$$
$$I^B(2) = p_1 = 0.90.$$

For this series structure, the component with the lowest reliability (i.e. component 2) is seen to be the most important. It is straightforward to show that this result applies for all series structures of independent components. To improve the reliability of a series structure, we should therefore focus our attention on the weakest component. □

Example 7.4 (Parallel structure)
Consider a parallel structure of n independent components with reliabilities $\boldsymbol{p} = (p_1, p_2, \dots, p_n)$. The system reliability of the parallel structure is

$$h(\boldsymbol{p}) = 1 - \prod_{j=1}^{n} (1 - p_j) = 1 - (1 - p_i) \prod_{j \neq i} (1 - p_j). \tag{7.9}$$

Because p_j for $j \neq i$ are treated as constants, the derivative becomes

$$I^B(i) = \frac{\partial h(\boldsymbol{p})}{\partial p_i} = \prod_{j \neq i} (1 - p_j). \tag{7.10}$$

Consider a parallel structure of $n = 2$ independent components and assume that $p_1 = 0.90$ and $p_2 = 0.70$. Birnbaum's metric of importance of the two components becomes

$$I^B(1) = 1 - p_2 = 0.30$$
$$I^B(2) = 1 - p_1 = 0.10.$$

For the parallel structure, the component with the highest reliability (i.e. component 1) is seen to be the most important. It is straightforward to show that this result applies for all parallel structures of independent components. To improve the reliability of a parallel structure, we should therefore–according to Birnbaum's metric–improve the most reliable component. □

Examples 7.3 and 7.4 indicate that components in a series structure are generally more important than components in a parallel structure.

7.4.1 Birnbaum's Metric in Fault Tree Analysis

The derivation of Birnbaum's metric of basic events in fault trees is similar to the derivation for components in RBDs.

Chapter 4 showed that a series structure corresponds to an OR-gate and that a parallel structure corresponds to an AND-gate in a fault tree. The derivations for a single AND-gate and a single OR-gate are shown in Examples 7.5 and 7.6, respectively.

Example 7.5 (Fault tree with a single AND-gate)

Consider a fault tree with a single AND-gate connecting n independent basic events $E_1, E_2, \dots, E_n$ with basic event probabilities $q_1, q_2, \dots, q_n$ (at a given time t). The TOP event probability is

$$Q_0 = \prod_{j=1}^{n} q_j = q_i \prod_{j \neq i} q_j. \tag{7.11}$$

Birnbaum's metric of importance of basic event E_i is

$$I^B(i) = \frac{\partial Q_0}{\partial q_i} = \prod_{j \neq i} q_j. \tag{7.12}$$

Remember that the q_js for $j \neq i$ are considered to be constants when deriving Birnbaum's metric. The basic event with the lowest basic event probability is by Birnbaum's metric the most important under an AND-gate. For an AND-gate, the TOP event (i.e. the output event of the AND-gate) occurs only when all the basic events under the AND-gate occur. The TOP event is avoided if at least one of the basic events do not occur. In this case, Birnbaum's metric tells us to focus on the basic event with the lowest probability of occurrence. □

Example 7.6 (Fault tree with a single OR-gate)
Consider a fault tree with a single OR-gate connecting n independent basic events $E_1, E_2, \dots, E_n$ with basic event probabilities $q_1, q_2, \dots, q_n$ (at a given time t). The TOP event probability is

$$Q_0 = 1 - \prod_{j=1}^{n}(1 - q_j) = 1 - (1 - q_i)\prod_{j \neq i}(1 - q_j). \tag{7.13}$$

Birnbaum's metric of importance of basic event E_i is

$$I^B(i) = \frac{\partial Q_0}{\partial q_i} = \prod_{j \neq i}(1 - q_j). \tag{7.14}$$

Again, the q_js for $j \neq i$ are considered to be constants when deriving Birnbaum's metric. The basic event with the highest basic event probability is by Birnbaum's metric the most important under an OR-gate. For an OR-gate, the TOP event (i.e. the output event of the OR-gate) occurs if any of the basic events under the OR-gate occurs. The TOP event is avoided only if none of the basic events occur. In this case, Birnbaum's metric tells us to focus on the basic event with the highest probability of occurrence. □

Similar to components in a structure, basic events under an OR-gate are generally more important – according to Birnbaum's metric – than basic events under an AND-gate.

7.4.2 A Second Definition of Birnbaum's Metric

Section 6.2.4 used *pivotal decomposition* to show that the system reliability $h(\boldsymbol{p})$ may be written as a linear function of p_i for $i = 1, 2, \dots, n$ when the n components are independent.

$$\begin{aligned} h(\boldsymbol{p}) &= p_i h(1_i, \boldsymbol{p}) + (1 - p_i) h(0_i, \boldsymbol{p}) \\ &= p_i[h(1_i, \boldsymbol{p}) - h(0_i, \boldsymbol{p})] + h(0_i, \boldsymbol{p}), \end{aligned} \tag{7.15}$$

where $h(1_i, \boldsymbol{p})$ is the (conditional) probability that the system is functioning when it is known that component i is functioning (at time t), and $h(0_i, \boldsymbol{p})$ is the (conditional) probability that the system is functioning when component i is in a failed state (at time t). From (7.15), Birnbaum's metric is

$$I^B(i) = \frac{\partial h(\boldsymbol{p})}{\partial p_i} = h(1_i, \boldsymbol{p}) - h(0_i, \boldsymbol{p}). \tag{7.16}$$

Remark 7.2 (Straight line)
Equation (7.15) shows that the system reliability $h(\boldsymbol{p})$ is a linear function of p_i as long as the reliability of the other components are considered to be constants.

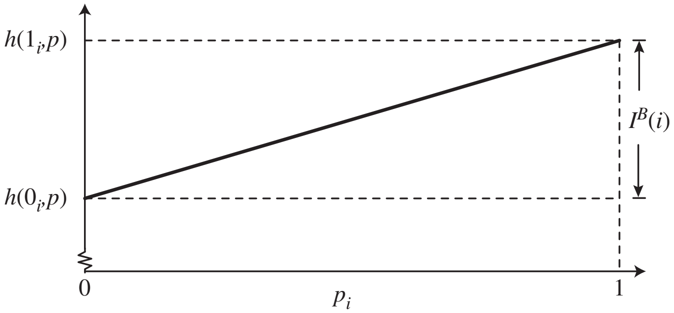


Figure 7.2 Illustration of Birnbaum's metric of reliability importance.

The derivative of a straight line, as shown in Figure 7.2, is a constant that may be calculated by considering the change over the whole interval [0, 1].

$$I^B(i) = \frac{h(1_i, \boldsymbol{p}) - h(0_i, \boldsymbol{p})}{1} = h(1_i, \boldsymbol{p}) - h(0_i, \boldsymbol{p}).$$

□

The result in (7.16) is derived from Definition 7.3, but many standards and guidelines prefer to define Birnbaum's metric by Eq. (7.16). We therefore formulate a second definition of Birnbaum's metric.

Definition 7.4 (Birnbaum's metric of importance - 2)
Birnbaum's metric for the importance of (1) component i or (2) basic event E_i at time t is

$$(1)\ \ I^B(i \mid t) = h[1_i, \boldsymbol{p}(t)] - h[0_i, \boldsymbol{p}(t))] \tag{7.17}$$

$$(2)\ \ I^B(i \mid t) = Q_0(t \mid E_i^*) - Q_0(t \mid E_i). \tag{7.18}$$

□

Definition 7.4 shows that Birnbaum's metric $I^B(i \mid t)$ of component i only depends on the structure of the system and the reliabilities of the other components. $I^B(i \mid t)$ is independent of the actual reliability $p_i(t)$ of component i. This may be regarded as a weakness of Birnbaum's metric.

The reason why many standards and guidelines prefer Definition 7.4 is twofold:

(1) Birnbaum's metric is easier to calculate from Definition 7.4, because we do not need to determine the derivative. Many fault tree programs use this approach. First, $Q_0(t \mid E^*)$ is calculated by setting $q_i(t) = 0$. Then $Q_0(t \mid E_i)$ is calculated by setting $q_i(t) = 1$. A simple subtraction then yields $I^B(i \mid t)$. This means that two separate recalculations of the TOP event probability are required to determine Birnbaum's metric for each basic event E_i for $i = 1, 2, \ldots, n$.

(2) The second reason is that Definition 7.4 can be used also for noncoherent structures and for systems (and fault trees) with dependent components (basic events).

The calculation of Birnbaum's metric in Definition 7.4 is briefly illustrated for a series and a parallel structure in Examples 7.7 and 7.8. Again, the time t is suppressed.

Example 7.7 (Series structure)
Reconsider the series structures in Example 7.3. When component i is functioning (i.e. $p_i = 1$), the series structure will function if and only if all the other $n-1$ components are functioning.

$$h(1_i, \boldsymbol{p}) = \prod_{j \neq i} p_j.$$

When component i is *not* functioning (i.e. $p_i = 0$), the series structure cannot function and $h(0_i, \boldsymbol{p}) = 0$. Birnbaum's metric of importance of component i in a series structure is

$$I^B(i) = h(1_i, \boldsymbol{p}) - h(0_i, \boldsymbol{p}) = \prod_{j \neq i} p_j.$$

□

Example 7.8 (Parallel structure)
Reconsider the series structures in Example 7.4. When component i is functioning ($p_i = 1$), the system is always functioning, that is $h(1_i, \boldsymbol{p}) = 1$. When component 1 is in a failed state ($p_i = 0$), the system is parallel structure of the other $n-1$ components, with reliability

$$h(1_i, \boldsymbol{p}) = 1 - \prod_{j \neq i} (1 - p_j).$$

Birnbaum's metric of importance of component i in a parallel structure is

$$I^B(i) = h(1_i, \boldsymbol{p}) - h(0_i, \boldsymbol{p}) = 1 - \left[1 - \prod_{j \neq i} (1 - p_j) \right] = \prod_{j \neq i} (1 - p_j).$$

□

The derivation of Birnbaum's metric is seen to be straightforward for both series and parallel structures and does not involve finding derivatives.

7.4.3 A Third Definition of Birnbaum's Metric

Section 6.2.4 shows that $h[\cdot_i, \boldsymbol{p}(t)] = E[\phi(\cdot_i, \boldsymbol{X}(t)]$, such that (7.17) can be written

$$\begin{aligned} I^B(i \mid t) &= h[1_i, \boldsymbol{p}(t)) - h(0_i, \boldsymbol{p}(t)] \\ &= E[\phi(1_i, \boldsymbol{X}(t)] - E[\phi(0_i, \boldsymbol{X}(t)] \\ &= E[\phi(1_i, \boldsymbol{X}(t) - \phi(0_i, \boldsymbol{X}(t)]. \end{aligned}$$

When $\phi[\boldsymbol{X}(t)]$ is a coherent structure, $(\phi[1_i, \boldsymbol{X}(t)] - \phi[0_i, \boldsymbol{X}(t)])$ can only take on the values 0 and 1. Birnbaum's metric (7.6) can therefore be written as

$$I^B(i \mid t) = \Pr(\phi[1_i, \boldsymbol{X}(t)] - \phi[0_i, \boldsymbol{X}(t)] = 1). \tag{7.19}$$

This means that $I^B(i \mid t)$ is equal to the probability that component i is critical for the system at time t (see Definition 7.1). A third definition of Birnbaum's metric is hence,

Definition 7.5 (Birnbaum's metric of importance – 3)
Birnbaum's metric of importance of component i at time t is equal to the probability that the system is in such a state at time t that component i is critical for the system. □

Definition 7.5 is not spelled out in the fault tree notation because this definition is seldom – if ever – used in fault tree analysis. We illustrate the use of the third definition by the same two examples as for the two first definitions. Again, the time t is suppressed.

Example 7.9 (Series structure)
Reconsider the series structure of n independent components in Example 7.3. For component i in the series structure to be critical, all the other $n - 1$ components have to function. Birnbaum's metric is hence

$$I^B(i) = \Pr(\text{Component } i \text{ is critical for the system}) = \prod_{j \neq i} p_i,$$

which is the same result as obtained by the two first definitions of Birnbaum's metric. □

Example 7.10 (Parallel structure)
Reconsider the parallel structure of n independent components in Example 7.4. For component i in the parallel structure to be critical, all the other $n - 1$ components must be in a failed state. Birnbaum's metric is hence

$$I^B(i) = \Pr(\text{Component } i \text{ is critical for the system}) = \prod_{j \neq i} (1 - p_i),$$

which is the same result as obtained by the two first definitions of Birnbaum's metric. □

For more complicated structures, it may be cumbersome to find all the system states that make component i critical, and the third approach may not be an efficient approach.

7.4.4 Computation of Birnbaum's Metric for Structural Importance

Birnbaum's metric of reliability importance $I_\phi^B(i)$ in Definition 7.2 can be determined from Birnbaum's metric of importance as follows: Assume that the reliabilities $p_j(t) = 1/2$ for all $j \neq i$. Again, we suppress the time t. The different realizations of the stochastic vector $(\cdot_i, \boldsymbol{X}) = (X_1, \ldots, X_{i-1}, \cdot, X_{i+1}, \ldots, X_n)$ all have probability $1/2^{n-1}$ because the state variables are assumed to be independent. Then

$$\begin{aligned} I^B(t) &= E[\phi(1_i, \boldsymbol{X}) - \phi(0_i, \boldsymbol{X})] = \sum_{(\cdot i, \boldsymbol{x})} [\phi(1_i, \boldsymbol{x}) - (\phi(0_i, \boldsymbol{x})] \Pr[(\cdot_i, \boldsymbol{X}) = (\cdot_i, \boldsymbol{x})] \\ &= \frac{1}{2^{n-1}} \sum_{(\cdot i, \boldsymbol{x})} [\phi(1_i, \boldsymbol{x}) - \phi(0_i, \boldsymbol{x})] = \frac{\eta_\phi}{2^{n-1}} = I_\phi^B(i), \end{aligned} \tag{7.20}$$

where $\eta_\phi(i)$ is defined in (7.1).

This means that when all the component reliabilities $p_j(t) = 1/2$ for $j \neq i$, then Birnbaum's metric for reliability importance of component i and his metric of structural importance for component i coincide.

$$I_\phi^B(i) = I^B(i)\Big|_{p_j = \frac{1}{2},\, j \neq i} = \frac{\partial h[\boldsymbol{p}]}{\partial p_i}\Big|_{p_j = \frac{1}{2}\; j \neq i}. \tag{7.21}$$

Equation (7.21) is hence an easy way to calculate structural importance.

7.4.5 Variants of Birnbaum's Metric

(1) Assume that component i has a failure rate λ_i. In some situations, we may be interested in studying how much the system reliability will change by making a small change to the failure rate λ_i. The sensitivity of the system reliability with respect to changes in λ_i is obtained by the *chain rule*.

$$\frac{\partial h[\boldsymbol{p}(t)]}{\partial \lambda_i} = \frac{\partial h[\boldsymbol{p}(t)]}{\partial p_i(t)} \frac{\partial p_i(t)}{\partial \lambda_i} = I^B(i \mid t) \frac{\partial p_i(t)}{\partial \lambda_i}. \tag{7.22}$$

(2) Consider a system where component i has reliability $p_i(t)$ that is a function of a parameter θ_i. The parameter θ_i may be the failure rate, the repair rate, or the test frequency of component i. To improve the system reliability, we may want to change the parameter θ_i (by buying a higher-quality component, or changing the maintenance strategy). Assume that we are able to determine the cost of the improvement as a function of θ_i, that is $c_i = c(\theta_i)$, and that this function is strictly increasing or decreasing such that we can find its inverse function. The effect of an extra investment related to component i may now be measured by

$$\frac{\partial h[\boldsymbol{p}(t)]}{\partial c_i} = \frac{\partial h[\boldsymbol{p}(t)]}{\partial \theta_i} \frac{\partial \theta_i}{\partial c_i} = I^B(i \mid t)\, \frac{\partial p_i(t)}{\partial \theta_i} \frac{\partial \theta_i}{\partial c_i}.$$

(3) In a practical reliability study of a complicated system, one of the most time-consuming tasks is to find adequate estimates for the input parameters (e.g. failure rates, repair rates). In some cases, we may start with rather rough estimates, calculate Birnbaum's metric of importance for the various components, or the parameter sensitivities, and then spend most time finding high-quality data for the most important components. Components with a very low value of Birnbaum's metric will have a negligible effect on the system reliability, and extra efforts finding high-quality data for such components may be considered a waste of time.

7.5 Improvement Potential

Consider a system with reliability $h[\boldsymbol{p}(t)]$ at time t. In some cases, it may be of interest to know how much the system reliability increases if component i ($i = 1, 2, \dots, n$) is replaced by a *perfect* component, that is a component with $p_i(t) = 1$. The difference between $h[1_i, \boldsymbol{p}(t)]$ and $h[\boldsymbol{p}(t)]$ is called the *improvement potential* of component i and denoted by $I^{\text{IP}}(i \mid t)$.

Definition 7.6 (Improvement potential)
The improvement potential for component i at time t is

$$I^{\text{IP}}(i \mid t) = h[1_i, \boldsymbol{p}(t)] - h[\boldsymbol{p}(t)] \qquad \text{for } i = 1, 2, \dots, n. \tag{7.23}$$

□

If the time t is given, and we suppress t in the formulas to simplify the notation, the improvement potential is written as

$$I^{\text{IP}}(i) = h(1_i, \boldsymbol{p}) - h(\boldsymbol{p}).$$

When the RBD is established, and all the input parameters (i.e. $\boldsymbol{p}$) are available, the *base case* system reliability $h(\boldsymbol{p})$ is usually calculated. The improvement potential of component i can be obtained by a simple recalculation of the system reliability, but this time with $p_i = 1$, which means that component i is perfect and cannot fail.

With fault tree notation, this may be written as

$$I^{\text{IP}}(i \mid t) = Q_0 - Q_0(E_i^*). \tag{7.24}$$

As for the reliability case above, the base case TOP event probability Q_0 is calculated first. The improvement potential of basic event E_i is obtained by recalculating the TOP event probability $Q_0(E_i^*)$ under the assumption that basic event E_i cannot occur.

If, for example, basic event E_i represents the fault of a safety barrier, the improvement potential tells how much the TOP event probability can be reduced by replacing the current barrier with a barrier that is 100% reliable.

7.5.1 Relation to Birnbaum's Metric

Birnbaum's metric of importance, $I^B(i) = h(1_i, \boldsymbol{p}) - h(\boldsymbol{p})$, is illustrated by the slope of the line in Figure 7.2 and can alternatively be expressed as

$$I^B(i) = \frac{h(1_i, \boldsymbol{p}) - h(\boldsymbol{p})}{1 - p_i} \qquad \text{for } i = 1, 2, \dots, n. \tag{7.25}$$

The improvement potential of component i can therefore be expressed by Birnbaum's metric as

$$I^{\text{IP}}(i) = I^B(i)\ (1 - p_i), \tag{7.26}$$

and Birnbaum's metric can be expressed by the improvement potential as

$$I^B(i) = \frac{I^{\text{IP}}(i)}{1 - p_i}. \tag{7.27}$$

With fault tree notation, we get

$$I^{\text{IP}}(i) = I^B(i)\ q_i, \tag{7.28}$$

and

$$I^B(i) = \frac{I^{\text{IP}}(i)}{q_i} = \frac{Q_0 - Q_0(E_i^*)}{q_i}. \tag{7.29}$$

For very large fault trees, (7.29) is a faster way to find Birnbaum's metric of basic event E_i than (7.18) because only one recalculation of the TOP event probability is required for each basic event.

7.5.2 A Variant of the Improvement Potential

The improvement potential of component i is the difference between the system reliability with a *perfect* component i, and the system reliability with the actual component i. In practice, it is not possible to improve component i to be 100% reliable. Let us assume that it is possible to improve p_i to the new value $p_i^{(n)}$ representing, for example the state-of-the-art for this type of components. We may then calculate the realistic, or *credible improvement potential* (CIP) of component i, defined by

$$I^{\text{CIP}}(i) = h(p_i^{(n)}, \boldsymbol{p}) - h(\boldsymbol{p}), \tag{7.30}$$

where $h(p_i^{(n)}, \boldsymbol{p})$ is the system reliability when component i is replaced with a new component with reliability $p_i^{(n)}$. Because the system reliability $h(\boldsymbol{p})$ is a linear function of p_i and because Birnbaum's metric is the slope of the line in Figure 7.2, we can write (7.30) as

$$I^{\text{CIP}}(i) = I^B(i)\ (p_i^{(n)} - p_i). \tag{7.31}$$

7.6 Criticality Importance

Criticality importance (CR) is a component importance metric that is particularly suitable for prioritizing maintenance tasks. Criticality importance is related to Birnbaum's metric. As a motivation for the definition of criticality importance, recall from Section 7.2 that component i is *critical* for the system if the other components of the system are in such states that the system is functioning if and only if component i is functioning. To say that component i is critical is thus a statement about the other components in the system, and not a statement about component i.

Again, we assume that the time t is given and therefore suppress t in the formulas. Let $C(1_i, \boldsymbol{X})$ be the event that the system at time t is in a state where component i is critical. According to (7.19), the probability of this event is equal to Birnbaum's metric of component i at time t.

$$\Pr[C(1_i, \boldsymbol{X})] = I^B(i). \tag{7.32}$$

Because the components are assumed to be independent, they fail at distinct points in time. We have also assumed that system failure will occur at the same time as one of the component failures. We say that component i *causes* system failure when the system fails when component i fails. For component i to cause system failure at time t, component i must be critical for the system immediately before time t and then fail at time t.

Because the components of the system are independent, event $C(1_i, \boldsymbol{X})$ is independent of the state of component i at time t. The probability that component i is critical for the system just before time t and then fails at time t, is hence

$$\Pr[C(1_i, \boldsymbol{X}) \cap (X_i = 0)] = I^B(i)\ (1 - p_i). \tag{7.33}$$

Assume that we know that the system failed, that is, $\phi(\boldsymbol{X}) = 0$. The conditional probability that component i caused system failure when we know that the system is failed is then

$$\begin{aligned}&\Pr(\text{Component } i \text{ caused system failure} \mid \text{The system is failed})\\ &\quad = \Pr[C(1_i, \boldsymbol{X}) \cap (X_i = 0) \mid \phi(\boldsymbol{X}) = 0].\end{aligned} \tag{7.34}$$

Because event $C(1_i, \boldsymbol{X}) \cap (X_i = 0)$ implies that $\phi(\boldsymbol{X}) = 0$, we can use (7.33) to obtain

$$\frac{\Pr[C(1_i, \boldsymbol{X}) \cap (X_i = 0)]}{\Pr(\phi(\boldsymbol{X}) = 0)} = \frac{I^B(i)\ (1 - p_i)}{1 - h(\boldsymbol{p})}. \tag{7.35}$$

This result is called the criticality importance, and we give the formal definition at time t as follows:

Definition 7.7 (Criticality importance – 1)
Separate definitions are given for (1) component i and (2) for basic event E_i.

(1) The *criticality importance* $I^{\text{CR}}(i \mid t)$ of component i at time t is the probability that component i caused system failure at time t, when we know that the system failed at time t.

$$I^{\text{CR}}(i \mid t) = \frac{I^B(i \mid t)\,[1 - p_i(t)]}{1 - h[\boldsymbol{p}(t)]}. \tag{7.36}$$

(2) The *criticality importance* $I^{\text{CR}}(i \mid t)$ of basic event E_i in a fault tree is the probability that E_i caused the TOP event to occur at time t, when we know that the TOP event occurs at time t.

$$I^{\text{CR}}(i \mid t) = \frac{I^B(i \mid t)\; q_i(t)}{Q_0(t)}. \tag{7.37}$$

□

When the component that has caused system failure is repaired, the system will start functioning again. This is why the criticality importance metric may be used to prioritize maintenance tasks in complicated systems.

Equation (7.37) shows that Birnbaum's metric can be expressed by the criticality importance metric as

$$I^B(i) = \frac{Q_0}{q_i}\, I^{\text{CR}}(i). \tag{7.38}$$

By using (7.28), $I^{\text{IP}}(i) = I^B(i) q_i$, the criticality importance metric may be written as

$$I^{\text{CR}}(i) = \frac{I^B(i)\; q_i}{Q_0} = \frac{I^{\text{IP}}(i)}{Q_0}.$$

The criticality importance metric may therefore, alternatively, be defined as follows:

Definition 7.8 (Criticality importance – 2)
Separate definitions are given for (1) component i and (2) for basic event E_i.

(1) The *criticality importance* $I^{\text{CR}}(i)$ of component i is the probability that component i caused system failure, when we know that the system fails.

$$I^{\text{CR}}(i) = \frac{h(1_i, \boldsymbol{p}) - h(\boldsymbol{p})}{1 - h(\boldsymbol{p})}. \tag{7.39}$$

(2) The *criticality importance* $I^{\text{CR}}(i)$ of basic event E_i in a fault tree is the probability that E_i caused the TOP event to occur, when we know that the TOP event occurs.

$$I^{\text{CR}}(i) = \frac{Q_0 - Q_0(E_i^*)}{Q_0}. \tag{7.40}$$

□

Definition 7.8 can be applied to fault trees with dependent basic events. If we are able to calculate the TOP event probability, we are also able to calculate the criticality importance. The second benefit of this definition is that it is easier to calculate because only one extra recalculation of the TOP event probability is required for each basic event.

By combining (7.6) and (7.37), the criticality importance $I^{\text{CR}}(i \mid t)$ of component i may be written as

$$I^{\text{CR}}(i) = \frac{\partial Q_0}{\partial q_i}\,\frac{q_i}{Q_0} = \frac{\partial Q_0 / Q_0}{\partial q_i / q_i}.$$

This may also be written as

$$\frac{\partial Q_0}{Q_0} = I^{\text{CR}}(i)\,\frac{\partial q_i}{q_i}. \tag{7.41}$$

Equation (7.38) helps to answer questions such as "If we make a small improvement (e.g. 5%) to the basic event probability $q_i(t)$, what will the (relative) effect on the TOP event probability $Q_0(t)$ be?"

7.7 Fussell–Vesely's Metric

Fussell and Vesely suggested the following metric for the importance of component i (see Fussell, 1975):

Definition 7.9 (Fussell–Vesely's metric – 1)
Fussell–Vesely's metric of importance, $I^{\text{FV}}(i \mid t)$ is the probability that at least one minimal cut set that contains component i, is failed at time t, given that the system is failed at time t. □

An alternative definition is provided later in this section. We say that a minimal cut set is failed when all the basic events in the minimal cut set occur – or, more formally, when the associated minimal cut parallel structure is failed.

Fussell–Vesely's metric takes into account the fact that a component may contribute to a system failure without being critical. The component contributes to a system failure when a minimal cut set, containing the component, is failed.

7.7.1 Derivation of Formulas for Fussell–Vesely's Metric

Consider a fault tree with n distinct basic events and k minimal cut sets $K_1, K_2, \dots, K_k$. Let F_j denote that minimal cut set j is failed, for $j = 1, 2, \dots, k$. Because the basic events are independent, the probability of F_j is

$$\check{Q}_j = \Pr(F_j) = \prod_{\ell \in K_j} q_\ell. \tag{7.42}$$

For any coherent fault tree, basic event E_i is a member of at least one minimal cut set. The number n_i of minimal cut sets where E_i is a member may range from 1 up to n. Let K_j^i be such a minimal cut set and let F_j^i denote that minimal cut set K_j^i is failed. The probability of this event is

$$\check{Q}_j^i = \Pr(F_j^i) = \prod_{\ell \in K_j^i} q_\ell. \tag{7.43}$$

For the TOP event to occur, at least one of the minimal cut sets must fail. The TOP event (i.e. system fault) can therefore be written as TOP $= \bigcup_{j=1}^k F_j$.

With this notation, Fussell–Vesely's metric may be written as the conditional probability

$$I^{\text{FV}}(i) = \Pr\left(\bigcup_{\nu=1}^{n_i} F_\nu^i \mid \bigcup_{j=1}^{k} F_j\right) = \frac{\Pr\left(\bigcup_{\nu=1}^{n_i} F_\nu^i\right)}{\Pr\left(\bigcup_{j=1}^{k} F_j\right)} = \frac{\Pr\left(\bigcup_{\nu=1}^{n_i} F_\nu^i\right)}{Q_0}, \tag{7.44}$$

because a failed minimal cut set will always lead to the TOP event.

We can now use the *upper bound approximation* (6.93) to determine both the nominator and the denominator in (7.44).

$$\Pr\left(\bigcup_{j=1}^{n_i} F_j^i\right) \lesssim 1 - \prod_{j=1}^{n_i}(1 - \check{Q}_j^i),$$

where we have replaced the counting variable ν with j.

Fussell–Vesely's metric for the importance of basic event E_i can therefore be calculated as

$$I^{\text{FV}}(i) \approx \frac{1 - \prod_{j=1}^{n_i}(1 - \check{Q}_j^i)}{Q_0}. \tag{7.45}$$

A slightly more crude approximation is $1 - \prod_{j=1}^{n_i}(1 - \check{Q}_j^i) \lesssim \sum_{j=1}^{n_i} \check{Q}_j^i$. With this approximation, Fussell–Vesely's metric may be calculated as

$$I^{\text{FV}}(i) \approx \frac{\sum_{j=1}^{n_i} \check{Q}_j^i}{Q_0}. \tag{7.46}$$

As shown in Chapter 4, any coherent fault tree can be drawn with an OR-gate under the TOP event, and where the minimal cut sets are inputs to this OR-gate. Assume that we delete all the minimal cut sets that do not contain basic event E_i and remain with the n_i minimal cut sets where E_i is a member. Further, let Q_0^i be the TOP event probability of this modified fault tree. Q_0^i may then be approximated by

$$Q_0^i \lesssim 1 - \prod_{j=1}^{n_i}(1 - \check{Q}_j^i) \lesssim \sum_{j=1}^{n_i} \check{Q}_j^i.$$

Here, Q_0^i can be interpreted as the contribution to the TOP event probability from those minimal cut sets, where basic event E_i is a member, and we may express Fussell–Vesely's metric as

$$I^{\text{FV}}(i) \approx \frac{Q_0^i}{Q_0}, \tag{7.47}$$

that is, the relative contribution to the TOP event probability from those minimal cut sets where basic event E_i is a member.

For complicated systems, Fussell–Vesely's metric is considerably faster and easier to calculate (even by hand) than Birnbaum's metric and the criticality importance metric. When Fussell–Vesely's metric is to be calculated by hand, formula (7.46) is normally used. The formula is simple to use and at the same time gives a good approximation when the basic event probabilities are small.

Example 7.11 (Bridge structure)

Consider the bridge structure in Figure 6.18. As shown in Example 4.5, the minimal cut sets of this structure are $K_1 = \{1, 2\}, K_2 = \{4, 5\}, K_3 = \{1, 3, 5\}$ and $K_4 = \{2, 3, 4\}$. Assume the following component unreliabilities:

Comp. i	p_i	$q_i = 1 - p_i$
1	0.99	0.01
2	0.98	0.02
3	0.95	0.05
4	0.97	0.03
5	0.98	0.02

The minimal cuts fail with the following probabilities

$\check{Q}_1 =$	$q_1q_2 = 0.01 \cdot 0.02$	$= 2 \times 10^{-4}$
$\check{Q}_2 =$	$q_4q_5 = 0.03 \cdot 0.02$	$= 6 \times 10^{-4}$
$\check{Q}_3 =$	$q_1q_3q_5 = 0.01 \cdot 0.05 \cdot 0.02$	$= 1 \times 10^{-5}$
$\check{Q}_4 =$	$q_2q_3q_4 = 0.02 \cdot 0.05 \cdot 0.03$	$= 3 \times 10^{-5}$

If we draw a corresponding fault tree with TOP event "System failed," the TOP event probability is

$$Q_0 \approx 1 - \prod_{j=1}^{4}(1 - \check{Q}_j) = 8.4 \times 10^{-4}.$$

For this example, the cruder approximations $Q_0 \approx \sum_{j=1}^{4} \check{Q}_j$ is very accurate. If not rounded off, the difference between the two approximations is approximately 1.52×10^{-7}.

To find Fussell–Vesely's metric of, for example basic event E_2 (i.e. component 2), observe that component 2 is a member of the two minimal cut sets K_1 and K_4. The contribution to the TOP event probability from these two cut sets is

$$Q_0^2 \approx 1 - (1{-}\check{Q}_1)(1{-}\check{Q}_4) = 2.3 \times 10^{-4}.$$

The FV importance of component 2 is therefore

$$I^{\text{FV}}(2) \approx \frac{1 - (1{-}\check{Q}_1)(1{-}\check{Q}_4)}{Q_0} \approx 0.274. \tag{7.48}$$

Fussell–Vesely's metric for the other basic events (components) can be determined in the same way to obtain:

Comp. *i*	$I^{\text{FV}}(i)$
1	0.250
2	0.274
3	0.048
4	0.750
5	0.726

Component 3 has a much lower Fussell–Vesely importance than the other four components. This is to be expected because component 3 is member of minimal cut sets of order 3, whereas the other components are also members of a minimal cut set of order 2. Generally, we find that components (and basic events) that are members of minimal cut sets of the lowest order are the most important. □

7.7.2 Relationship to Other Metrics for Importance

In Eq. (7.46), $\check{Q}_j^i$ is the probability that minimal cut set j, which contains component i, is failed. From (7.42), we have that $\check{Q}_j^i = \prod_{\ell \in K_j^i} q_\ell$, where we can put $q_i(t)$ outside the product and get

$$\check{Q}_j^i = q_i \left(\prod_{\ell \in K_j^i, \ell \neq i} q_\ell \right) = q_i \, \check{Q}_j^{i-}. \tag{7.49}$$

Here, $\check{Q}_j^{i-}$ is the probability that minimal cut set j – that contains basic event E_i (component i), but where basic event E_i is removed – is failed. We may now rewrite (7.46) and get

$$I^{\mathrm{FV}}(i) \approx \frac{q_i}{Q_0} \sum_{j=1}^{n_i} \check{Q}_j^{i-}. \tag{7.50}$$

The TOP event probability $Q_0(t)$ may, according to (6.95) be approximated by

$$Q_0 \approx \sum_{j=1}^{k} \check{Q}_j. \tag{7.51}$$

Equation (7.47) may be used to find an approximation to Birnbaum's metric for basic event E_i. We therefore have to take the partial derivative of Q_0 with respect to q_i. The partial derivative of $\check{Q}_j$ is zero for all minimal cut sets, where E_i is not a member, and the partial derivative of a $\check{Q}_j^i$ where i is a member is from (7.45)

$$I^B(i) = \frac{\partial Q_0}{\partial q_i} \approx \sum_{j=1}^{n_i} \check{Q}_j^{i-}.$$

The criticality importance metric is then given by

$$I^{\mathrm{CR}}(i) = \frac{q_i}{Q_0}\, I^B(i) \approx \frac{q_i}{Q_0(t)} \sum_{j=1}^{n_i} \check{Q}_j^{i-}. \tag{7.52}$$

By comparing with (7.50), we see that

$$I^{\mathrm{FV}}(i) \approx I^{\mathrm{CR}}(i), \tag{7.53}$$

for systems where the approximation (7.51) is adequate. This means that the following, alternative definition of Fussell–Vesely's metric may be used.

Definition 7.10 (Fussell–Vesely's metric – 2)
Fussell–Vesely's metric of importance, $I^{\mathrm{FV}}(i)$ is approximately given by

$$I^{\mathrm{FV}}(i) \approx \frac{Q_0 - Q_0(E_i^*)}{Q_0}.$$

□

This second definition can be used also for fault trees with dependent basic events.

Birnbaum's metric may be (approximately) expressed by Fussel–Vesely's metric as

$$I^B(i) \approx \frac{q_i}{Q_0} I^{\mathrm{FV}}(i). \tag{7.54}$$

This way of determining Birnbaum's metric is listed in several guidelines.

Remark 7.3 Consider a system with minimal cut sets $K_1, K_2, \dots, K_k$. A necessary criterion for component i to be critical for the system is that all the components, except for component i, in at least one minimal cut set containing component i are in a failed state. This is not a sufficient criterion for component i to be critical because we have to require the remaining cut sets be functioning. This fact highlights the similarity and the difference between the definitions of criticality importance $I^{\text{CR}}(i)$, and Fussell–Vesely's metric $I^{\text{FV}}(i)$. We realize that we always have that

$$I^{\text{CR}}(i) \lesssim I^{\text{FV}}(i). \tag{7.55}$$

□

Example 7.12 (Example 7.11 cont.)

Reconsider the bridge structure in Example 7.11. The TOP event probability may be expressed as

$$\begin{aligned} Q_0 = {} & q_1q_2 + q_4q_5 + q_1q_3q_5 + q_2q_3q_4 - q_1q_2q_4q_5 - q_1q_2q_3q_5 \\ & - q_1q_2q_3q_4 - q_1q_3q_4q_5 - q_2q_3q_4q_5 + 2q_1q_2q_3q_4q_5. \end{aligned} \tag{7.56}$$

With the same input data as in Example 7.12, $Q_0 = 8.38 \times 10^{-4}$, that is slightly smaller than obtained by the upper bound approximation.

Birnbaum's metric is obtained by taking the derivative, for example,

$$I^B(1) = \frac{\partial Q_0}{\partial q_1} = q_2 + q_3q_5 - q_2q_4q_5 - q_2q_3q_5 - q_2q_3q_4 - q_3q_4q_5 + 2q_2q_3q_4q_5.$$

Similar for the other basic events. The criticality importance of basic event E_i is calculated as

$$I^{\text{CR}}(i) = \frac{q_i}{Q_0}\, I^B(i).$$

By including the Fussell–Vesely metric from Example 7.11 the following results are obtained:

Component *i*	$I^B(i)$	$I^{\text{CR}}(i)$	$I^{\text{VF}}(i)$
1	0.020	0.249	0.250
2	0.011	0.273	0.274
3	7.72×10^{-4}	0.046	0.048
4	0.020	0.750	0.750
5	0.030	0.726	0.726

Observe that Fussell–Vesely's metric is a good approximation to criticality importance for this example. □

7.8 Differential Importance Metric

The *differential importance metric* (DIM) is proposed by Borgonovo and Apostolakis (2001). The DIM of basic event E_i is denoted DIM(i). The idea of DIM(i) is to compare a small change (Δq_i) of the basic event probability q_i with the change of the TOP event probability Q_0. DIM(i) is defined as follows:

Definition 7.11 (Differential importance metric)
The DIM is given by

$$\text{DIM}(i) = \frac{\frac{\partial Q_0}{\partial q_i}\,\Delta q_i}{\sum_{j=1}^{n} \frac{\partial Q_0}{\partial q_j}\,\Delta q_j}. \tag{7.57}$$

□

The numerical value of DIM(i) depends on how the changes Δq_j, for $j = 1, 2, \dots, n$ are selected. Borgonovo and Apostolakis (2001) propose two different options:

(1) $\Delta q_j = \Delta q_k$ for all j, k. This is the simplest possible option and may be acceptable when all the q_js are of the same order of magnitude.

(2) $\dfrac{\Delta q_j}{q_j} = \dfrac{\Delta q_k}{q_k}$ for all j, k. In this option, all the basic event probabilities are changed with the same percentage. This option is more acceptable when the q_js are very different, for example when one basic event represent an operator error with probability 0.10 and another basic event is the fault of a safety item with basic event probability 10^{-5}.

7.8.1 Option 1

For option 1, Δq_j has the same value for all $j = 1, 2, \dots, n$ and can therefore be cancelled from the expression, such that $\text{DIM}_1(i)$ becomes

$$\text{DIM}_1(i) = \frac{\frac{\partial Q_0}{\partial q_i}\,\Delta q_i}{\sum_{j=1}^{n} \frac{\partial Q_0}{\partial q_j}\,\Delta q_j} = \frac{\frac{\partial Q_0}{\partial q_i}}{\sum_{j=1}^{n} \frac{\partial Q_0}{\partial q_j}}.$$

Because Birnbaum's metric $I^B(j)$ was defined as

$$I^B(j) = \frac{\partial Q_0}{\partial q_j} \qquad \text{for } j = 1, 2, \dots, n.$$

$\text{DIM}_1(j)$ can be expressed as a function of Birnbaum's metric for the n basic events.

$$\text{DIM}_1(i) = \frac{I^B(i)}{\sum_{j=1}^{n} I^B(j)}. \tag{7.58}$$

DIM_1 will therefore give the same importance ranking of the basic event as Birnbaum's metric, but the numerical value of DIM_1 will be relative to the sum of Birnbaum's metric for all the n basic events and therefore add up to 1.

$$\sum_{i=1}^{n} \text{DIM}_1(i) = 1.$$

7.8.2 Option 2

For option 2, $\Delta q_j/q_j$ has the same value for all $j = 1, 2, \dots, n$, and we can therefore cancel all factors $\Delta q_j/q_j$. If we also divide by Q_0, $\text{DIM}_2(i)$ becomes

$$\text{DIM}_2(i) = \frac{\frac{\partial Q_0}{\partial q_i}\,\Delta q_i}{\sum_{j=1}^{n} \frac{\partial Q_0}{\partial q_j}\,\Delta q_j} = \frac{\frac{\partial Q_0}{\partial q_i}\,q_i}{\sum_{j=1}^{n} \frac{\partial Q_0}{\partial q_j}\,q_j} = \frac{\frac{\partial Q_0}{\partial q_i}\,\frac{q_i}{Q_0}}{\sum_{j=1}^{n} \frac{\partial Q_0}{\partial q_j}\,\frac{q_j}{Q_0}}.$$

Because the criticality importance metric $I^{\text{CR}}(i)$ can be expressed as

$$I^{\text{CR}}(j) = \frac{\partial Q_0}{\partial q_j}\,\frac{q_j}{Q_0}.$$

$\text{DIM}_2(j)$ can be expressed as a function of the criticality importance of the various basic events.

$$\text{DIM}_2(j) = \frac{I^{\text{CR}}(i)}{\sum_{j=1}^{n} I^{\text{CR}}(j)}. \tag{7.59}$$

DIM_2 gives the same importance ranking as the criticality importance metric, but the numerical value of DIM_2 will be relative to the sum of the criticality importance for all the n basic events and therefore add up to 1.

$$\sum_{i=1}^{n} \text{DIM}_2(i) = 1.$$

Because $I^{\text{CR}}(i) \approx I^{\text{FV}}(i)$, we also have that

$$\text{DIM}_2(j) \approx \frac{I^{\text{FV}}(i)}{\sum_{j=1}^{n} I^{\text{FV}}(j)}, \tag{7.60}$$

which is easier to calculate than Eq. (7.59).

A main advantage of DIM is that it is *additive*, such that DIM of a group of several basic events can be determined by adding the individual DIMs. This means that we can find the DIM of a larger module (e.g. equipment, subsystem) by summing the DIM of the components (or basic events) that constitute the module.

Example 7.13 (Simple structure)

Consider the structure of three independent components that is shown by the RBD in Figure 7.3. Component 1 has reliability $p_1 = 0.99$, component 2 has reliability

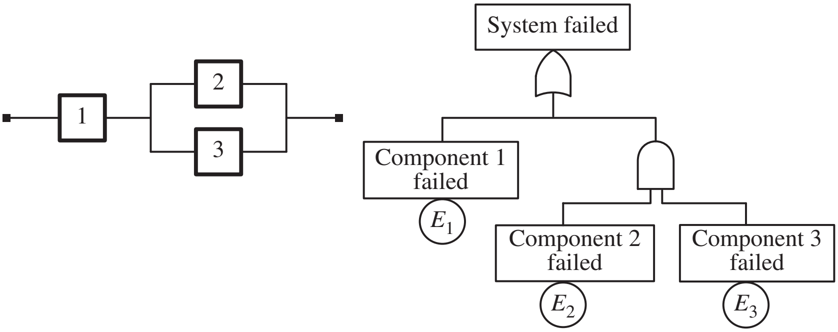


Figure 7.3 Structure with three components: RBD and associated fault tree representation.

$p_2 = 0.95$, and component 3 has reliability $p_3 = 0.93$. The associated fault tree, shown in Figure 7.3, has three basic events E_1, E_2, and E_3 with basic event probabilities $q_1 = 0.01$, $q_2 = 0.05$, and $q_3 = 0.07$.

The TOP event probability of the fault tree is

$$Q_0 = q_1 + q_2q_3 - q_1q_2q_3 = 0.0135.$$

Birnbaum's metric for the two basic events (and components) are

$$I^B(1) = \frac{\partial Q_0}{\partial q_1} = 1 - q_2q_3 = 0.9965$$

$$I^B(2) = \frac{\partial Q_0}{\partial q_2} = q_3 - q_1q_3 = 0.0693$$

$$I^B(3) = \frac{\partial Q_0}{\partial q_3} = q_2 - q_1q_2 = 0.0495.$$

DIM_1 for the two basic events (and components) are

$$\text{DIM}_1(1) = \frac{I^B(1)}{I^B(1) + I^B(2) + I^B(3)} = 0.8935 = 89.35\%$$

$$\text{DIM}_1(2) = \frac{I^B(2)}{I^B(1) + I^B(2) + I^B(3)} = 0.0621 = 6.21\%$$

$$\text{DIM}_1(3) = \frac{I^B(3)}{I^B(1) + I^B(2) + I^B(3)} = 0.0444 = 4.44\%,$$

such that $\text{DIM}_1(1) + \text{DIM}_1(2) + \text{DIM}_1(3) = 1$.

The criticality importance of the basic events (and components) are

$$I^{\text{CR}}(1) = \frac{\partial Q_0}{\partial q_1}\frac{q_1}{Q_0} = 0.7401$$

$$I^{\text{CR}}(2) = \frac{\partial Q_0}{\partial q_2}\frac{q_2}{Q_0} = 0.2573$$

$$I^{\text{CR}}(3) = \frac{\partial Q_0}{\partial q_3}\frac{q_3}{Q_0} = 0.2573.$$

DIM_2 for the basic events (and components) are

$$\text{DIM}_2(1) = \frac{I^{\text{CR}}(1)}{I^{\text{CR}}(1) + I^{\text{CR}}(2) + I^{\text{CR}}(3)} = 0.5898 = 58.98\%$$

$$\text{DIM}_2(2) = \frac{I^{\text{CR}}(2)}{I^{\text{CR}}(1) + I^{\text{CR}}(2) + I^{\text{CR}}(3)} = 0.2051 = 20.51\%$$

$$\text{DIM}_2(3) = \frac{I^{\text{CR}}(3)}{I^{\text{CR}}(1) + I^{\text{CR}}(2) + I^{\text{CR}}(3)} = 0.2051 = 20.51\%,$$

such that $\text{DIM}_2(1) + \text{DIM}_2(2) + \text{DIM}_2(3) = 1$. □

The reader can refer to Do et al. (2008, 2010) for further work on the DIM in the context of dynamic systems including inter-component, functional dependencies, or more generally, systems described by Markov models at steady state.

7.9 Importance Metrics for Safety Features

In this section, we introduce two importance metrics for a *safety feature* in a system. Safety feature *i* is assumed to be represented as an event E_i in a fault tree. The event E_i may be a basic event or an intermediate event. In the latter case, E_i may be the TOP event of a subfault tree and may sometimes represent a complicated safety system. The interpretation of the term "safety feature" is wide. It may be a technical item or a human action, in fact, every fault tree event that represent a protective feature.

The importance metrics are

- Risk achievement worth (RAW)
- Risk reduction worth (RRW)

The two metrics were introduced for the nuclear power industry (e.g. see NUREG/CR-3385, 1986) and are still most used in nuclear applications. The main purpose of these two importance metrics is to support decision-making related to the following questions:

- What is the risk reduction obtained by safety feature *i*?
- How much can the risk be reduced by installing safety feature *i*?
- Which safety feature (among several candidates) should be installed – and where?
- Is it sufficiently safe to remove or by-pass safety feature *i* when the system is in operation?

The metrics are in principle time-dependent, but we suppress the reference to a time t to simplify the notation. We only present the metrics with fault tree terminology.

The term *risk* is usually defined as a function of the consequence of an accident scenario and the probability or frequency of the accident scenario. In this case, the consequence is disregarded (or assumed to be the same in all cases), and the risk is measured as the probability (or frequency) of the TOP event. As used in this section, risk reduction means reducing the TOP event probability (or frequency).

7.9.1 Risk Achievement Worth

The importance metric *risk achievement worth* (RAW) is defined as (e.g. see Cheok et al. 1998).

Definition 7.12 (Risk achievement worth)
The RAW of basic event i is defined as the ratio

$$I^{\mathrm{RAW}}(i) = \frac{Q_0(E_i)}{Q_0} \qquad \text{for } i = 1, 2, \dots, n, \tag{7.61}$$

□

where $Q_0(E_i)$ is the TOP event probability when we know that basic event E_i occurs (with probability 1). This may, for example, represent that we know that a safety feature i has been taken out of service or that it fails.

If we assume that a safety feature always have a positive effect on the system's safety, we must have $Q_0(E_i) \geq Q_0$. Consequently, $I^{\mathrm{RAW}}(i) \geq 1$ for all coherent fault trees.[2]

We introduce the term *risk achievement* for basic event E_i as

$$\mathrm{RA}(i) = Q_0(E_i) - Q_0. \tag{7.62}$$

RA(i) tells how much the risk (i.e. the TOP event probability) can be reduced by installing safety feature i (with its actual reliability).

[2] In the risk literature, the RAW is often denoted RAW or RAW(i) instead of $I^{\mathrm{RAW}}(i)$.

Equation (7.61) can be rewritten as

$$I^{\mathrm{RAW}}(i) - 1 = \frac{Q_0(E_i) - Q_0}{Q_0} = \frac{\mathrm{RA}(i)}{Q_0},$$

such that

$$\mathrm{RA}(i) = [I^{\mathrm{RAW}}(i) - 1]Q_0. \tag{7.63}$$

Example 7.14 (A numerical example)
Assume that the RAW importance of safety feature i is found to be $I^{\mathrm{RAW}}(i) = 1.25$. This means that the risk achievement of i is from (7.58)

$$\mathrm{RA}(i) = 0.25\ Q_0.$$

This means that by installing the safety feature i the TOP event probability is reduced by 25%. □

Example 7.15 (Barrier against demands)
A safety feature i is considered to be installed in a process system as barrier against a specific type of demands that occur at random with frequency ν_0. Further, assume that an accident occurs if the TOP event is present when a demand occurs. The accident frequency is therefore $\nu_{\mathrm{acc}} = \nu_0\ Q_0$ (see Section 9.2 for a detailed explanation).

Let $Q_0(E_i)$ be the TOP event probability without the safety feature i, and Q_0 the TOP event probability with the safety feature (with its actual reliability). The risk achievement of installing safety feature i is $Q_0(E_i) - Q_0$. The accident frequency is now

$$\begin{aligned} &\text{With the safety feature } i: && \nu_{\mathrm{acc}}^{+} = \nu_0\ Q_0 \\ &\text{Without the safety feature } i: && \nu_{\mathrm{acc}}^{-} = \nu_0\ Q(E_i). \end{aligned}$$

This means that

$$\nu_{\mathrm{acc}}^{-} = \frac{Q_0(E_i)}{Q_0}\ \nu_{\mathrm{acc}}^{+} = I^{\mathrm{RAW}}(i)\ \nu_{\mathrm{acc}}^{+}. \tag{7.64}$$

If, for example, the RAW importance of safety feature i is found to be $I^{\mathrm{RAW}}(i) = 1.25$, the accident frequency without the safety feature in place (e.g. that it is removed for maintenance) will increase to $\nu_{\mathrm{acc}}^{-} = 1.25\nu_{\mathrm{acc}}^{+}$, or 25% higher that it would be with the safety feature in place.

In nuclear applications, one is mainly concerned with the *core damage frequency* (CDF). The CDF is the accident frequency ν_{acc} for core damage accidents.

If a safety feature i is disconnected from the main safety system, the CDF becomes

$$\mathrm{CDF}_i = I^{\mathrm{RAW}}(i)\ \mathrm{CDF}_0,$$

where CDF_0 is the base core damage frequency, and CDF_i is the core damage frequency when safety feature i is not present. □

7.9.2 Risk Reduction Worth

The importance metric *risk reduction worth* (RRW) is defined as follows:

Definition 7.13 (Risk reduction worth)
The RRW of basic event E_i is defined as the ratio

$$I^{\text{RRW}}(i) = \frac{Q_0}{Q_0(E_i^*)} \quad \text{for} \;\; i = 1, 2, \ldots, n. \tag{7.65}$$

□

If we assume that the safety feature has a positive effect on the system's safety, $Q_0 \geq Q_0(E_i^*)$ and consequently that $I^{\text{RRW}}(i) \geq 1$. Recall that $Q_0(E_i^*)$ is the conditional TOP event probability when safety feature i is available and 100% reliable (i.e. will always function as intended). Also recall that Q_0 is the TOP event probability when safety feature i is installed with its actual reliability. Two small examples may help explain the conditional probability:

- The current safety feature is replaced with a safety feature that can never fail.
- The event E_i is related to an operator error or some external events that are removed from the system by a system modification that avoids the operator intervention or protects the system from external stresses.

We introduce the term *risk reduction* for basic event E_i as

$$\text{RR}(i) = Q_0 - Q_0(E_i^*). \tag{7.66}$$

The risk reduction, RR(i), tells how much the TOP event probability may be reduced by replacing the current safety feature i by a perfect safety function with the same functionality, or by *designing out* the problem that the safety feature is protecting against.

Equation (7.65) can be rewritten as

$$I^{\text{RRW}}(i) - 1 = \frac{Q_0 - Q_0(E_i^*)}{Q_0(E_i^*)} = \frac{\text{RR}(i)}{Q_0(E_i^*)},$$

such that

$$\text{RR}(i) = (I^{\text{RRW}}(i) - 1)Q_0(E_i^*). \tag{7.67}$$

Example 7.16 (A numerical example)
Assume that the RRW importance of safety feature i is found to be $I^{\text{RRW}}(i) = 1.25$. This means that the risk reduction of i is

$$\text{RR}(i) = 0.25\ Q_0(E_i^*).$$

This means that by installing safety feature i (with its actual reliability) provides a risk reduction that is 25% of the TOP even probability obtained with a 100% reliable safety feature (with the same functionality). □

7.9.3 Relationship with the Improvement Potential

Assume that the systems are considered at a given time t, but suppress t in the formulas to simplify the notation.

Recall that the improvement potential was defined as

$$I^{\mathrm{IP}}(i) = Q_0 - Q_0(E_i^*), \tag{7.68}$$

which is the same definition as for RRW. Mathematically, the IP and the RRW are identical

$$I^{\mathrm{IP}}(i) = I^{\mathrm{RRW}}(i), \tag{7.69}$$

but the two metrics are used for different purposes. Whereas $I^{\mathrm{IP}}(i)$ is mainly used in relation to avoiding potential component failures in the design phase of a system, the $I^{\mathrm{RRW}}(i)$ is used as support for decision-making related to installation or removal of safety features.

Example 7.17 Reconsider the process safety system in Example 7.15 with reliability $h(\boldsymbol{p})$. Let us assume that we contemplate improving component i and would like to know the maximum potential improvement, by replacing component i with a *perfect* component with reliability $p_i = 1$. The (conditional) system reliability will then be $1 - h(1_i, \boldsymbol{p})$, which we can use (7.65) to express as

$$1 - h(1_i, \boldsymbol{p}) = \frac{1 - h(\boldsymbol{p})}{I^{\mathrm{RRW}}(i)}.$$

If we, as an example, find that $I^{\mathrm{RRW}}(i) = 2$, then the system unreliability we would obtain by replacing component i with a perfect component would be 50% of the initial unreliability $1 - h(\boldsymbol{p})$. □

Remark 7.4 We observe from (7.52) that the criticality importance $I^{\mathrm{CR}}(i \mid t)$ is close to a linear function of $q_i(t)$, at least for systems with a high level of redundancy. This is because Birnbaum's metric is not a function of $q_i(t)$ and because $q_i(t)$ will have a rather low influence on $Q_0(t)$ in highly redundant systems. The linearity is, however, not adequate for very simple systems with two components. □

7.10 Barlow–Proschan's Metric

Barlow and Proschan (1975) observe that Birnbaum's metric gives the importance at fixed points of time, leaving for the analyst to determine which points are important. To compensate for this shortcoming they proposed a new metric, which is now known as *Barlow–Proschan's metric* for the importance $I^{\mathrm{BP}}(i)$ of component i. Before defining Barlow and Proschan's metric, we consider some intermediate results.

Let $S^*(t, t + dt)$ be the event that a system failure occurs in $(t, t + dt)$ and let $B_i^*(t, t + dt)$ be the event that a failure of component i occurs in $(t, t + dt)$. If component i is critical for the system at time t, the two events will occur at the same time, and component i is said to *cause* the system failure.

The conditional probability that the system failure is caused by component i, given that a system failure occurs in $(t, t + dt)$ is

$$\Pr[B_i^*(t, t + dt) \mid S^*(t, t + dt)] = \frac{\Pr[B_i^*(t, t + \Delta t) \cap S^*(t, t + dt)]}{\Pr[S^*(t, t + dt)]}. \tag{7.70}$$

According to the third definition of Birnbaum's importance metric, $I^B(i \mid t)$ is the probability that component i is *critical* at time t. When we know that the system has failed in $(t, t + dt)$, that is, the event $S^*(t, t + dt)$ has occurred, the simultaneous occurrence of $B_i^*(t, t + dt)$ is the same as the event "component i caused system failure at time t."

When component i is *nonrepairable*, the probability that component i causes system failure at time t must be $I^B(i \mid t)f_i(t)dt$, where $f_i(t)$ is the probability density function for the time-to-failure of component i.

Because any system failure must be caused by (i.e. coincide with) failure of one of the components, the probability of $S^*(t, t + dt)$ can be written as

$$\Pr[S^*(t, t + dt)] = \sum_{i=1}^{n} I^B(i \mid t)f(t)\, dt. \tag{7.71}$$

The conditional probability (7.70) can hence be written as

$$\frac{I^B(i \mid t)f(t)\, dt}{\sum_{i=1}^{n} I^B(i \mid t)f(t)\, dt}. \tag{7.72}$$

The conditional probability that the system failure is caused by component i in the time interval $(0, t_0)$ is

$$\frac{\int_0^{t_0} I^B(i \mid t)f(t)\, dt}{\sum_{i=1}^{n} \int_0^{t_0} I^B(i \mid t)f(t)\, dt}. \tag{7.73}$$

Letting $t_0 \to \infty$, the denominator tends to one because any system must fail sooner or later. We are now ready to define Barlow and Proschan's metric:

Definition 7.14 (Barlow–Proschan's metric for a nonrepairable component)
Barlow–Proschan's metric of importance for a nonrepairable component i is

$$I^{\mathrm{BP}}(i) = \int_0^\infty I^B(i \mid t) f_i(t)\, dt \qquad \text{for } i = 1, 2, \dots, n. \tag{7.74}$$

□

Similarly, when component i is repairable, the probability that component i fails in $(t, t + dt)$ is $w_i(t)dt$, where $w_i(t)$ is the ROCOF of component i at time t. Barlow and Proschan's metric of importance of the repairable component i is hence:

Definition 7.15 (Barlow–Proschan's metric for a repairable component)
Barlow–Proschan's metric of importance for a repairable component i is

$$I^{\mathrm{BP}}(i) = \int_0^\infty I^B(i \mid t)\ w_i(t)\ dt. \tag{7.75}$$

□

Observe that it is obvious from (7.72) that

$$\sum_{i=1}^n I^{\mathrm{BP}}(i) = 1.$$

This means that $I^{\mathrm{BP}}(i)$ is the percentage of system failures that are caused by component i (i.e. among the failures that occur).

Example 7.18 (Series structure)
Reconsider the series structure of n independent and nonrepairable components in Example 7.3. Assume that all components have constant failure rates λ_i, for $i = 1, 2, \dots, n$. Birnbaum's metric of component i is from Example 7.9

$$I^B(i \mid t) = \prod_{j \neq i} e^{-\lambda_j t}.$$

The probability density for the time-to-failure of component i is $f_i(t) = \lambda_i e^{-\lambda_i t}$ and Barlow–Proschan's metric of component i is

$$\begin{aligned} I^{\mathrm{BP}}(i) &= \int_0^\infty I^B(i \mid t) f_i(t)\, dt = \int_0^\infty \prod_{j \neq i} e^{-\lambda_j t} \lambda_i e^{-\lambda_i t}\, dt = \lambda_i \int_0^\infty \prod_{j=1}^n e^{-\lambda_j t}\, dt \\ &= \lambda_i \int_0^\infty e^{-\sum_{j=1}^n \lambda_j t}\, dt = \frac{\lambda_i}{\sum_{j=1}^n \lambda_j}. \end{aligned}$$

This means that Barlow–Proschan's metric is the percentage of system failures that are caused by component i for the series structure. Observe that when all the components have the same failure rate λ, Barlow–Proschan's metric of each component is $1/n$. □

Example 7.19 (Parallel structure)
Reconsider the parallel structure of n independent and nonrepairable components in Example 7.4. Assume that all components have constant failure rates λ_i, for $i = 1, 2, \dots, n$. Birnbaum's metric of component i is from Example 7.10.

$$I^B(i \mid t) = \prod_{j \neq i} (1 - e^{-\lambda_j t}).$$

The probability density for the time-to-failure of component i is $f_i(t) = \lambda_i e^{-\lambda_i t}$ and Barlow–Proschan's metric of component i is

$$I^{\text{BP}}(i) = \int_0^\infty I^B(i \mid t) f_i(t)\, dt = \int_0^\infty \prod_{j \neq i} (1 - e^{-\lambda_j t}) \lambda_i e^{-\lambda_i t}\, dt.$$

To solve this integral is time-consuming, and we do not come up with any nice and closed formula for $I^{\text{BP}}(i)$. □

The problem of finding a nice and closed formula for the parallel structure in Example 7.19 also applies to most structures that are not a purely series structure and this makes Barlow–Proschan's metric of limited interest. The computation of Barlow–Proschan's metric is discussed by Eryilmaz (2016) for systems of identical components.

7.11 Problems

7.1 Show that a 2oo3:G structure of independent components with component reliabilities $p_1 \geq p_2 \geq p_3$ fulfills:
(a) if $p_1 \geq 0.5$, $I^B(1) \geq I^B(2) \geq I^B(3)$
(b) if $p_1 \leq 0.5$, $I^B(1) \leq I^B(2) \leq I^B(3)$

7.2 Let $p_S(t) = 1 - Q_0(t)$ be the system reliability, and let $p_i(t) = 1 - q_i(t)$ be the reliability of component i, for $i = 1, 2, \dots, n$. Verify that

$$\frac{dp_S(t)}{dp_i(t)} = \frac{dQ_0(t)}{dq_i(t)}.$$

7.3 Consider the nonrepairable structure in Figure 7.4

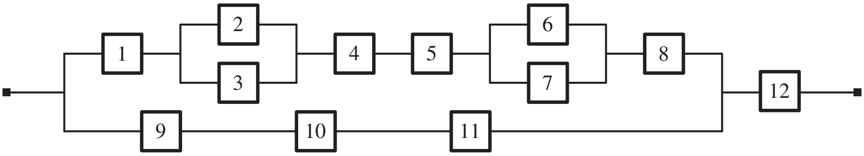


Figure 7.4 RBD (Problem 7.3).

(a) Show that the corresponding structure function may be written as:

$$\begin{aligned}\phi(\mathrm{X}) = [&X_1(X_2 + X_3 - X_2X_3)X_4X_5(X_6 + X_7 - X_6X_7)X_8 \\ &+ X_9X_{10}X_{11} - X_1(X_2 + X_3 - X_2X_3)X_4X_5(X_6 + X_7 - X_6X_7) \\ &X_8X_9X_{10}X_{11}]X_{12}.\end{aligned}$$

(b) Determine the system reliability when the different component reliabilities are given as follows:

$p_1 = 0.970$	$p_5 = 0.920$	$p_9 = 0.910$
$p_2 = 0.960$	$p_6 = 0.950$	$p_{10} = 0.930$
$p_3 = 0.960$	$p_7 = 0.959$	$p_{11} = 0.940$
$p_4 = 0.940$	$p_8 = 0.900$	$p_{12} = 0.990.$

(c) Determine the reliability importance of component 8 by using Birnbaum's metric and the criticality importance metric.

(d) Similarly, determine the reliability importance of component 11, using the same metrics as in (c). Compare and comment on the results obtained.

7.4 Find Birnbaum's reliability importance and structural importance of component 7 of the structure in Figure 7.4.

7.5 Find the reliability importance for component 7 of the structure in Figure 7.4 by using Fussell–Vesely's metric.

7.6 Consider the nonrepairable structure in Figure 7.5

(a) Determine the structure function.

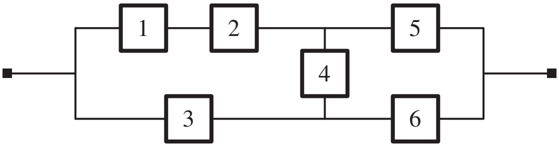


Figure 7.5 RBD (Problem 7.6).

(b) Assume the components to be independent and determine the reliability importance according to Birnbaum's metric for components 2 and 4 when $p_i = 0.99$ for $i = 1, 2, \dots, 6$.

7.7 Consider the nonrepairable structure in Figure 7.5. Assume that the six components are independent, and let the reliability at time t of component i be $p_i(t)$, for $i = 1, 2, \dots, 6$.

(a) Determine Birnbaum's metric of importance of component 3.
(b) Determine the Criticality importance of component 3.
(c) Determine Fussell–Vesely's metric of component 3.
(d) Select realistic values for the component reliabilities and discuss the difference between criticality importance and Fussell–Vesely's metric for this particular system. Show that the relation (7.55) is fulfilled.

7.8 Let (C, ϕ) be a coherent structure of n independent components with state variables $X_1, X_2, \dots, X_n$. Consider the following modular decomposition of (C, ϕ):

(i) $C = \bigcup_{j=1}^{r} A_j$ where $A_i \cap A_j = \emptyset \quad$ for $\; i \neq j$
(ii) $\phi(\mathrm{x}) = \omega(\chi_1(\mathrm{x}^{A_1}), \chi_2(\mathrm{x}^{A_2}), \dots, \chi_r(\mathrm{x}^{A_r}))$

Assume that $k \in A_j$ and show that

- the Birnbaum metric of importance of component k is equal to the product of
- the Birnbaum metric of importance of module j relative to the system, and
- the Birnbaum metric of importance of component k relative to module j.

Is the same relation valid for the other metrics?

References

Barlow, R.E. and Proschan, F. (1975). Importance of system components and fault tree events. *Stochastic Processes and their Applications* 3: 153–173.

Birnbaum, Z.W. (1969). On the importance of different components in a multicomponent system. In: *Multivariate Analysis II* (ed. P.R. Krishnaiah), 581–592. New York: Academic Press.

Borgonovo, E. and Apostolakis, G.E. (2001). A new importance measure for risk-informed decision making. *Reliability Engineering and System Safety* 72: 193–212.

van der Borst, M. and Schoonakker, H.A. (2001). An overview of PSA importance measures. *Reliability Engineering and System Safety* 72: 241–245.

Cheok, M.C., Parry, G.W., and Sherry, R.R. (1998). Use of importance measures in risk-informed regulatory applications. *Reliability Engineering and System Safety* 60: 213–226.

Do, P., Barros, A., and Bérenguer, C. (2008). A new result on the differential importance measures of Markov systems. *9th International Probabilistic Safety Assessment and Management Conference - Proceedings PSAM 9*, pp. 18–23.

Do, P., Barros, A., and Bérenguer, C. (2010). From differential to difference importance measures for Markov reliability models. *European Journal of Operational Research* 204 (3): 513–521. doi: https://doi.org/10.1016/j.ejor.2009.11.025.

Eryilmaz, S. (2016). Computing Barlow-Proschan's importance in combined systems. *IEEE Transactions on Reliability* 65 (1): 159–163.

Fussell, J.B. (1975). How to hand-calculate system reliability and safety characteristics. *IEEE Transactions on Reliability* R-24 (3): 169–174.

Kuo, W. and Zhu, X. (2012). *Importance Measures in Reliability, Risk, and Optimization: Principles and Optimization*. Hoboken, NJ: Wiley.

La Rovere, S., Vestrucci, P., Sperandii, M., and Mandurino, C. (2013). Differential importance measure for components subjected to aging phenomena. *Journal of Quality and Reliability Engineering* 2013: 1–11.

NASA (2011). Probabilistic Risk Assessment Procedures: Guide for NASA Managers and Practitioners, *Guide NASA/SP-2011-3421*. Washington, DC: U.S. National Aeronautics and Space Administration.

NUREG/CR-3385 (1986). Measures of Risk Importance and their Applications. *Report NUREG/CR-3385*. Washington, DC: U.S. Nuclear Regulatory Commission.

Vesely, W.E. (1998). Supplemental viewpoints on the use of importance measures in risk-informed regulatory applications. *Reliability Engineering and System Safety* 60: 257–259.

Vu, H.C., Do, P., and Barros, A. (2016). A stationary grouping maintenance strategy using mean residual life and the Birnbaum importance measure for complex structures. *IEEE Transactions on Reliability* 65 (1): 217–234.

8

Dependent Failures

8.1 Introduction

Chapter 6 deals with systems where n components fail independent of each other, but this assumption of independence is not always realistic. The current chapter starts with a brief repetition of the definitions of statistical independence, dependence, and correlation. Two main types of dependent failures have a special relevance in system reliability: cascading failures and common-cause failures (CCFs). Of these, cascading failures are treated very briefly, whereas CCFs are given a more thorough treatment. The most commonly used models for CCFs are presented through examples and discussed.

8.1.1 Dependent Events and Variables

Two events E_1 and E_2 are (statistically) independent if

$$\Pr(E_1 \mid E_2) = \Pr(E_1) \quad \text{and} \quad \Pr(E_2 \mid E_1) = \Pr(E_2),$$

which means that

$$\Pr(E_1 \cap E_2) = \Pr(E_1)\Pr(E_2). \tag{8.1}$$

Independence implies, in terms of conditional probabilities, that the probability of E_1 is not changed by knowing that event E_2 has occurred, and vice versa.

Consider two independent components 1 and 2 and let E_1 denote that component 1 is functioning and E_2 denote that component 2 is functioning. Further, let E_i^* denote that component i is in a failed state, for $i = 1, 2$. When E_1 and E_2 are independent, E_1 and E_2^* are also independent.[1] This means that if component 2 fails (E_2^*), the state of component 1 is not at all influenced by this failure. In practice, this assumption may not always be realistic.

[1] The proof is left to the reader as Problem 8.2.

System Reliability Theory: Models, Statistical Methods, and Applications, Third Edition.
Marvin Rausand, Anne Barros, and Arnljot Høyland.

Companion Website: www.wiley.com/go/SystemReliabilityTheory3e

When two components are not independent, they are *dependent*, and the dependence can take many different forms. In general, we say that two events E_1 and E_2 are dependent when $\Pr(E_1 \cap E_2) \neq \Pr(E_1)\Pr(E_2)$, which means that

$$\Pr(E_1 \mid E_2) \neq \Pr(E_1) \quad \text{and} \quad \Pr(E_2 \mid E_1) \neq \Pr(E_2).$$

When $\Pr(E_2 \mid E_1) \neq \Pr(E_2)$, two cases may be distinguished:

- $\Pr(E_2 \mid E_1) > \Pr(E_2)$. This means that the probability of E_2 increases when E_1 has occurred, and we say that E_2 is *positively dependent* of E_1.
- $\Pr(E_2 \mid E_1) < \Pr(E_2)$. This means that the probability of E_2 is reduced when E_1 has occurred, and we say that E_2 is *negatively dependent* on E_1.

As for events, we say that two (discrete) random variables X_1 and X_2 are independent when

$$\Pr(X_1 = x_1 \cap X_2 = x_2) = \Pr(X_1 = x_1)\Pr(X_2 = x_2) \qquad \text{for all } x_1 \text{ and } x_2,$$

and X_1 and X_2 are dependent if

$$\Pr(X_1 = x_1 \cap X_2 = x_2) \neq \Pr(X_1 = x_1)\Pr(X_2 = x_2),$$

for at least one combination of x_1 and x_2.

Remark 8.1 (Mutually exclusive versus independent)
Two events, E_1 and E_2, are said to be *mutually exclusive* when $E_1 \cap E_2 = \emptyset$, which means that $\Pr(E_1 \mid E_2) = 0$. By comparing with (8.1), we observe that the two events E_1 and E_2 cannot be both independent and mutually exclusive. The two properties are sometimes confused, and the reader should be aware of the difference. □

8.1.2 Correlated Variables

The degree of correlation between two random variables X_1 and X_2 may be measured by their *covariance*, defined as

$$\text{cov}(X_1, X_2) = E([X_1 - E(X_1)][X_2 - E(X_2)]). \tag{8.2}$$

Pearson's *correlation coefficient*[2] $\rho(X_1, X_2)$ is obtained by scaling the covariance with the standard deviation (SD) of the two variables.

$$\rho(X_1, X_2) = \frac{\text{cov}(X_1, X_2)}{\text{SD}(X_1)\text{SD}(X_2)}, \tag{8.3}$$

where $\text{SD}(X_i) = E([X_i - E(X_i)]^2)$ for $i = 1, 2$.

[2] Named after the British mathematician Karl Pearson (1857–1936).

The correlation coefficient $\rho(X_1, X_2)$ is bounded by the interval $[-1, 1]$. When $\rho(X_1, X_2) = 0$, the two variables are said to be *uncorrelated*, and when $\rho(X_1, X_2) = \pm 1$, they are totally correlated. Positive correlation implies that large (small) values of one variable correspond to large (small) values of the other variable, and negative correlation implies that large (small) values of one variable correspond to small (large) values of the other variable.

Correlation is not the same as independence. The following implications apply:

$$\text{Independent variables} \;\Rightarrow\; \text{Uncorrelated variables}$$
$$\text{Uncorrelated variables} \;\not\Rightarrow\; \text{Independent variables.}$$

Consider a structure of two components that are correlated in such a way that both components tend to fail within a very short time interval. This does not necessarily mean that the two components are dependent and that one failure causes the failure of the other component. It may be that there is an external stress event that causes both components to fail. We should be aware of that correlation does not necessarily imply causation.

$$\text{Correlation} \;\not\Rightarrow\; \text{Causation.}$$

The definition of independence may be extended to n random variables as follows:

$$\Pr\left(\bigcap_{i=1}^{n} X_i = x_i\right) = \prod_{i=1}^{n} \Pr(X_i = x_i) \quad \text{for all } x_1, x_2, \dots, x_n. \tag{8.4}$$

If (8.4) is not fulfilled for all $x_1, x_2, \dots, x_n$, the random variables $X_1, X_2, \dots, X_n$ are dependent.

The correlation coefficients between more than two random variables are set up as a *correlation matrix*, where the entries are pair-wise correlation coefficients as defined above.

Remark 8.2 (Dependence versus interdependence)
The terms "dependence" and "interdependence" are often used in the literature without any clear distinction in meaning. According to our view, an event B *depends* on an event A when A influences B, such that, $A \rightarrow B$. It may, on the other hand, be physically impossible for event B to influence A. The two event A and B are *interdependent* when A influences B and B influences A, such that $A \leftrightarrow B$. See also Section 4.8. □

Several types of dependence were mentioned in Chapter 6, such as the dependence between events in an event tree and the influencing attributes of Bayesian networks.

8.2 Types of Dependence

Consider a structure of n dependent components and assume that the structure is functioning at time t_0. There can be several types of (statistical) dependence.

(1) If component i fails, this failure may increase the probability that another component j will also fail, such that $\Pr(X_j = 0 \mid X_i = 0) > \Pr(X_j = 0 \mid X_i = 1)$. This is typically the situation when two or more components are sharing a load. When one of the components fails, the remaining components have to carry a higher load and the probability of failure increases. In some systems – especially those built as a network – this type of dependence can lead to a long sequence of failures. The failures are often said to show a *domino effect* and the sequence of failures is called a *cascading failure*.

(2) Several components may fail at the same time, or within a limited time interval, due to a shock or some common stress. Relevant types of shocks may include lightning, storm, falling objects, and many more. Common stresses may include humid environment, maintenance errors, installation errors, and many more. Shocks may lead to simultaneous failures, whereas stresses, such as increased humidity, may lead to failures rather close in time. This type of dependent failures is called *common-cause failures*.

(3) A dependent failure where several components in a structure fail in the same way (i.e. with the same failure mode) is called a *common-mode failure*. A common-mode failure is a specific type of CCF where several subsystems fail in the same way for the same reason. The failures may occur at different times and the common cause could be a design defect or a repeated event.

(4) In some cases, the failure of a component may lead to a more benign operating environment for another component. This is the case when a component that produces extreme heat, or heavy vibrations fails. After the failure, the nearby component(s) get an improved environment such that the probability of failure is reduced. This type of dependence is sometimes called *negative dependence* and is not discussed any further in this book.

The remainder of the chapter first provides a brief introduction to cascading failures and thereafter a more thorough treatment of CCFs. For other kind of dependences, the reader may refer to Zhang et al. (2017, 2018a,b,c) for examples.

8.3 Cascading Failures

A cascading failure may be defined as follows:

Definition 8.1 (Cascading failure)
An uncontrolled process in a structure of connected items in which the failure of one or a few items trigger the failure of other items, and so on. □

The cascading sequence of failures in often called a *domino effect*. Cascading failures may happen in many types of systems, such as electric power transmission systems, computer networks, and transportation systems. Such failures have especially been a problem in power transmission and many of the blackouts or power outages have been caused by cascading failures. Several cascading failures have taken place in computer networks, such as the Internet.

Cascading failures may be initiated by a random failure or event, or a deliberate action by a threat actor. For power transmission systems, a cascading failure is called a *cascading outage*. Initiating events that may start a cascading outage in a power transmission system include the following:

- Strong winds
- Heavy snowfall or freezing rain
- Lightning
- Other natural threats, such as avalanche and flooding
- Mechanical failure, e.g. relay or cable joint failure
- Contact between conductors and vegetation
- Maintenance, caused by the isolation of the maintained item, or because of maintenance errors
- Human errors
- Sabotage
- …and many more

Cascading outages in power transmission systems are thoroughly discussed by Sun et al. (2019).

When cascading outages occur, they are usually analyzed thoroughly, and several methods and tools are available for this purpose. To identify and analyze potential (i.e. future) cascading failures in power transmission systems is a very difficult task, if at all possible. There are a multitude of initiating events, and how they occur and where they occur strongly influence the further development. The cascading effects can follow many different trajectories and spread extremely fast. Some analysts consider cascading failures to be *emergent properties* of complex transmission systems. The systems do not comply with the Newtonian–Cartesian paradigm (see Chapter 2) and are hence not possible to analyze properly (see also Perrow 1984).

Potential cascading failures in simple network systems may partly be studied by Markov methods (see Chapter 11). In power transmission, several Monte Carlo simulation programs have been developed to study cascading failures, or selected trajectories of such failures.

Some analysts claim that power transmission systems are vulnerable to cascading outages because of the protection policies, where

- Upper and lower thresholds are assigned to many types of items.
- A variety of relays are used to remove items from service when their thresholds are crossed.
- The system is operated close to threshold values to optimize profit.

Example 8.1 (Fukushima nuclear disaster)

The Fukushima nuclear accident on 11 March 2011 was a classical example of a cascading failure. The accident started by a magnitude 9.0 earthquake off the eastern coast of Japan. The earthquake generated a tsunami that flushed into the nuclear plant and virtually destroyed the whole plant. All the six reactors were shut down as designed when triggered by the earthquake, but the plant's seawater cooling pumps were damaged by the tsunami and the emergency electrical generators were flooded. As a result, the plant was left without means to cool the reactors and the spent nuclear fuel that was stored on-site. The resulting explosions and fires released high levels of radioactive contamination into the air, ocean, and on land. For more information about the Fukushima accident, you may consult Little (2012) or search the Internet. □

8.3.1 Tight Coupling

The seminal book Perrow (1984) classifies systems on a scale from loosely coupled to tightly coupled. Tightly coupled systems are, according to Perrow, vulnerable to major failures or accidents.

The main characteristics of a tightly coupled system include

- Time-dependent processes that cannot wait.
- Rigidly ordered processes (sequence A must follow B) – direct and immediate connection and interaction between components.
- Fast and time-dependent processes – they happen quickly and cannot be turned off or isolated.
- Only one path to a successful outcome.
- Little or no slack (requiring precise quantities of specific resources for successful operation).
- Little opportunity for mitigation or defense once an initial disturbance or fault occurs.
- Rapid response.
- Fast propagation of disturbances – operators do not have time or ability to determine what is wrong.
- Limited substitutions.

The listed characteristics are not disjoint. We observe that these characteristics are typical for a system that is vulnerable to cascading failures.

8.4 Common-Cause Failures

CCF analysis was introduced in the nuclear power industry in the 1970s. This industry has had a continuous focus on CCFs and has been in the forefront of the development of CCF models, and on collection and analysis of data related to CCFs. The aerospace industry has also given these failures close attention, and the offshore oil and gas industry has at least since the 1990s focused on CCFs related to reliability assessment of safety systems. CCF analysis is now mandated by IEC 61508 for safety-instrumented systems (see Chapter 13).

CCFs are mainly relevant for redundant structures and is often restricted to *voted groups*, that is a subsystem of n items that are configured as a $k$oo$n$:G structure, where $k < n$. As illustrated in Example 8.5, CCF is not a problem for series structures (i.e. when $k = n$). In voted groups, the minimal paths are often called *channels*. A channel is a structure of one or more components that can independently perform a required function.

In the rest of this chapter, the study object is a voted group of n components (or channels) that is configured as a $k$oo$n$:G structure. CCF analysis is most relevant for safety protection systems, such as safety-instrumented systems (see Chapter 13). A CCF may be defined as follows:

Definition 8.2 (Common-cause failure)
Failures, that are the direct result of a shared cause, after which two or more components in a multicomponent structure are in fault state at the same time, leading to failure of the structure. □

To be classified as a CCF, this definition requires the CCF to lead to structure failure. A 2oo4:G structure is functioning as long as at least two components are functioning. If exactly two components are in a failed state at the same time, this is a *multiple fault with a shared cause*, but it is not a CCF because the structure is still functioning.

To be a CCF, it is not required that the components fail exactly at the same time. What is important, is that the failed components are in a failed state at the same time. Nonsimultaneous failures may sometimes be detected and repaired before the next failure occurs, thus avoiding CCF.

The relationship between independent (or individual) failures and CCFs of a structure of two components is illustrated in Figure 8.1. The number of components that fail due to the common cause is called the *multiplicity* of the CCF.

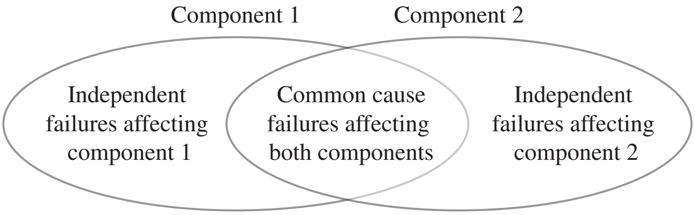

Component 1
Component 2
Independent failures affecting component 1
Common cause failures affecting both components
Independent failures affecting component 2

Figure 8.1 Relationship between independent failures and CCFs of a structure with two components.

Remark 8.3 (Increased stress)
Increased stress may not only lead to a CCF, but may also increase the independent failure rates of the affected components. In some cases, the increased stress may imply that the constant failure rate assumption is no longer valid and that the components have increasing failure rate functions. □

8.4.1 Multiple Failures that Are Not a CCF

As mentioned above, a multiple failure with a shared cause does not need to be a CCF. It is sometimes useful to have a specific term. The following term is therefore introduced:

Definition 8.3 (Multiple failure with a shared cause, MFSC)
Failure, that is a direct result of a shared cause, in which two or more items are in failed state simultaneously. □

An MFSC is also called a *CCF event*, but the authors prefer the term MFSC because the term CCF event may be confused with a CCF. Observe that when an MFSC leads to system failure, then the MFSC is a CCF of the system. CCFs have particularly been focused in systems where there is a high risk for fatal accidents. Methods for controlling and preventing such failures have been developed during safety analyses within the aviation and the nuclear power industry.

8.4.2 Causes of CCF

The causes of a CCF may be split into *shared causes* and *coupling factors*. A shared cause is a cause of a component failure (e.g. high humidity), whereas a coupling factor explains why several components are affected by the same cause. The relation between a CCF, the shared cause, and the coupling factors is illustrated in Figure 8.2 for a parallel structure of two components, that is, a voted group with two single-component channels.

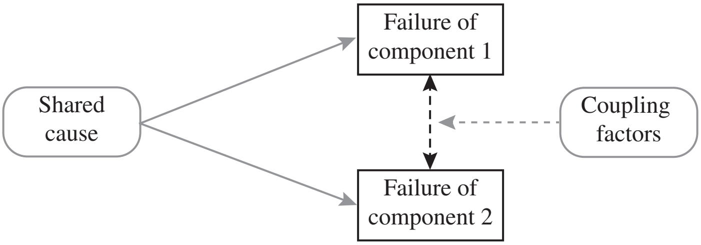


Figure 8.2 A shared cause combined with coupling factors lead to CCF of a parallel structure of two components.

Shared causes. A number of studies have investigated the shared causes of CCF events, and several classification schemes have been proposed and used to categorize these events. Several studies of CCFs have shown that a majority of the root causes are related to human actions and procedural deficiencies. A study of centrifugal pumps in nuclear power plants indicates that the causes of nearly 70% of all CCFs are of this category (Miller et al. 2000).

Coupling factors. A coupling factor is a property that makes multiple channels susceptible to failure from a single shared cause. Such properties include the following:

- Same design
- Same hardware
- Same software
- Connections to the same network (e.g. Internet)
- Same installation staff
- Same maintenance or operation staff
- Same procedures
- Same environment
- Same location

More detailed taxonomies of coupling factors are available in NEA (2004), NUREG/CR-5485 (1998), and Childs and Mosleh (1999). Studies of CCFs in nuclear power plants indicate that the majority of coupling factors contributing to CCFs are related to operational aspects (Miller et al. 2000).

To save money and ease operation and maintenance, the technical solutions in many industries become increasingly standardized. This applies both to hardware and software and increases the presence of coupling factors. SINTEF, the Norwegian research organization, has made several studies of the impacts of this type of standardization on Norwegian offshore oil and gas installations, where new operational concepts and reduced manning levels are feeding this trend (Hauge et al. 2006).

When studying structures that are vulnerable to CCFs, it is often helpful to identify components with similar vulnerabilities to CCFs. Such a group of components is called a common-cause component group (CCCG) and may be defined as follows:

Definition 8.4 (Common-cause component group, CCCG)
A group of components that share one or more coupling factors, making them susceptible to CCFs. □

CCCGs' are discussed thoroughly in NUREG/CR-5485 (1998).

8.4.3 Defenses Against CCF

A number of possible defense measures against CCFs have been proposed among which are the following:

Separation or segregation. The separation can be both physical and logical and enhances the independence of the components and reduces the susceptibility to both CCFs and cascading failure.

Diversity. Different components and different technologies reduce the likelihood of coupling factors and the susceptibility to CCFs.

Robustness. A robust design has a higher ability to withstand environmental stresses (e.g. see DOE-STD-1195 2011).

Component reliability. High component reliability reduces the number of both individual and dependent failures, and thereby the number of maintenance actions and human interventions (which are recognized causes of CCFs)

Simplicity of design. A simple design is easier to understand and maintain and reduces the number of intervention errors.

Analysis. Failure modes, effects, and criticality analysis (FMECA) and other reliability analyses can identify causes of CCFs and propose measures to reduce the likelihood of CCFs

Procedures and human interface. Clear procedures and an adequate human–machine interface reduce the likelihood of human errors.

Competence and training. Designers, operators, and maintainers can help to reduce CCFs by understanding shared causes and coupling factors.

Environmental control. The susceptibility to CCFs can be reduced by weather proofing.

Diagnostics and coverage. A diagnostic system with high coverage can reveal the first nonsimultaneous CCFs and bring the system to a safe state before the next failure occurs.

Remark 8.4 (Condition monitoring and software)

Condition monitoring is increasingly used in many technical items. To serve its purpose, the condition monitoring equipment is connected to a computer network, and more and more often to the Internet. This way, an item expert may sit in a distant place and survey the item condition and recommend maintenance actions and when these actions should be carried out. Even if items are diversified, the condition monitoring equipment and/or the associated software may be the similar, thus leading to coupling factors. Dedicated cyberattacks may also take down several items at the same time. The same applies for other software-based functions that are implemented in modern technical items that are connected to the Internet. □

8.5 CCF Models and Analysis

Two different methods can be used to model CCFs, an explicit method and an implicit method. There are two main categories of CCFs:

(1) CCFs due to clear deterministic causes.
(2) Residual CCFs that are not considered explicitly in the system models (e.g. fault trees) because we do not see any clear deterministic causes, we do not fully understand how they can occur, or it is not possible to obtain reliability data.

The first category should, as far as possible, be modeled explicitly.

A number of implicit CCF models have been proposed and O'Connor (2013) classifies these models into three categories:

(1) Direct method – basic parameter model (BPM)
(2) Ratio models
(3) Shock models

8.5.1 Explicit Modeling

Assume that a specific cause of a CCF can be identified and defined. By the explicit method, this cause of dependence is included into the system logic models. This modeling is illustrated in Figure 8.3, for a 1oo2:G structure of two

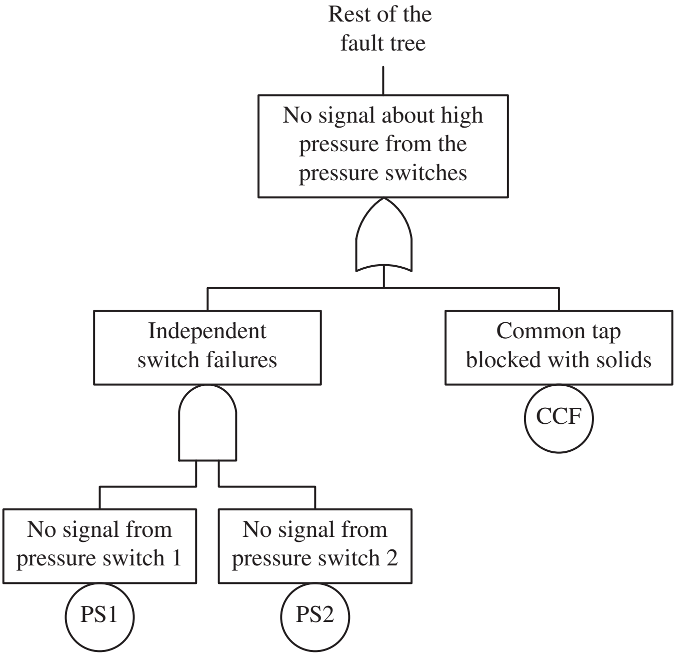

Rest of the fault tree
No signal about high pressure from the pressure switches
Independent switch failures
Common tap blocked with solids
CCF
No signal from pressure switch 1
No signal from pressure switch 2
PS1
PS2

Figure 8.3 Explicit modeling of a CCF in a system of two pressure switches.

pressure switches that are installed on a pressure vessel. The two switches are installed on a common tap (thin pipe) from the pressurized vessel. The 1oo2:G structure can fail in two different ways: (i) as two independent failures or (ii) as a CCF because the common tap is blocked by some sort of solids.

Examples of causes that may be modeled explicitly are the following:

- Human errors
- Utility failures (e.g. electricity, cooling, and heating)
- Environmental events (e.g. earthquake, and lightning).

8.5.2 Implicit Modeling

Some causes of dependencies are difficult or even impossible to identify and model explicitly. These are called residual causes and are catered for by an implicit model. The residual causes cover many different shared causes and coupling factors, such as common manufacturer, common environment, and maintenance errors. There are so many causes that an explicit representation of all of them in a fault tree or an event tree would not be manageable.

When establishing the implicit model, it is important to remember which causes were covered in the explicit model so that they are not counted twice.

8.5.3 Modeling Approach

Modeling and analysis of CCFs as part of a reliability study should, in general, comprise at least the following steps:

(1) *Development of system logic models*. This activity comprises system familiarization, system functional failure analysis, and establishment of system logic models (e.g. fault trees and reliability block diagrams (RBDs)).

(2) *Identification of common-cause component groups*. The groups of components with similar vulnerabilities to CCFs are identified.

(3) *Identification of shared causes and coupling factors*. The shared causes and coupling factors are identified and described for each CCCG. Suitable tools are checklists and root cause analysis.

(4) *Assessment of component defenses*. The CCCGs are evaluated with respect to their defenses against the root causes that were identified in the previous step.

(5) *Explicit modeling*. Explicit CCF causes are identified for each CCCG and included into the system logic model.

(6) *Implicit modeling*. Residual CCF causes that were not covered in the previous step are included in an implicit model as discussed later in this chapter. The parameters of this model have to be estimated based on checklists (e.g. see IEC 61508 2010, Part 6) or from available data.

(7) *Quantification and interpretation of results*. The results from the previous steps are merged into an overall assessment of the system.

In most cases, we are not able to find high-quality input data for the explicitly modeled CCF causes. However, even with low-quality input data, or guesstimates, the result is usually more accurate than by including the explicit causes into a general (implicit) CCF model.

The CCF models that are discussed in the rest of this section are limited to covering implicit causes of CCF.

8.5.4 Model Assumptions

The following assumptions apply for the CCF models presented in the remainder of this chapter.[3]

(1) The study object is a voted group of n identical components. The voted group is written $k$oo$n$:G for functioning structures and $k$oo$n$:F for failed structures.
(2) There is a complete symmetry in the n components, and each component has the same constant failure rate.
(3) All specific combinations, where k components are failing and $n-k$ components are not failing, have the same probability to occur.
(4) Removing j of the n components has no effect on the probabilities of failure of the remaining $n-j$ components.

These assumptions imply that we do not have to specify completely new parameters for each n. The parameters defined to handle CCF for $n = 2$ are retained for $n = 3$, and so on.

8.6 Basic Parameter Model

The BPM, was proposed by Fleming et al. (1983), and can be applied to a $k$oo$n$:F voted group of identical components. A component that has failed can be an individual (single) fault or one fault in a set of multiple faults. The variable of interest in the BPM is the *multiplicity* of the fault and its distribution.

To illustrate the approach, consider a voted group of $n = 3$ identical components (e.g. gas detectors), which may have hidden failures that are revealed in a proof test (see Chapter 13). Assume that all the n components have been proof-tested at time t. Let E_i^* denote that component i is found to be functioning and that component E_i is found to be in a failed state, for $i = 1, 2, 3$. A specific component, say component 1, can be involved in four disjoint fault scenarios:

[3] The following treatment of implicit CCF modeling is a reworked and updated version of Hokstad and Rausand (2008).

(1) Component 1 is failed, as an individual (single) fault, that is, $E_1 \cap E_2^* \cap E_3^*$.
(2) Component 1 is, together with component 2, involved in a double fault, that is, $E_1 \cap E_2 \cap E_3^*$.
(3) Component 1 is, together with component 3, involved in a double fault, that is, $E_1 \cap E_2^* \cap E_3$.
(4) Component 1 has, together with components 2 and 3, a triple fault, that is, $E_1 \cap E_2 \cap E_3$.

Similar expressions can be established for components 2 and 3.

8.6.1 Probability of a Specific Multiplicity

Let $g_{i,n}$ be the probability of a *specific* combination of functioning and failed components, such that (exactly) i components are in fault state, and $n - i$ components are functioning. The probability of a specific single (individual) fault in a voted group of three identical components is

$$g_{1,3} = \Pr(E_1 \cap E_2^* \cap E_3^*) = \Pr(E_1^* \cap E_2 \cap E_3^*) = \Pr(E_1^* \cap E_2^* \cap E_3). \tag{8.5}$$

The probability of a specific double fault is

$$g_{2,3} = \Pr(E_1 \cap E_2 \cap E_3^*) = \Pr(E_1 \cap E_2^* \cap E_3) = \Pr(E_1^* \cap E_2 \cap E_3), \tag{8.6}$$

and the probability of a triple fault is

$$g_{3,3} = \Pr(E_1 \cap E_2 \cap E_3). \tag{8.7}$$

These probabilities are shown in the Venn diagram in Figure 8.4. Let $Q_{k:3}$ be the probability that a voted group of three identical components has a (unspecified) fault with multiplicity k, for $k = 1, 2, 3$. These probabilities are the following:

$$\begin{aligned} Q_{1:3} &= \binom{3}{1} g_{1,3} = 3g_{1,3} \\ Q_{2:3} &= \binom{3}{2} g_{2,3} = 3g_{2,3} \\ Q_{3:3} &= \binom{3}{3} g_{3,3} = g_{3,3}. \end{aligned} \tag{8.8}$$

These probabilities can also be deducted from Figure 8.4.

A 1oo3:F voted group fails when at least one of its three components fails. The probability that the 1oo3:F voted group fails is then

$$Q_{1\text{oo}3:\text{F}} = Q_{1:3} + Q_{2:3} + Q_{3:3} = 3g_{1,3} + 3g_{2,3} + g_{3,3}.$$

Similarly, a 2oo3:F voted group fails when at least two of its three components fail. The probability that the 2oo3:F voted group fails is then

$$Q_{2\text{oo}3:\text{F}} = Q_{2:3} + Q_{3:3} = 3g_{2,3} + g_{3,3}$$

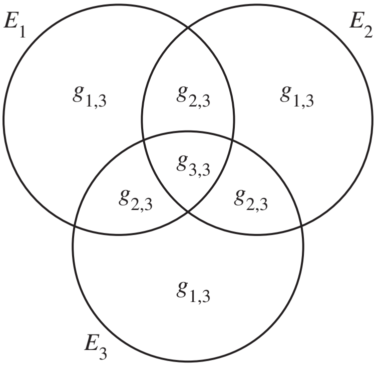


Figure 8.4 Probabilities of different multiplicities for a voted group of three identical channels.

and, a 3oo3:F voted group fails when all its three components fail and the probability that the 3oo3:F voted group fails is

$$Q_{3oo3:F} = Q_{3:3} = g_{3,3}.$$

When proof-testing a voted group of three identical components, the probability that a particular component (say component 1) has failed is

$$Q = g_{1,3} + 2g_{2,3} + g_{3,3}, \tag{8.9}$$

where Q is the total probability that a specific component is found to have failed, comprising both individual and multiple faults.

This probability may be written as

$$Q = \sum_{i=1}^{3} \binom{3-1}{i-1} g_{i,3}, \tag{8.10}$$

and it is not so difficult to show that for a voted group of n components, this expression becomes

$$Q = \sum_{i=1}^{n} \binom{n-1}{i-1} g_{i,n}. \tag{8.11}$$

8.6.2 Conditional Probability of a Specific Multiplicity

Assume that a particular component is proof-tested at time t and found to be in a failed state in a voted group of three identical components. Without loss of generality, we can assume that this is component 1. As above, let Q be the probability of this event. When such a fault is observed, the multiplicity of faults is either 1, 2, or 3. Let $f_{i,3}$ be the *conditional* probability that the fault has multiplicity i when we know that a specific component has failed, for $i = 1, 2, 3$. For a triple fault, the fault of component 1 is included in the triple fault, and we have

$$f_{3,3} = \Pr(E_1 \cap E_2 \cap E_3 \mid E_1) = \frac{\Pr(E_1 \cap E_2 \cap E_3)}{\Pr(E_1)} = \frac{g_{3,3}}{Q}. \tag{8.12}$$

For a double fault, the fault of component 1 is included in two of the three possible fault combinations in (8.8). By using the same argument as above, the conditional probability of a double fault involving components 1 and 2 is

$$f_{2,3}^{(1,2)} = \frac{g_{2,3}}{Q},$$

and the conditional probability of a double fault involving components 1 and 3 is

$$f_{2,3}^{(1,3)} = \frac{g_{2,3}}{Q}.$$

The superscripts in $f_{2,3}^{(1,2)}$ and $f_{2,3}^{(1,3)}$ indicate which components are involved in the fault. The conditional probability of a double fault involving component 1 and one of the other components is

$$f_{2,3} = f_{2,3}^{(1,2)} + f_{2,3}^{(1,3)} = \frac{2g_{2,3}}{Q}. \tag{8.13}$$

For a single fault, the fault of component 1 is included in only one fault combination in (8.8). The conditional probability that the fault of component 1 is a single fault is

$$f_{1,3} = \frac{g_{1,3}}{Q}. \tag{8.14}$$

These probabilities are easily seen from Figure 8.4. If it is known, for example, that component 1 is in a failed state, the probabilities have to be found in the circle representing fault of component 1 (i.e. E_1).

The above probabilities can be estimated from observed data – if sufficient data are available – as

$$Q_{n:i} = \frac{m_i}{m_{\text{tot}}}, \tag{8.15}$$

where m_i is the observed number of faults of multiplicity i and m_{tot} is the total number of proof tests of the voted group. It is assumed that each time the voted group is tested, all the n components in the group are tested.

The BPM cannot estimate CCFs for voted groups for which data is unavailable, and for this reason is rarely used directly. A more detailed discussion of the BPM may be found in Hokstad and Rausand (2008) and NUREG/CR-5485 (1998).

8.7 Beta-Factor Model

The beta-factor model was introduced by Fleming (1975) and is the simplest and most commonly used model for CCF analysis. The model is applicable for a voted group of n identical components, requires only two parameters, and is the main CCF model in IEC 61508.

The idea of the beta-factor model is to split the constant failure rate, λ, of a component into two parts, one part, $\lambda^{(i)}$, covering the individual failures of the component, and another part, $\lambda^{(c)}$, covering CCFs.

$$\lambda = \lambda^{(i)} + \lambda^{(c)}. \tag{8.16}$$

The beta-factor, β, is introduced as

$$\beta = \frac{\lambda^{(c)}}{\lambda}, \tag{8.17}$$

and is the fraction of all the component failures that are CCFs.

Example 8.2 (Interpretation of the beta-factor)
Consider a system component with constant failure rate λ that has failed 100 times. If the beta-factor is $\beta = 0.10$, approximately 90 of the failures are individual (i.e. independent) failures and ten failures are CCFs that also involve the other components in the system. □

The parameter β can be interpreted as the conditional probability that a component failure is in fact a CCF.

$$\beta = \Pr(\text{CCF} \mid \text{Failure of component}).$$

The individual failure rate and the CCF rate can be expressed by the total component failure rate λ and the factor β as

$$\begin{aligned}\lambda^{(c)} &= \beta\lambda \\ \lambda^{(i)} &= (1-\beta)\lambda.\end{aligned}$$

The beta-factor model is based on the assumption that when a CCF occurs, it affects all the components of the voted group, such that we either have individual failures or a total failure affecting all components.

The data inputs to the beta-factor model are the total failure rate λ and the beta-factor β. If λ is kept constant and we make adjustments to the system design such that β is reduced, the result of this adjustment is that the CCF failure rate $\beta\lambda$ is reduced, whereas the individual failure rate $(1-\beta)\lambda$ increases.

The beta-factor model may be regarded as a *shock model* where shocks occur randomly according to a homogeneous Poisson process (HPP) with rate $\lambda^{(c)}$. Each time a shock occurs, all the system components fail at the same time, irrespective of the status of the components. Each component may hence fail due to two independent causes: shocks and component-specific (individual) causes. We may choose different beta-factors for the different component failure modes.

8.7.1 Relation to the BPM

Assume that a particular component of the voted group is proof-tested and found in a failed state. The probability of this event is in the BPM denoted Q. The fault is either an individual fault or a total fault involving all the n components of the system. An individual fault occurs with probability $g_{1,n}$ and a total CCF occurs with probability $g_{n,n} = Q_{n:n}$, such that

$$Q = g_{1,n} + Q_{n:n}. \tag{8.18}$$

The parameter β is defined as the fraction of the total failure probability attributable to dependent failures, such that

$$\beta = \frac{Q_{n:n}}{Q} = \frac{Q_{n:n}}{Q_{n:n} + g_{1,n}},$$

and we get

$$\begin{aligned} g_{1,n} &= (1-\beta)Q \\ g_{i,n} &= 0 \quad \text{for } i = 2, 3, \dots, n-1 \\ g_{n,n} &= \beta Q. \end{aligned} \tag{8.19}$$

The conditional probabilities of specific multiplicities, given the fault of a particular component is

$$\begin{aligned} f_{1,n} &= 1-\beta \\ f_{i,n} &= 0 \quad \text{for } i = 2, 3, \dots, n-1 \\ f_{n,n} &= \beta. \end{aligned} \tag{8.20}$$

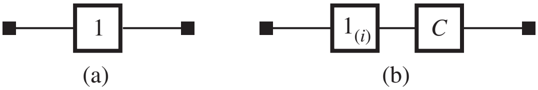
Remark 8.5 (Unreliable components have higher beta-factor)
For a fixed β, the rate of CCFs, $\lambda^{(c)} = \beta\lambda$, in the beta-factor model is seen to increase with the failure rate λ. Therefore, systems with many failures have many CCFs. Because repair and maintenance is often claimed to be a prime cause of CCFs, it is relevant to assume that systems requiring a lot of repair also have many CCFs. □

8.7.2 Beta-Factor Model in System Analysis

In line with (8.16), the RBD in Figure 8.5 interprets a component 1 as a series structure of two blocks, the first block $1_{(i)}$ represents component 1 exposed to individual failure and block C represents component 1 exposed to CCF.

The use of the beta-factor model in system analysis is illustrated in the following examples.

Figure 8.5 A component represented as a series structure of two blocks.



Example 8.3 (Parallel structure of two identical components)
Consider the parallel structure of two identical components with constant total failure rate λ in Figure 8.6 (i.e. a 1oo2:G voted group). The structure is exposed to CCFs modeled by a beta-factor model. Part (a) of the figure shows the traditional RBD for a parallel structure. In part (b), each component is split into two blocks as shown in Figure 8.6. Because the block C, representing a component exposed to CCF, is identical for both components, the RBD in part (b) can be redrawn to the RBD in part (c). The contribution from CCFs can therefore be modeled in an RBD as a block C in series with the rest of the structure.

The survivor function of the 1oo2:G structure is

$$R_S(t) = [2e^{-(1-\beta)\lambda t} - e^{-2(1-\beta)\lambda t}]\, e^{-\beta t} = 2e^{-\lambda t} - e^{-(2-\beta)\lambda t}. \tag{8.21}$$

The mean time-to-failure (MTTF) of the 1oo2:G structure is

$$\text{MTTF}_{1\text{oo}2:\text{G}} = \frac{2}{\lambda} - \frac{1}{(2-\beta)\lambda}. \tag{8.22}$$

The fractions of individual (independent) failures and CCFs for this structure are shown in Figure 8.7. Because the failure rates are constant, the number of failures that occur in a time interval of length t is determined from a HPP. If the parallel structure is observed during a long time interval $(0, t)$, the numbers of observed failures are

Mean number of individual failures	$2(1-\beta)\lambda t$
Mean number of double (i.e. CCF) failures	$\beta\lambda t$

□

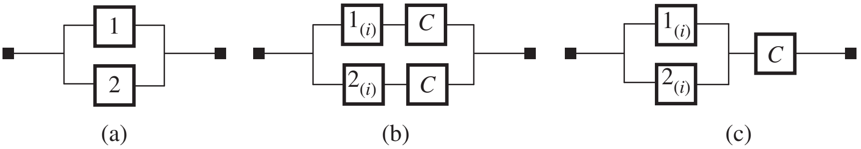
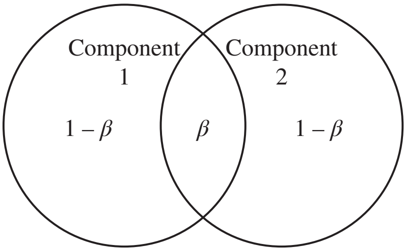


Figure 8.6 A parallel structure modeled by the beta-factor model.

Figure 8.7 Fractions of different types of failures for a structure of two components when using a beta-factor model.



Example 8.4 (2oo3:G structure of identical components)

Consider a 2oo3:G structure of identical components with constant failure rate λ and beta-factor β. The structure may be represented by the RBD in Figure 8.8. The survivor function of the 2oo3:G structure is

$$\begin{aligned} R(t) &= (3e^{-2(1-\beta)\lambda t} - 2e^{-3(1-\beta\lambda t})e^{-\beta\lambda t} \\ &= 3e^{-(2-\beta)\lambda t} - 2e^{-(3-2\beta)\lambda t}. \end{aligned} \tag{8.23}$$

The MTTF of the 2oo3:G structure is

$$\text{MTTF}_{2\text{oo}3:\text{G}} = \frac{3}{(2-\beta)\lambda} - \frac{2}{(3-2\beta)\lambda}. \tag{8.24}$$

The MTTF is shown as a function of β in Figure 8.9. Observe that:

(a) When the three components are independent (i.e. $\beta = 0$), the $\text{MTTF}_{2\text{oo}3:\text{G}}$ is shorter than the MTTF_1 of a single component. We get $\text{MTTF}_{2\text{oo}3:\text{G}} = (5/6)\ \text{MTTF}_1 \approx 0.833\ \text{MTTF}_1$.

(b) When $\beta = 1$, all the three components fail when one of them fails and $\text{MTTF}_{2\text{oo}3:\text{G}} = \text{MTTF}_1$.

(c) Setting MTTF in (8.24) equal to 1 and solving for β yields $\beta = 0.5$. This means that with $\beta = 0.5$, the MTTF of the 2oo3:G structure is equal to the MTTF of a single component.

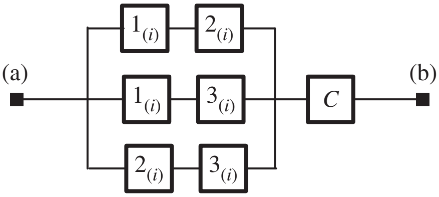


Figure 8.8 RBD for a 2oo3:G structure modeled by the beta-factor model.

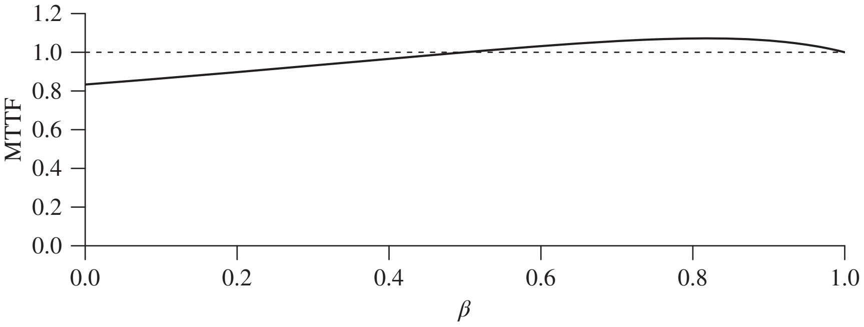
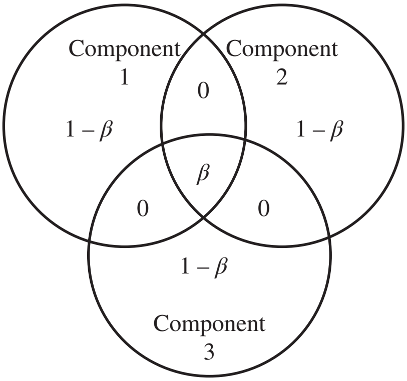


Figure 8.9 The MTTF of a 2oo3:G structure modeled as a function of the beta-factor β, for $\lambda = 1$.

Figure 8.10 Fractions of different types of failures for a system with three components when using a beta-factor model.



The fractions of individual (independent) failures and CCFs for this structure are shown in Figure 8.10. Because the failure rates are constant, the number of failures that occur in a time interval of length t is determined from a HPP. If the parallel structure is observed during a long time interval $(0, t)$, the numbers of observed failures are

Number of individual failures	$3(1-\beta)\lambda t$
Number of double failures	0
Number of triple (i.e. CCF) failures	$\beta\lambda t$

Observe that double failures are not allowed with the beta-factor model. □

Example 8.5 (Series structure of *n* identical components)
Consider a series structure of n identical components with constant failure rate λ. The structure is exposed to CCFs, and these are modeled by a beta-factor model with parameter β. The survivor function of the series structure is

$$R_S(t) = e^{-n(1-\beta)\lambda t} e^{-\beta\lambda t} = e^{-[n-(n-1)\beta]\lambda t}. \tag{8.25}$$

The MTTF of the series structure is

$$\text{MTTF}_{1\text{oo}n:\text{G}} = \frac{1}{[n-(n-1)\beta]\lambda} = \frac{n}{n-(n-1)\beta}\frac{1}{n\lambda}. \tag{8.26}$$

Observe that $1/n\lambda$ is the $\text{MTTF}^{(i)}$ of the series structure for independent components (i.e. for $\beta = 0$). The MTTF of the series structure with beta-factor β is obtained by multiplying $\text{MTTF}^{(i)}$ with the scaling factor (sf):

$$\text{sf} = \frac{n}{n-(n-1)\beta}.$$

Numerical example: Let $n = 10$. The sf is calculated for some selected values of β as

β	0	0.05	0.10	0.15	0.50	1.00
sf	1.000	1.047	1.098	1.156	1.818	10

Observe that the MTTF increases with β. When $\beta = 1$, the series structure behaves as a single component with failure rate λ and MTTF $= 1/\lambda$. We observe that the reliability of a series structure increases with increasing β when using the beta-factor model. This is an obvious result as seen by the beta-factor model, but some readers may find it a bit strange. □

The beta-factor model is simple, and it is easy to understand the practical interpretation of the parameter β. A serious limitation of the beta-factor model is that it does not allow that only a certain fraction of the components fails. The model seems quite adequate for parallel structures of two components, but may not be fully adequate for more complicated structures. NUREG/CR-4780 states that:

> Although historical data collected from the operation of nuclear power plants indicate that common cause events do not always fail all redundant components, experience from using this simple model shows that, in many cases, it gives reasonably accurate (or slightly conservative) results for redundancy levels up to about three or four items. However, beyond such redundancy levels, this model generally yields results that are conservative.

The beta-factor is further discussed with many examples in Chapter 13. Input data to the beta-factor model are discussed in Chapter 16.

8.7.3 Beta-Factor Model for Nonidentical Components

The beta-factor model presented above is defined for identical components, but many systems are diversified with components that are nonidentical. In this case, it is more difficult to define and interpret the beta-factor. An approach that is sometimes used is to define β as a percentage of the *geometric average* (see box) of the failure rates of the various components of the group (e.g. see Hauge et al. 2013).

Arithmetic Versus Geometric Average

Consider a data set $\{a_1, a_2, \dots, a_n\}$.

The *arithmetic average* of the data set is $\overline{a} = \frac{1}{n} \sum_{i=1}^{n} a_i$

The *geometric average* of the data is $a^* = \left(\prod_{i=1}^{n} a_i\right)^{\frac{1}{n}} = \sqrt[n]{a_1 a_2 \ \cdots \ a_n}$

For a data set of two entries: $a_1 = 1$ and $a_2 = 10$, we have

$\overline{a} = (1 + 10)/2 = 5.5$ and $a^* = \sqrt{1 \cdot 10} = \sqrt{10} \approx 3.16$

Example 8.6 (Parallel structure of *n* nonidentical components)
Consider a parallel structure of n nonidentical components. The structure is exposed to CCFs, and we assume that this can be modeled by a beta-factor model. Let λ_i be the (total) failure rate of component i, for $i = 1, 2, \dots, n$. The geometric average of the n failure rates is

$$\lambda = \left(\prod_{i=1}^{n} \lambda_i \right)^{1/n}. \tag{8.27}$$

The beta-factor β can be determined as a fraction of this average failure rate λ, and the individual failure rate of component i becomes

$$\lambda_i^{(i)} = \lambda_i - \beta\lambda. \tag{8.28}$$

The survivor function of the structure is

$$R_S(t) = \left[1 - \prod_{i=1}^{n} (1 - e^{-(\lambda_i - \beta\lambda)t}) \right] e^{-\beta\lambda t}.$$

□

This approach in Example 8.6 may be acceptable when all the failure rates are in the same order of magnitude. When the failure rates are very different, this approach can lead to unrealistic results, as illustrated in Example 8.7.

Example 8.7 (Beta-factor with very different failure rates)
Consider a parallel structure of two components. The failure rate of component 1 is $\lambda_1 = 1 \times 10^{-4}\ \text{h}^{-1}$, and the failure rate of component 2 is $\lambda_2 = 1 \times 10^{-8}\ \text{h}^{-1}$. The two components are exposed to CCFs that can be modeled by a beta-factor model. The geometric average of the two failure rates is, according to (8.27),

$$\lambda = (\lambda_1 \lambda_2)^{1/2} = \sqrt{10^{-4} \cdot 10^{-8}} = 1 \times 10^{-6}\ \text{h}^{-1}.$$

If we suggest a β of 10%, the CCF rate becomes $\lambda^{(c)} = \beta\lambda = 10^{-7}\ \text{h}^{-1}$. This is clearly impossible as the total failure rate of the strongest component is $\lambda_2 = 10^{-8}\ \text{h}^{-1}$, and the rate of CCFs of a component can never be higher than the total failure rate.

This example shows that the suggested approach cannot be suitable when the components have very different failure rates. □

Another problematic issue is illustrated in Example 8.8.

Example 8.8 (2oo3:G voted group with different failure rates)
Consider a 2oo3:G structure, where components 1 and 2 are identical, with failure rate $\lambda_{12} = 5 \times 10^{-7}\ \text{h}^{-1}$, and component 3 is different with failure rate $\lambda_3 = 2 \times 10^{-6}\ \text{h}^{-1}$ (a possible example of such a system might be a system of two smoke detectors and one flame detector).

If the third component had been of the same type as components 1 and 2, we would have used a beta-factor model with $\beta_{12} = 0.10$ for the whole group. The third component is different from components 1 and 2 and the likelihood of a CCF involving all the three components is considered to be very low, such that a beta-factor model with $\beta_{\text{all}} = 0.01$ might be suggested.

Because the group is voted 2oo3:G, it is sufficient that two components fail for the group to fail. If a CCF involving components 1 and 2 occurs, the group fails. The rate of group CCFs should consequently be

$$\lambda_{\text{S}}^{(c)} \geq \lambda_{12}^{(c)}\, \beta_{12} = 5 \times 10^{-8}\ \text{h}^{-1}.$$

How this situation should be treated by the approach suggested above is far from obvious. □

8.7.4 *C*-Factor Model

The C-factor model was introduced by Evans et al. (1984) and is essentially the same model as the beta-factor model but defines the fraction of CCFs in another way. In the C-factor model, the CCF rate is defined as $\lambda^{(c)} = C\lambda^{(i)}$, that is as a fraction of the individual failure rate $\lambda^{(i)}$. The total failure rate may then be written as $\lambda = \lambda^{(i)} + C\lambda^{(i)}$. In this model, the individual failure rate $\lambda^{(i)}$ is kept constant and the CCF rate is added to this rate to give the total failure rate.

8.8 Multi-parameter Models

The beta-factor model has only one parameter, β, in addition to the component failure rate λ and is said to be a single-parameter model. This section presents briefly four CCF models with more than one parameter. These are sometimes called multiparameter models. The four models described are the following:

- Binomial failure rate (BFR) model
- Multiple Greek letter (MGL) model
- Alpha-factor model
- Multiple beta-factor (MBF) model

8.8.1 Binomial Failure Rate Model

The BFR model was introduced by Vesely (1977) based on the assumptions listed in Section 8.5.4. The BFR model is based on the premise that CCFs result from *shocks* to the voted group (Evans et al. 1984). The shocks occur randomly according to a HPP with rate ν. Whenever a shock occurs, each of the individual components is assumed to fail with probability p, independent of the states of the other components. The number Z of components failing as a consequence of the shock is thus binomially distributed (n, p). The probability that the multiplicity, Z, of failures due to a shock is equal to z is

$$\Pr(Z = z) = \binom{n}{z} p^z (1-p)^{n-z} \quad \text{for } z = 0, 1, \dots, n. \tag{8.29}$$

The mean number of components that fail in one shock is $E(Z) = np$. The following two conditions are assumed:

- Shocks and independent failures occur independently of each other.
- All failures are immediately discovered and repaired, with the repair time being negligible.

As a consequence, the time between independent component failures, in the absence of shocks, is exponentially distributed with failure rate $\lambda^{(i)}$, and the time between shocks is exponentially distributed with rate ν. The number of independent failures in any time period of length t is therefore Poisson distributed with parameter $\lambda^{(i)}t$, and the number of shocks in any time period of length t is Poisson distributed νt.

The component failure rate caused by shocks thus equals $p\nu$, and the total failure rate of one component equals

$$\lambda = \lambda^{(i)} + p\nu. \tag{8.30}$$

By using this model, we have to estimate the independent failure rate $\lambda^{(i)}$ and the two parameters ν and p. The parameter ν relates to the degree of "stress" on the group, whereas p is a function of the built-in component protection against external shocks. Observe that the BFR model is identical to the beta-factor model when the group has only two components.

The assumption that the components fail independently when a shock occurs is a rather serious limitation, and this assumption is often not satisfied in practice. The problem can, to some extent, be remedied by defining one fraction of the shocks as being "lethal" shocks, that is shocks that automatically cause all the components to fail, that is $p = 1$. If all the shocks are "lethal," one is back to the beta-factor model. Observe that this case, $p = 1$, corresponds to the situation in which there is no built-in protection against these shocks.

Situations where independent failures occur together with nonlethal as well as lethal shocks are often realistic. Such models are rather complicated, even when the nonlethal and the lethal shocks occur independently of each other.

Example 8.9 (2oo3:G structure of identical components)

Consider a 2oo3:G structure of identical components with individual failure $\lambda^{(i)} = 5.0 \times 10^{-6}$ h^{-1}.

The voted group is exposed to random shocks that occur according to a HPP with rate ν that has been estimated to be $\nu = 1.0 \times 10^{-5}\ \mathrm{h}^{-1}$. When a shock occurs, each component fails with probability $p = 0.20$. The components are assumed to fail independently when a shock occurs such that the number Z of components that fail due to a shock is binomially distributed.

$$\begin{aligned}
\Pr(Z = 0) &= \binom{3}{0} p^0(1-p)^{3-0} = (1-p)^3 = 0.5120\\
\Pr(Z = 1) &= \binom{3}{1} p^1(1-p)^{3-1} = 3p(1-p)^2 = 0.3840\\
\Pr(Z = 2) &= \binom{3}{2} p^2(1-p)^{3-2} = 3p^2(1-p) = 0.0960\\
\Pr(Z = 3) &= \binom{3}{3} p^3(1-p)^0 = p^3 = 0.0080.
\end{aligned}$$

The voted group only fails due to shocks when $Z = 2$ or $Z = 3$. This means that the probability that a shock results in a group failure is $p_s = \Pr(Z = 2) + \Pr(Z = 3) = 0.1040$. Random shocks giving group failures therefore occur according to a HPP with rate $\nu_s = \nu p_s = 1.04 \times 10^{-6}\ \mathrm{h}^{-1}$. □

Observe that shocks can occur without any failures (i.e. $Z = 0$). This makes it difficult to estimate ν directly from failure data because shocks may occur unnoticed when no component fails.

8.8.2 Multiple Greek Letter Model

The *MGL* model represents a generalization of the beta-factor model to obtain an approach that is less conservative for highly redundant structures (e.g. see Fleming and Kalinowski 1983; NUREG/CR-4780 1989). The assumptions made for the MGL model are the same as listed in Section 8.5.4.

Assume that a voted group of n identical components are proof-tested at time t. We may assume that potential faults are hidden faults. Let E_1 be the event that a specified component, say component 1, is found to be failed and let Z be the multiplicity of the fault. New parameters (Greek letter) are defined as follows:

$\beta = \Pr[Z \geq 2 \mid E_1 \cap (Z \geq 1)] = \Pr[Z \geq 2 \mid E_1]$
$\gamma = \Pr[Z \geq 3 \mid E_1 \cap (Z \geq 2)]$
$\delta = \Pr[Z \geq 4 \mid E_1 \cap (Z \geq 3)]$
Additional Greek letters are introduced for higher multiplicities of failures.

Expressed verbally:

- If we have detected one fault in the group, then β is the probability that there are at least one more fault.

- If we have detected two faults, then γ is the probability that there are at least one more fault.
- If we have detected three faults, then δ is the probability that there are at least one more fault.

The extra parameters account for

- Higher component redundancies
- Fault multiplicities greater than one and less than n.

Observe that the beta-factor model is a special case of the MGL model when $n = 2$, and also when all the parameters of the model, except for β, are equal to 1.

In the MGL model, the probabilities $Q_{k:n}$ are expressed in terms of the total component failure probability, Q, which includes effects of all (independent and CCF) contributions to that component failure, and a set of conditional probabilities of all possible ways a CCF of component can be shared with other components in the same group, given that component failure has occurred.

We do not go into the details of the MGL model, but suffice by illustrating the approach by studying a voted group of $n = 3$ identical components. Further details may be found in NUREG/CR-4780 (1989).

System with Three Identical Components

Consider a structure of $n = 3$ identical components. The probabilities $g_{k,3}$ of the various multiplicities of failures are shown in Figure 8.11 for $k = 1, 2, 3$. Without loss of generality, we may consider component 1, such that event E_1 denotes that a fault of component 1 is detected. The probability of event E_1 is

$$Q = \Pr(E_1) = g_{1,3} + 2g_{2,3} + g_{3,3}, \tag{8.31}$$

as shown in Figure 8.11. All three components have the same probability of failure.

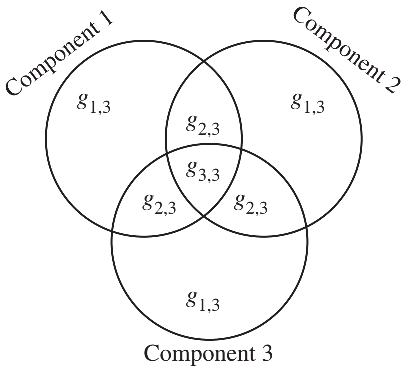


Figure 8.11 Probabilities of failures with different multiplicities.

As a first step of the development, we restrict our attention to the circle representing component 1 in Figure 8.11. The parameter β can be expressed as

$$\beta = \Pr(Z \geq 2 \mid Z \geq 1) = \frac{\Pr(Z \geq 2)}{\Pr(Z \geq 1)} = \frac{2g_{2,3} + g_{3,3}}{Q}. \tag{8.32}$$

For a structure of three components, $Z \geq 3$ is identical to $Z = 3$, and the parameter γ can be expressed as

$$\gamma = \Pr(Z = 3 \mid Z \geq 2) = \frac{\Pr(Z = 3)}{\Pr(Z \geq 2)} = \frac{g_{3,3}}{2g_{2,3} + g_{3,3}}. \tag{8.33}$$

Combining (8.32) and (8.33) yields

$$g_{3,3} = \beta\gamma Q. \tag{8.34}$$

Entering this result into (8.32) yields

$$g_{2,3} = \frac{1}{2}\beta(1-\gamma)Q. \tag{8.35}$$

The probability that E_1 is an individual (single) fault is

$$g_{1,3} = Q - 2g_{2,3} - G_{3,3} = Q[1 - \beta(1-\gamma) - \beta\gamma] = (1-\beta)Q. \tag{8.36}$$

The probability that the structure of three components has a single, double, and triple faults is

$$\begin{aligned} Q_{1:3} &= \binom{3}{1} g_{1,3} = 3(1-\beta)Q \\ Q_{2:3} &= \binom{3}{2} g_{2,3} = \frac{3}{2}\beta(1-\gamma)Q \\ Q_{3:3} &= \binom{3}{3} g_{3,3} = \beta\gamma Q. \end{aligned} \tag{8.37}$$

The probability that a 2oo3:F structure fails is therefore

$$Q_{2\text{oo}3:\text{F}} = Q_{2:3} + Q_{3:3} = [3\beta(1-\gamma) + \beta\gamma]Q = [3\beta - 2\beta\gamma]Q. \tag{8.38}$$

8.8.3 Alpha-Factor Model

The *alpha-factor model* is described by Mosleh and Siu (1987) for structures of n identical components. Assume that a failure event is observed at time t, be it an individual or a multiple failure. Let Q_{tot} be the probability of this event. With the notation introduced in Section 8.6, the probability is

$$Q_{\text{tot}} = 3g_{1,3} + 3g_{2,3} + g_{3,3}.$$

The alpha-factor model is based on a sequence of n parameters, $\alpha_1, \alpha_2, \dots, \alpha_n$, defined as

$$\begin{aligned} \alpha_k = {} & \Pr(\text{Exactly } k \text{ components fail} \mid \text{Failure event occurs}) \\ & \quad \text{for } k = 1, 2, \dots, n. \end{aligned} \tag{8.39}$$

This implies that $\sum_{k=1}^{n} \alpha_k = 1$. Observe that the alpha-factor model reduces to beta-factor model when $\alpha_1 = 1 - \beta$, $\alpha_n = \beta$, and $\alpha_i = 0$ for all $i = 2, 3, \dots, n-1$.

The formulas for the alpha-factor model depend on the type of testing performed, whether it is simultaneous testing and staggered testing. Interested readers may find the appropriate formulas in Mosleh and Siu (1987).

The alpha-factor model is recommended for CCF analysis in aerospace applications (NASA 2011) and is also recommended by the US NRC for nuclear applications.

Structure with Three Identical Components

We illustrate the alpha-factor model by a structure of $n = 3$ identical components. The starting point for the model is a failure event E in the structure. The failure can be an individual failure or a multiple fault of any (possible) multiplicity. In our case, with only three components, the probability of E is, from the results in Section 8.6

$$\Pr(E) = Q_{\text{tot}} = 3Q_{1:3} + 3Q_{2:3} + Q_{3:3},$$

a result that is easily seen from Figure 8.11. From the same figure, it is seen that

$$\begin{aligned} \alpha_1 &= \frac{3Q_{1:3}}{Q_{\text{tot}}} &\Rightarrow\quad Q_{1:3} &= \frac{\alpha_1}{3} Q_{\text{tot}} \\ \alpha_2 &= \frac{3Q_{2:3}}{Q_{\text{tot}}} &\Rightarrow\quad Q_{2:3} &= \frac{\alpha_2}{3} Q_{\text{tot}} \\ \alpha_3 &= \frac{Q_{3:3}}{Q_{\text{tot}}} &\Rightarrow\quad Q_{1:3} &= \alpha_3 Q_{\text{tot}}. \end{aligned} \tag{8.40}$$

Observe that Q_{tot} has to be estimated separately and is not a result from using the alpha-factor model. The alpha-factor model gives only the distribution of the failure multiplicities when a failure event E occurs.

The probability that a 2oo3:F structure fails is

$$Q_{2\text{oo}3:\text{F}} = Q_{2:3} + Q_{3:3} = \left[\frac{\alpha_2}{3} + \alpha_3\right] Q_{\text{tot}}.$$

8.8.4 Multiple Beta-Factor Model

The *MBF* model is developed by the research organization SINTEF and is thoroughly described by Hokstad and Rausand (2008). The MBF model is developed for application to safety-instrumented systems and is similar to the MGL model.

To illustrate the MBF model, consider a 2oo3:G structure of three identical components. As shown earlier, this structure fails with probability

$$Q_{2\text{oo}3} = 3g_{2,3} + g_{3,3}.$$

In line with the MGL model (with $\gamma = \beta_2$), this probability can be written as

$$Q_{2oo3} = (3 - 2\beta_2)\beta Q. \tag{8.41}$$

The factor in front of βQ is in the MBF model considered as a correction factor, C_{2oo3} to give

$$Q_{2oo3} = C_{2oo3}\beta Q. \tag{8.42}$$

The same approach is used for all $k$oo$n$:G configurations and suggested values for the correction factors C_{koon} for all relevant values of k and n are provided in Hokstad and Rausand (2008). Several SINTEF reports (e.g. Hauge et al. 2013, 2015) are available, providing theoretical background for the model and practical help.

Hokstad and Rausand (2008) give a more thorough survey of CCF models, including some additional models, and also presents ideas on how to estimate the parameters of the models. See also NASA (2011, chapter 10).

8.9 Problems

8.1 Let E_i denote that component i is functioning and let E_i^* denote that component i is failed, for $i = 1, 2$. Assume that the events E_1 and E_2 are independent. Show that this implies that events E_1 and E_2^* are also independent.

8.2 Consider the two events A and B that both have positive probabilities. Show that if $\Pr(A \mid B) = \Pr(A)$ then $\Pr(B \mid A) = \Pr(B)$.

8.3 A coin is tossed three times. Determine the probability of getting exactly two heads, when it is given that

(a) The first outcome was a head.
(b) The first outcome was a tail.
(c) The two first outcomes were heads.

8.4 If the occurrence of event A makes event B more likely, does the occurrence of event B make event A more likely? Justify your answer.

8.5 If $\Pr(A^*) = 0.35$ and $\Pr(B \mid A) = 0.55$, what is $\Pr(A \cap B)$?

8.6 Discuss basic CCF concepts.

(a) Carefully review the definition of a CCF in Definition 8.2. What type of criticism can be raised against this definition? Do you have any suggestions for improvements?

(b) What is a root cause and a coupling factor, and why are these two terms useful when explaining why CCFs occur?

(c) It is possible to argue that a CCF is sometimes a systematic failure and sometimes a random failure? Why is this the case?

8.7 Consider a 1oo4:G, a 2oo4:G, and a 3oo4:G structure.

(a) Compare and discuss how vulnerable with respect to CCFs and spurious activations the three structures are.

(b) Assume that we decide to use a beta-factor model to include CCFs. When determining the factor β, we usually end up using the same β value for all the three structures. Discuss the realism of this, and also the realism of the beta-factor model.

8.8 Consider a 2oo3:G structure of identical components with constant failure rate λ. The system is exposed to common-cause failures that may be modeled by a beta-factor model. In Figure 8.9, it is shown that the MTTF of the system has a minimum for $\beta = 0$. Determine the value of β for which MTTF attain its maximum. Explain why MTTF as a function of β has this particular shape.

8.9 Consider a bridge structure of five components. Assume that all the five components are identical and have constant failure rate λ. The system is exposed to common-cause failures that may be modeled by a beta-factor model. Determine the MTTF of the bridge structure as a function of β, and make a sketch of MTTF as a function of β when $\lambda = 5 \times 10^{-4}$ failures/h, and no repair is carried out.

8.10 *C*-Factor model.

(a) Describe and discuss the main differences between the beta-factor model and the *C*-factor model.

(b) In some cases, it may be argued that the *C*-factor model is more realistic than the beta-factor model. Why is this the case?

8.11 Consider a 2oo3:G structure of identical components. The system is exposed to common cause failures that may be modeled by a binomial failure rate (BFR) model. The "individual" failure rate of the components is $\lambda^{(i)} = 5 \times 10^{-5}$ failures/h. Nonlethal shocks occur with frequency $\nu = 10^{-5}$ non-lethal shocks/h. When a nonlethal shock occurs, the components may fail independently with probability $p = 0.20$. Lethal shocks occur with frequency $\omega = 10^{-7}$ lethal shocks/h. When a lethal shock occurs, all the three components will fail simultaneously. The lethal and the nonlethal shocks are assumed to be independent.

(a) Determine the mean time between system failures, $\text{MTBF}^{(i)}$, caused by "individual" component failures, when you assume that the system is only repaired when a system failure occurs. In such a case the system is repaired to an as-good-as-new condition.

(b) Determine the mean time between system failures, MTBF_{NL} when you assume that the only cause of system failures is the nonlethal shocks.

(c) Determine the mean time between system failures, MTBF_{L}, when you assume that the only cause of system failures is the lethal shocks.

(d) Try to find the total mean time between system failures. Discuss the problems you meet during this assessment.

References

Childs, J.A. and Mosleh, A. (1999). A modified FMEA tool for use in identifying and assessing common cause failure risk in industry. *Proceedings Annual Reliability and Maintainability Symposium*.

DOE-STD-1195 (2011). *Design of safety significant safety instrumented systems used at DOE nonreactor nuclear facilities*. Washington, DC: U.S. Department of Energy.

Evans, M.G.K., Parry, G.W., and Wreathall, J. (1984). On the treatment of common-cause failures in system analysis. *Reliability Engineering* 9: 107–115.

Fleming, K.N. (1975). A Reliability Model for Common Mode Failures in Redundant Safety Systems. *Tech. Rep. GA-A13284*. San Diego, CA: General Atomic Company.

Fleming, K.N. and Kalinowski, A.M. (1983). An Extension of the Beta Factor Method to Systems with High Levels of Redundancy. *Technical Report PLG-0289*. Pickard, Lowe, and Garrick Inc.

Fleming, K.N., Mosleh, A., and Kelley, A.P. (1983). Analysis of dependent failures in risk assessment and reliability evaluation. *Nuclear Safety* 24 (5): 637–657.

Hauge, S., Hoem, Å.S., Hokstad, P.R. et al. (2015). Common Cause Failures in Safety Instrumented Systems. *Report A26922*. Trondheim, Norway: SINTEF.

Hauge, S., Kråknes, T., Håbrekke, S., and Jin, H. (2013). Reliability prediction methods for safety instrumented systems, PDS method handbook, *Handbook*. Trondheim, Norway: SINTEF.

Hauge, S., Onshus, T., Øien, K. et al. (2006). Independence of Safety Systems on Offshore Oil and Gas Installations – Status and Challenges (in Norwegian). *STF50 A06011*. Trondheim, Norway: SINTEF.

Hokstad, P.R. and Rausand, M. (2008). Common cause failure modeling: status and trends. In: *Handbook of Performability Engineering*, Chapter 39 (ed. K.B. Misra), 621–640. London: Springer.

IEC 61508 (2010). *Functional safety of electrical/electronic/programmable electronic safety-related systems*. Parts 1-7, *International standard*. Geneva: International Electrotechnical Commission.

Little, R.G. (2012). Managing the risk of aging infrastructure, *Project report*. International Risk Governance Council (IRGC). https://irgc.org (accessed 25 March 2020).

Miller, A.G., Kaufer, B., and Carlson, L. (2000). Activities on component reliability under the OECD Nuclear Energy Agency. *Nuclear Engineering and Design* 198: 325–334.

Mosleh, A. and Siu, N.O. (1987). A multi-parameter, event-based common cause failure model. *Transactions of the 9th International Conference Structural Mechanics in Reactor Technology*, Lausanne, Switzerland.

NASA (2011). Probabilistic Risk Assessment Procedures: Guide for NASA Managers and Practitioners, *Guide NASA/SP-2011-3421*. Washington, DC: U.S. National Aeronautics and Space Administration.

NEA (2004). International Common-Cause Failure Data Exchange. ICDE general coding guidelines, *Technical report R(2004)4*. Paris: Nuclear Energy Agency.

NUREG/CR-4780 (1989). Procedures for Treating Common-Cause Failures in Safety and Reliability Studies, Volume 2, Analytical Background and Techniques. *Report NUREG/CR-4780*. Washington, DC: U.S. Nuclear Regulatory Commission.

NUREG/CR-5485 (1998). Guidelines on Modeling Common-Cause Failures in Probabilistic Risk Assessment, *Guideline NUREG/CR-5485*. Washington, DC: U.S. Nuclear Regulatory Commission.

O'Connor, A.N. (2013). A general cause based methodology for analysis of dependent failures in system risk and reliability assessments. PhD thesis. College Park, ML: University of Maryland.

Perrow, C. (1984). *Normal Accidents: Living With High-Risk Technologies*. New York: Basic Books, Inc.

Sun, K., Hou, Y., Sun, W., and Qi, J. (2019). *Power System Control Under Cascading Failures*. Hoboken, NJ: Wiley and IEEE Press.

Vesely, W.E. (1977). Estimating common cause failure probabilities in reliability and risk analyses: Marshall-Olkin specializations. In: *Nuclear Systems Reliability Engineering and Risk Assessment* (ed. J.B. Fussell and G.R. Burdick), 314–341. Philadelphia, PA: SIAM.

Zhang, N., Fouladirad, M., and Barros, A. (2017). Maintenance analysis of a two-component load-sharing system. *Reliability Engineering and System Safety* 167: 67–74.

Zhang, N., Fouladirad, M., and Barros, A. (2018a). Evaluation of the warranty cost of a product with type III stochastic dependence between components. *Applied Mathematical Modelling* 59: 39–53.

Zhang, N., Fouladirad, M., and Barros, A. (2018b). Optimal imperfect maintenance cost analysis of a two-component system with failure interactions. *Reliability Engineering and System Safety* 177: 24–34.

Zhang, N., Fouladirad, M., and Barros, A. (2018c). Warranty analysis of a two-component system with type I stochastic dependence. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of risk and reliability* 232 (3): 274–283.

9

Maintenance and Maintenance Strategies

9.1 Introduction

Many technical systems need to be maintained to preserve high operational reliability during their useful life. The influence of maintenance is clearly seen in Chapter 6 where maintenance metrics, such as the mean downtime (MDT) and the mean time to repair (MTTR), enter directly into the formulas that determine the reliability metrics. Other maintenance aspects do not have such a direct and visible effect, but may strongly affect failure rates and other system reliability metrics.

This chapter deals with aspects of maintenance that influence the operational reliability but does not provide a general introduction to maintenance. Two popular maintenance strategies, reliability-centered maintenance (RCM) and total productive maintenance (TPM), are presented. Management and economic aspects of maintenance are not covered adequately in this book.

First, we need to introduce some new terms. We consider a system where at least some of the items are subjected to some sort of maintenance. The smallest items that are maintained or replaced as a unit (i.e. without being disassembled on site) are called *maintainable items*. These items are the lowest level in the system hierarchy to which a *maintenance task* is allocated. A maintainable item is also called a *least replaceable assembly/unit*. What is defined to be a maintainable item may vary from company to company and between different application areas. A maintainable item is always a part of a specific system, which we refer to as the study object.

The persons who plan, execute, and document the maintenance tasks are referred to as the *maintenance personnel* or the maintenance crew. A computerized maintenance management system (CMMS) is often used to support and document the maintenance management.

System Reliability Theory: Models, Statistical Methods, and Applications, Third Edition.
Marvin Rausand, Anne Barros, and Arnljot Høyland.

Companion Website: www.wiley.com/go/SystemReliabilityTheory3e

In many industries, maintenance tasks are specified as maintenance *work orders*. A maintenance work order is a written request that lists the maintenance work to be carried out at a specified date and time. The work order

- Gives an explanation of the work to be carried out.
- Provides the maintenance personnel with detailed instructions on the work to be performed and which tools and other resources to use.
- Documents the labor, materials, and resources used to complete the work.
- Tracks all maintenance and repair work that has been performed on each maintainable item.

Most CMMSs have a module for making and managing maintenance work orders.

9.1.1 What is Maintenance?

The term "maintenance" is defined in Chapter 1 as: "the combination of all technical and management tasks intended to retain an item in, or restore it to, a state in which it can perform as required" (IEV 192-06-01).

Several international standards define and outline the main aspects of maintenance. Such general maintenance standards include

- EN 13306 (2017) "Maintenance terminology"
- ISO 55000 (2014) "Asset management – Overview, principles and terminology"
- IEC 60300-3-14 (2004) "Dependability management – Application Guides – Maintenance and maintenance support"
- ISO 17359 (2018) "Condition monitoring and diagnostics of machines – General guidelines"

A high number of standards covering maintenance of specific systems are also available.

In this book, maintenance is related to the required performance of item functions. Maintenance carried out for other purposes, such as to preserve aesthetic appearance, is hence not covered. Relevant objectives of maintenance include

- Prevent breakdowns
- Reduce downtime
- Reduce total costs
- Improve safety
- Improve equipment efficiency
- Improve production
- Reduce energy use
- Prevent/reduce pollution of the environment
- Extend useful life of equipment

9.2 Maintainability

The term "maintainability" was defined in Chapter 1 as a characteristic of an item's design and installation that determines the ease and rapidity maintenance tasks can be accomplished using prescribed practices and procedures. The maintainability of an item may be different for different failure modes. Some authors distinguish between two types of maintainability: serviceability and repairability.

The mean number of hours a study object is inoperative while undergoing maintenance is a determining factor for the operational reliability of the study object and much efforts are therefore devoted to getting an optimal maintainability.

Maintainability engineering is concerned with designing an item (system or device) such that all maintenance tasks can be performed easily, rapidly, and at low cost. Maintainability is a function of the equipment design and installation, personnel availability with the required skill levels, adequacy of maintenance procedures, test equipment, spare parts, and the physical environment under which the maintenance task is performed. Several standards and guidelines are developed to support maintainability engineering, and examples include (IEC 60706 2006; SAE JA1010 2011; MIL-HDBK-470A 1997). The maintainability may be influenced by factors such as (e.g. see IEC TR 62278-3 2010):

- The simplicity of the design and the use of standardized and interchangeable items and modules.
- The accessibility of items for servicing and repair (e.g. without having to build a scaffolding structure).
- The skills required to perform the work.
- The availability and quality of diagnostics to identify and isolate faults.
- The modularization and stacking such that modules with high failure rate are easy to access.
- The standardization and availability of tools.
- The redundancy of failure-prone items and the feasibility of switchovers.
- The maintenance documentation and its availability and completeness.
- The availability and quality of spare parts.
- The software code quality (i.e. to what degree it is developed, documented, and maintained according to accepted software quality principles).
- The accessibility for cleaning and testing.

High maintainability generally improves the operational reliability of the system, but for some types of systems, high maintainability may also have negative effects and lead to increased system failure rate. This is, for example the case when splitting the system into smaller and more handy modules, requiring several more connectors that may fail.

Example 9.1 (Subsea oil/gas production system)

A subsea oil and gas production system consists of a high number of control and safety valves, sensors and electrical and hydraulic control systems. Modern systems may also have pump, compressors, and several other types of equipment. The production system is located on the seabed, often more than 2000m below the sea surface. The largest systems might fill almost a whole football stadium and are split into a high number of modules, which must be pulled/lifted to the surface for maintenance/repair. Each module is connected to the main structure and/or to other modules by a number of remotely operated electrical, hydraulic, and flow connectors. Many modules are stacked on the top of each other. To avoid having to pull modules because an item beneath the module has failed, it is very important to locate the most failure-prone modules on the top level. The process of arranging the modules to avoid having to pull well-functioning modules is called *stacking*. For large systems, stacking is a very laborious and demanding process that requires many detailed reliability analyses. □

Maintainability engineering provides predictions of how long it will take to repair the maintainable item when it fails. Using these predictions, the design can be analyzed to identify possible changes that would reduce the time required to perform maintenance. Several *maintainability metrics* are used. Among these are

- Mean time to repair (MTTR)
- Mean/median active repair time versus mean item downtime
- Mean system downtime
- Maximum active corrective maintenance time
- Mean preventive maintenance time
- Mean man-hours per repair task
- Maintenance hours per operating hours
- Mean time to restore system

Maintainability metrics are probabilistic and are determined in a similar way as other reliability metrics. The actual maintainability of the system is usually determined by *maintainability demonstration* of typical maintenance tasks.

9.3 Maintenance Categories

Maintenance tasks may be classified in many different ways. The classification in Figure 9.1 is in line with many standards (e.g. see EN 13306 2017). The main types of maintenance tasks are briefly described in the following:

(1) *Corrective maintenance* (CM) denotes all tasks that are carried out as a result of a detected item failure or fault, to restore the item to a specified condition. The goal of a CM task is to bring the item back to a functioning state as soon as possible, either by repairing or replacing the failed item. CM tasks

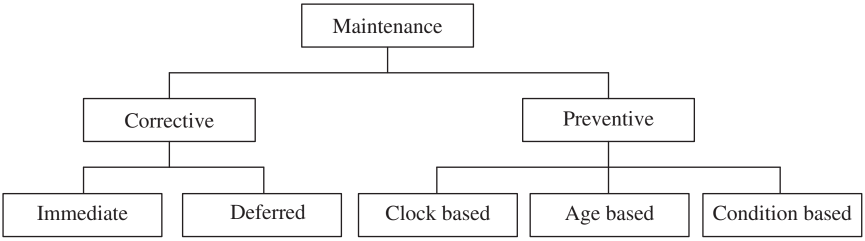


Figure 9.1 Classification of maintenance types.

may be carried out immediately when a fault is revealed or postponed until an opportunity occurs. In the last case, the CM task is said to be *deferred*. CM is also called *repair*, *reactive maintenance*, *run-to-failure maintenance*, or *breakdown maintenance* and can include any or all of the following steps: localization, isolation, disassembly, interchange, reassembly, alignment, and checkout (e.g. see MIL-HDBK-338B 1998). An argument for CM may be summarized by the old quotation: "if it ain't broke, don't fix it."

(2) *Preventive maintenance* (PM) is planned maintenance "carried out to mitigate degradation and reduce the probability of failure" (IEV 192-06-05). PM tasks may involve inspection, adjustments, lubrication, parts replacement, calibration, and repair of items that are beginning to wear out. PM tasks have traditionally been performed on a regular basis, regardless of whether or not functionality or performance is degraded. With the current possibility of massive data collection, many companies move to degradation-based PM tasks. There are several types of PM tasks:

  (a) *Age-based PM tasks* are carried out at a specified age of the item. The age may be measured as time in operation, or by other time concepts, such as number of kilometers driven for an automobile, or number of take-offs/landings for an aircraft. The *age replacement* policy discussed in Section 12.3 is an example of age-based maintenance.

  (b) *Clock-based PM tasks* are carried out at specified calendar times. The *block replacement* policy discussed in Section 12.3 is an example of clock-based maintenance. A clock-based maintenance policy is generally easier to manage than an age-based maintenance policy because the maintenance tasks can be scheduled to predefined times.

  (c) *Condition-based PM tasks* are based on measurements of one or more condition variables of the item. A PM task is initiated when a condition variable (or a function of several condition variables) approaches, or passes a threshold value. Examples of condition variables include vibration, temperature, and number of particles in the lubrication oil. The condition variables may be monitored continuously or at regular intervals.

  (d) *Opportunity-based tasks* are carried out when maintenance tasks of other items or a system shutdown provides an opportunity for carrying out maintenance on items that were not the cause of the opportunity. Opportunity maintenance is not listed as a separate type in some maintenance standards.

  (e) *Overhaul* is a comprehensive set of PM tasks that are carried out to maintain the level of a system's performance. Very often, the overhaul is allocated to periods with low demand for the system services. In the offshore oil and gas industry, whole platforms may be closed down for several weeks during summer to overhaul the equipment and optimize functions, to secure that the production will run smoothly during the high-demand period.

(3) *Predictive maintenance* extends condition-based maintenance by adding theory and methods used to predict the time when the item will fail. A PM task can then be planned for a suitable time before the maintainable item fails.

Maintenance tasks may be classified in many different ways as illustrated by the German (DIN 31051 2012), see box.

An alternative classification of maintenance

The German standard (DIN31051) distinguishes between four categories of maintenance.

Servicing – tasks to reduce the wear, such as lubrication, cleaning, adjustments, and calibration.

Inspection –tasks to determine and assess the actual state of the item, including the causes of this state and necessary consequences for further usage.

Repair – tasks to restore the function of a failed item. Also called corrective maintenance.

Improvement – tasks to improve the reliability and/or the maintainability of the item without changing its original function. This category includes replacement of worn parts.

Example 9.2 (Automobile service)

Automobile service consists of a set of maintenance tasks carried out at a specified time or after the automobile has run a certain distance. The service intervals and the content of each service are specified by the automobile manufacturer. Some modern automobiles display the time of the next service on the instrument panel and adjust the service-time based on additional usage parameters, such as number of starts. The service may include tasks such as, replace engine oil, replace filters, check/refill brake fluid, grease and lubricate components, check lights and wipers, and many more. □

Example 9.3 (Proof test)
Consider an automatic safety shutdown valve that is located on a pipeline in a process plant. The valve is normally in open position and is only used to shut down the flow through the pipeline if a safety-critical situation should occur. Because the valve is seldom closed, it may have hidden critical faults that prevent it from performing its safety function. Of this reason, the valve is proof-tested at regular intervals. The proof test covers more than testing that the valve is able to close. The valve is tested with pressure and flow-rate as close to the real shutdown situation as possible. □

Remark 9.1 (Modification)
Modification is an integrated set of tasks carried out to modify, or change, one or more functions of a system. After a modification, the system does not perform exactly the same functions as before the modification. A modification, or change, is *not* classified as a maintenance task, but this is often performed by the maintenance personnel. □

9.3.1 Completeness of a Repair Task

When a maintainable item is repaired, the repair task may be

(1) *Perfect repair.* The repair task returns the item to an *as-good-as-new* condition, which corresponds to replacing the item with a new item of the same type.
(2) *Imperfect repair.* The repair task returns the item to a functioning state that is inferior to the state of a new item. In most cases, the return state is better than the state just before the failure occurred, but it may sometimes be even worse (e.g. when new faults are introduced in the repair task).
(3) *As-bad-as-old.* The repair task returns the item to the same state as it had just before the failure. This is, for example, the case for a large system where the repair task is to repair a small component and do nothing with the rest of the system.

The completeness of a repair task is discussed further in Section 10.5.

9.3.2 Condition Monitoring

Condition monitoring may be defined as follows:

Definition 9.1 (Condition monitoring)
The process of systematic data collection and evaluation to identify changes in performance or condition of a maintainable item or a system, such that a remedial task may be planned in a cost-effective manner to maintain reliability. □

Common techniques for condition monitoring of mechanical systems include (Fedele 2011):

- Visual inspection
- Performance monitoring
- Monitoring of noise and vibrations
- Monitoring of wear debris
- Monitoring of temperature

9.4 Maintenance Downtime

There are two types of system downtime associated with a maintenance task.

(1) *Unplanned downtime* is the downtime caused by item failures or internal and external (random) events; for example human errors, environmental impacts, loss of utility functions, labor conflicts (strikes), and sabotage. In some applications (e.g. electro-power generation), the unplanned downtime is called the *forced outage* time.

(2) *Planned downtime* is the downtime caused by planned preventive maintenance, planned operations (e.g. change of tools), and planned breaks, holidays, and the like. What is to be included as planned downtime depends on how the mission period is defined. We may, for example, define the mission period as one year (8760 hours), or the net planned time in operation during one year, excluding all holidays and breaks, and all planned operational stops. In some applications, it is common to split the planned downtime into two types:
    (a) *Scheduled downtime* that is planned long time in advance (e.g. planned preventive maintenance, breaks, and holidays)
    (b) *Unscheduled planned downtime* initiated by condition monitoring, detection of incipient failures, and events that may require a preventive task to improve or maintain the quality of the system functions, or to reduce the probability of a future failure. The associated remedial tasks can sometimes be postponed (within some limits) and carried out when it is suitable from an operational point of view.

The scheduled downtime may often be regarded as deterministic and be estimated from the operational plans. The unscheduled planned downtime may be subject to random variations, but it is usually rather straightforward to estimate a mean value.

The unplanned downtime strongly depends on the cause of the downtime. Assume that we have identified n independent causes of unplanned downtime, and let D_i be the random downtime associated to cause i for $i = 1, 2, \ldots, n$. Let $F_{D_i}(d)$ denote the distribution function of D_i, and let p_i be the probability that a specific downtime has cause i. The distribution of the downtime D is then $F_D(d) = \sum_{i=1}^{n} p_i F_{D_i}(d)$, and the mean downtime is

$$\text{MDT} \approx \sum_{i=1}^{n} p_i \text{MDT}_i,$$

where $\text{MDT}_i = E(D_i)$ denotes the mean downtime associated with cause i for $i = 1, 2, \ldots, n$.

9.4.1 Downtime Caused by Failures

In the following, we confine ourselves to discussing the downtime caused by item failures and assume that the planned downtime and the unplanned downtime from other causes are treated separately. When we use the term "downtime" in the following, we tacitly assume that the downtime is caused by item failures.

The downtime of an item can usually be regarded as a sum of elements, such as access time, diagnosis time, active repair time, and checkout time. The elements are further discussed by Smith (2013). The length of the various elements are influenced by a number of system specific factors, such as ease of access, maintainability, and availability of maintenance personnel, tools, and spare parts. The downtime associated with a specific failure therefore has to be estimated based on knowledge of all these factors.

The MDT is the mean time the item is in a nonfunctioning state after a failure. The MDT is usually significantly longer than the MTTR and includes time to detect and diagnose the failure, logistic time, and time to test and startup of the item. When the item is put into operation again it is considered to be as-good-as-new. The mean uptime (MUT) of the item is equal to the MTTF. Both concepts may be used, but MUT is a more common term in maintenance applications. The mean time between consecutive occurrences of failures is denoted MTBF. The state variable and the various time concepts are illustrated in Figure 9.2.

For detailed reliability assessments, it is important to choose an adequate downtime distribution as basis for the estimation. Three distributions are commonly used: the exponential, the normal, and the lognormal distribution (Ebeling 2009). We briefly discuss the adequacy of these distributions.

Exponential Distribution

The exponential distribution is the most simple downtime distribution we can choose because it has only one parameter, the *repair rate* μ. The exponential distribution was discussed in detail in Section 5.4. Here, we briefly mention some of its main features.

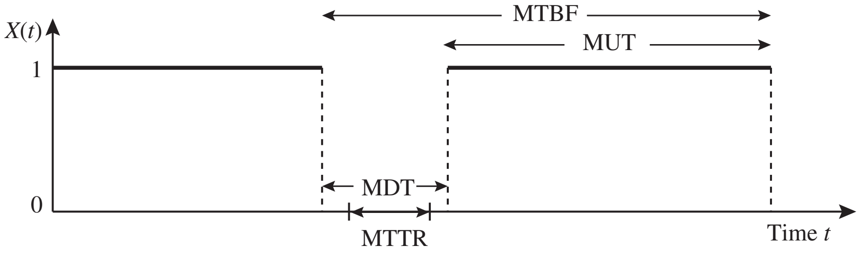


Figure 9.2 Average “behavior” of a repairable item and main time concepts.

The mean downtime is MDT $= 1/\mu$, and the probability that a downtime D is longer than a value d is $\Pr(D > d) = e^{-\mu d}$. The exponential distribution has no memory. This implies that if a downtime has lasted a time d, the mean *remaining* downtime is $1/\mu$ regardless of the value of d. This feature is not realistic for most downtimes, except for situations where the main part of the downtime is spent on searching for failures, and where failures are found more or less at random.

In many applications, the exponential distribution is chosen as a downtime distribution, not because it is realistic, but because it is easy to use.

Example 9.4 (Exponentially distributed downtime)
Consider a repairable item with downtime D related to a specific type of failures. The downtime is assumed to be exponentially distributed with repair rate μ. The MDT for this specific type of failures has been estimated to be five hours. The repair rate is then $\mu = 1/\text{MDT} = 0.20$ hours^{-1}. The probability that the downtime, D, is longer than seven hours is $\Pr(D > 7) = e^{-7\mu} \approx 0.247 = 24.7\%$. □

Normal Distribution

The rationale for choosing a normal (Gaussian) downtime distribution is motivated by the fact that the downtime may be considered as a sum of many independent elements. The normal distribution is discussed in Section 5.4. Estimation of MDT and the standard deviation are straightforward in the normal model. When using the normal distribution, the repair rate function $\mu(d)$ may be approximated by a straight line as a function of the elapsed downtime d. Therefore, the probability of being able to complete an ongoing repair task within a next short interval increases almost linearly with time.

Lognormal Distribution

The lognormal distribution is often used as a model for the repair time distribution. The lognormal distribution is discussed in Section 5.4. When using the lognormal distribution, the repair rate $\mu(d)$ increases up to a maximum, and thereafter decreases asymptotically down to zero as a function of the elapsed downtime d.

When an item has been down for a very long time, this indicates serious problems, for example that there are no spare parts available on the site, or that the maintenance crew is not able to get access to, or correct the failure. It is therefore natural to believe that the repair rate is decreasing after a certain period of time.

9.4.2 Downtime of a Series Structure

Consider a series structure of *n independent* items. Item i has constant failure rate λ_i. When item i fails, the mean system downtime is MDT_i, for $i = 1, 2, \dots, n$. The probability that the structure failure is caused by item i is $\lambda_i / \sum_{j=1}^{n} \lambda_j$, and the mean system downtime for an unspecified failure is

$$\text{MDT} \approx \frac{\sum_{i=1}^{n} \lambda_i \text{MDT}_i}{\sum_{j=1}^{n} \lambda_j} \tag{9.1}$$

The MDT is equal to the right-hand side of (9.1) only when the items are not independent. In most applications, equation (9.1) gives a good approximation.

Example 9.5 (Item with independent failure modes)
Consider an item with n independent failure modes. Failure mode i occurs with constant failure rate λ_i, and the mean downtime required to restore the item from failure mode i is MDT_i for $i = 1, 2, \dots, n$. The item may be considered as a series structure of n independent virtual items, where item i only can fail with failure mode i. The mean downtime of the item is therefore given by (9.1). □

Equation (9.1) may be used as an approximation for the mean downtime caused by an unspecified item failure of a nonseries structure of independent items. In this case, it is important to realize that MDT_i denotes the *system* downtime caused by failure of item i for $i = 1, 2, \dots, n$.

9.4.3 Downtime of a Parallel Structure

Consider a parallel structure of *n independent* items. The structure fails when all the n items are in a failed state. The mean system downtime MDT_S can be very different depending on the repair strategy. Several options are possible and among these are

(1) We wait until the structure has failed before starting any maintenance task and then we may
   (a) repair all the items at the same time, or
   (b) repair the item with the lowest MDT_i, in which case $\text{MDT}_S = \min_{1=1,2,\dots,n}\{\text{MDT}_i\}$.

(2) We begin repairing the items as soon as they fail or as soon as a given number of them have failed. The structure experiences a downtime if the last surviving item fails before the maintenance of at least one of the failed items is finished.

The system downtime is a random variable and its distribution and mean value may be rather difficult to determine because they cannot be derived directly based on a description of a limited number of scenarios. The use of a stochastic process is required to describe the possible states of the system. This can be done in some cases by using Markov processes (see Chapter 11) if the times-to-failure and repair times are exponentially distributed.

9.4.4 Downtime of a General Structure

For more complicated structures, there are no generic analytical formulas available that can give the mean structure downtime MDT_S. The structure downtime strongly depends on the maintenance policy when a failure occurs (as indicated for the parallel structure). Monte Carlo simulation may be used to obtain adequate estimates.

9.5 Reliability Centered Maintenance

As many modern maintenance practices, the RCM concept originated within the aircraft industry. RCM has now been applied with considerable success for more than 40 years; first within the aircraft industry, and later within the military forces, the nuclear power industry, the offshore oil and gas industry, and many other industries. Experiences from these industries show significant reductions in PM costs while maintaining, or even improving, the availability of the systems.

Definition 9.2 (Reliability-centered maintenance, RCM)
A systematic consideration of system functions, the way functions can fail, and a priority-based consideration of safety and economics that identifies applicable and effective PM tasks[1] □

The focus of RCM is on the system *functions*, and not on the system hardware, and the main objective of RCM is to reduce the maintenance cost, by focusing on the most important functions of the system, and avoiding or removing maintenance tasks that are not strictly necessary. If a maintenance program already exists, the result of an RCM analysis will often be to eliminate inefficient PM tasks.

[1] The definition is based on a definition proposed by the Electric Power Research Institute (EPRI).

The RCM concept is described in several standards, reports, and textbooks. Among these are Nowlan and Heap (1978), IEC 60300-3-11 (2009), SAE JA1012 (2011), and NASA (2008). The main ideas presented in the various sources are more or less the same, but the detailed procedures may be rather different.

The maintenance tasks considered in the RCM approach are related to failures and functional degradation. Maintenance carried out, for example to preserve or improve the aesthetic appearance of a system by cleaning and painting is outside the scope of RCM, at least when such maintenance has no effect on the system functions. However, planning of such tasks should be integrated with the planning of RCM relevant tasks.

9.5.1 What is RCM?

RCM is a technique for developing a PM program. It is based on the assumption that the inherent reliability of the equipment is a function of the design and the built quality. An effective PM program will ensure that the inherent reliability is maintained. It should be realized that RCM will never be a substitute for poor design, inadequate build quality or bad maintenance practices. RCM cannot improve the inherent reliability, of the system. This is only possible through redesign or modification.

The application of PM is often misunderstood. It is easy to erroneously believe that the more an item is routinely maintained, the more reliable it will be. Often, the opposite is the case due to maintenance-induced failures. RCM was designed to balance the costs and benefits, to obtain the most cost-effective PM program. To achieve this, the desired system performance standards have to be specified. PM will not prevent all failures, and therefore the potential consequences of each failure must be identified and the likelihood of failure must be known. PM tasks are chosen to address each failure by using a set of applicability and effectiveness criteria. To be effective, a PM task must provide a reduced expected loss related to personnel injuries, environmental damage, production loss, and/or material damage.

An RCM analysis basically provides answers to the following seven questions.

(1) What are the functions and associated performance standards of the equipment in its present operating context?
(2) In what ways can it fail to fulfill its functions?
(3) What is the cause of each functional failure?
(4) What happens when each failure occurs?
(5) In what way does each failure matter?
(6) What can be done to prevent each failure?
(7) What should be done if a suitable preventive task cannot be found?

Experience has shown that approximately 30% of the efforts of an RCM analysis is involved in defining functions and performance standards, that is, answering question 1.

9.5.2 Main Steps of an RCM Analysis

The RCM analysis may be carried out as a sequence of activities or steps, some of which are overlapping in time.

(1) Study preparation
(2) System selection and definition
(3) Functional failure analysis (FFA)
(4) Critical item selection
(5) Data collection and analysis
(6) Failure modes, effects and criticality analysis (FMECA)
(7) Selection of maintenance tasks
(8) Determination of maintenance intervals
(9) Preventive maintenance comparison analysis
(10) Treatment of noncritical items
(11) Implementation
(12) In-service data collection and updating

The various steps are discussed in the following.

Step 1: Study Preparation
In Step 1 an RCM project group is established. The project group must define and clarify the objectives and the scope of the analysis. Requirements, policies, and acceptance criteria with respect to safety and environmental protection should be made visible as boundary conditions for the RCM analysis.

Overall drawings and process diagrams, such as piping and instrumentation diagrams, must be made available. Possible discrepancies between the as-built documentation and the real plant must be identified. The resources that are available for the analysis are usually limited. The RCM group should therefore be sober with respect to what to look into, realizing that analysis cost should not dominate potential benefits.

Step 2: System Selection and Definition
Before a decision to perform an RCM analysis at a plant is taken, two questions should be considered.

(1) To which systems are an RCM analysis beneficial compared with more traditional maintenance planning?

(2) At what level of assembly (plant, system, subsystem) should the analysis be conducted?

All systems may in principle benefit from an RCM analysis. With limited resources, we must, however, make priorities, at least when introducing RCM in a new plant. We should start with the systems we assume will benefit most from the analysis. Most operating plants have developed some sort of assembly hierarchy. In the offshore oil and gas industry, this hierarchy is referred to as a tag number system. The following terms will be used for the levels of the assembly hierarchy:

Plant is a set of systems that function together to provide some sort of output. An offshore gas production platform is, for example considered to be a plant.

System is a set of subsystems that perform a main function in the plant (e.g. generate electro-power, supply steam). The gas compression system on an offshore gas production platform may, for example be considered as a system. Observe that the compression system may consist of several compressors with a high degree of redundancy. Redundant items performing the same main function should be included in the same system.

The system level is recommended as the starting point for the RCM analysis. This means that on an offshore oil/gas platform the starting point of the analysis should, for example be the gas compression system, and not the whole platform.

The systems may be broken down into subsystems, sub-subsystems, and so on. For the purpose of the RCM analysis, the lowest level of the hierarchy is called *maintainable items*.

Maintainable item is an item that is able to perform at least one significant function as a stand-alone item (e.g. pumps, valves, and electric motors). By this definition, a shutdown valve is, for example a maintainable item, whereas the valve actuator is not. The actuator is a supporting equipment to the shutdown valve and only has a function as part of the valve. The importance of distinguishing the maintainable items from their supporting equipment is clearly seen in the FMECA in Step 6. If a maintainable item is found to have no significant failure modes, then none of the failure modes or causes of the supporting equipment are important, and therefore do not need to be addressed. Similarly, if a maintainable item has only one significant failure mode, then the supporting equipment only needs to be analyzed to determine if there are failure causes that may affect that particular failure mode. Therefore, only the failure modes and effects of the maintainable items need to be analyzed in the FMECA in Step 6.

By the RCM approach, all maintenance tasks and maintenance intervals are decided for the maintainable items. When it comes to the execution of a particular maintenance task on a maintainable item, this will usually involve repair, replacement, or testing of an item or part of the maintainable item. These components/parts are identified in the FMECA in Step 6. The RCM analyst should always try to keep the analysis at the highest practical indenture level. The lower the level, the more difficult it is to define performance standards.

It is important that the maintainable items are selected and defined in a clear and unambiguous way in this initial phase of the RCM analysis because the following steps of the analysis are based on these items.

Step 3: Functional Failure Analysis

A specific system was selected in Step 2. The objectives of this step are to

(i) Identify and describe the system's required functions and performance criteria
(ii) Describe input interfaces required for the system to operate
(iii) Identify the ways in which the system might fail to function

Step 3(i): Identification of System Functions

The system will usually have a high number of different functions. It is essential for the RCM analysis that all the important system functions are identified. The analyst may benefit from using the approach outlined in Chapter 4.

Step 3(ii): Identification of Interfaces

The various system functions may be represented by functional block diagrams, to illustrate the input interfaces to a function. In some cases, we may want to split system functions into subfunctions on an increasing level of detail, down to functions of maintainable items. This may be accomplished by functional block diagrams, or reliability block diagrams.

Step 3(iii): Functional Failures

The next step is a FFA to identify and describe the potential system failure modes. In most of the RCM references, the system failure modes are called *functional failures*. Classifications schemes for failure modes were discussed in Chapter 3. Such schemes may be used to secure that all relevant functional failures are identified.

The functional failures are recorded on a specific FFA worksheet that is rather similar to a standard FMECA worksheet. An example of an FFA worksheet is shown in Figure 9.3. In the first column of the worksheet, the various operational modes of the system are recorded. For each operational mode, all the relevant system functions are recorded in column 2. The performance requirements to each function such as target values and acceptable deviations are listed in column 3. For each system function (in column 2) all the relevant functional failures are listed in column 4. In columns 5–8, a criticality ranking of each functional failure in that particular operational mode is given. The reason for including the criticality ranking is to be able to limit the extent of the further analysis by not wasting time

System: Performed by:
Drawing no.: Date: Page: of

Operational mode	System function	Functional requirements	Functional failure	Criticality				Frequency
				S	E	A	C	

Figure 9.3 Functional failure analysis (FFA) worksheet.

on insignificant functional failures. For complicated systems, such a screening is often important in order not to waste time and money.

The criticality must be judged on the plant level and should be ranked in the four consequence classes:

S: Safety of personnel
E: Environmental impact
A: Production availability
M: Material loss

For each of these consequence classes, the criticality may be ranked as for example high (H), medium (M), low (L), or negligible (N), where the definition of the categories will depend on the specific application. If at least one of the four entries are medium (M) or high (H), the criticality of the functional failure should be classified as significant, and be subject to further analysis.

The frequency of the functional failure may also be classified into four categories. The frequency classes may be used to prioritize between the significant functional failures. If all the four criticality entries of a functional failure are low or negligible, and the frequency is also low, then the failure is classified as insignificant and disregarded in the further analysis.

Step 4: Critical Item Selection

The objective of this step is to identify the maintainable items that are potentially critical with respect to the functional failures identified in Step 3 (iii). These maintainable items are denoted *functional significant items* (FSI). Observe that some of the less critical functional failures are disregarded at this stage of the analysis.

For simple systems, the FSIs may be identified without any formal analysis. In many cases, it is obvious which maintainable items have influence on the system functions.

For complicated systems with an ample degree of redundancy or with buffers, we may need a formal approach to identify the FSIs. Depending on the complexity of the system, importance ranking based on techniques such as fault tree analysis, reliability block diagrams, or Monte Carlo simulation may be suitable. In a petroleum production plant, there is often a variety of buffers and rerouting possibilities. For such systems, Monte Carlo next event simulation may often be the only feasible approach.

In addition to the FSIs, we should identify items with high failure rate, high repair costs, low maintainability, long lead time for spare parts, and items requiring external maintenance personnel. These maintainable items are denoted *maintenance cost significant items* (MCSI). The combination of the FSIs and the MCSIs are denoted *maintenance significant items* (MSI).

In the FMECA in Step 6, each of the MSIs will be analyzed to identify potential failure modes and effects.

Step 5: Data Collection and Analysis

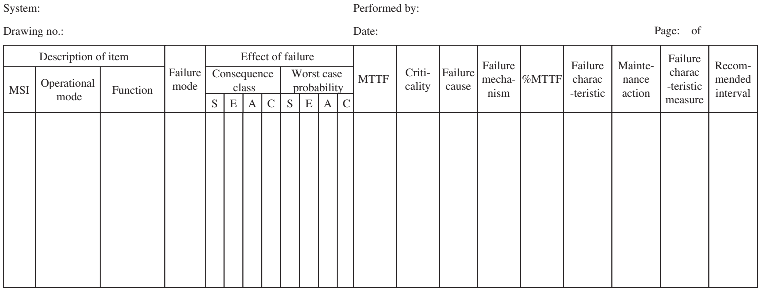
The various steps of the RCM analysis require a variety of input data, such as design data, operational data, and reliability data. Reliability data sources are discussed in Chapter 16. Reliability data is necessary in order to decide the criticality, to mathematically describe the failure process, and to optimize the time between PM tasks.

In some situations, there is a complete lack of reliability data. This is the case when developing a maintenance program for new systems. The maintenance program development starts long before the equipment enters service. Helpful sources of information may then experience data from similar equipment, recommendations from manufacturers, and expert judgments. The RCM method will even in this situation provide useful information.

Step 6: Failure Modes, Effects, and Criticality Analysis

The objective of this step is to identify the dominant failure modes of the MSIs identified in Step 4. A variety of different FMECA worksheets are proposed in the main RCM references. The FMECA worksheet used in our approach is presented in Figure 9.4, and is more detailed than most of the FMECA worksheets in the main RCM references. The various columns in our FMECA worksheet are as follows:

- *MSI*. In this column, we record the maintainable item number in the assembly hierarchy (tag number), optionally with a descriptive text.



System: Performed by:

Drawing no.: Date: Page: of

Description of item			Failure mode	Effect of failure								MTTF	Criticality	Failure cause	Failure mechanism	%MTTF	Failure characteristic	Maintenance action	Failure characteristic measure	Recommended interval
MSI	Operational mode	Function		Consequence class				Worst case probability												
				S	E	A	C	S	E	A	C									

Figure 9.4 RCM-FMECA worksheet.

- *Operational mode*. The MSI may have various operational modes, for example running and standby. The operational modes are listed, one by one.
- *Function*. The various functions for each operational mode of the MSI are listed.
- *Failure mode*. The failure modes for each function are listed.
  Effect of failure/severity class. The effect of a failure is described in terms of the *worst-case* outcome for the S, E, A, and C categories introduced in Step 3(iii). The criticality may be specified by the same four classes as in Step 3(iii), or by some numerical severity measure. A failure of an MSI will not necessarily give a worst-case outcome resulting from, redundancy, buffer capacities, and the like. Conditional likelihood columns are therefore introduced.
- *"Worst-case" probability*. The worst-case probability is defined as the probability that an equipment failure will give the worst-case outcome. To obtain a numerical probability measure, a system model is sometimes required. This will often be inappropriate at this stage of the analysis and a descriptive measure may be used.
- *MTTF*. Mean time-to-failure (MTTF) for each failure mode is recorded. Either a numerical measure or likelihood classes may be used.

The information described so far should be entered for all failure modes. A screening may now be appropriate, giving only dominant failure modes, that is, items with high criticality.

- *Criticality*. The criticality field is used to tag off the dominant failure modes according to some criticality measure. A criticality measure should take failure effect, worst-case probability, and MTTF into account. "Yes" is used to tag off the dominant failure modes.

For the dominant failure modes, the following fields are required:

- *Failure cause*. For each failure mode, there may be several failure causes. An MSI failure mode will typically be caused by one or more component failures. Observe that supporting equipment to the MSIs entered in the FMECA worksheet is for the first time considered in this step. In this context, a failure cause may therefore be a failure mode of a supporting equipment. A "fail to close" failure of a safety valve may, for example be caused by a broken spring in the failsafe actuator.
- *Failure mechanism*. For each failure cause, there is one or several failure mechanisms. Examples of failure mechanisms are fatigue, corrosion, and wear.
- *%MTTF*. The MTTF was entered on an MSI failure mode level. It is also interesting to know the (marginal) MTTF for each failure mechanism. To simplify, a percent is given, and the (marginal) MTTF may be estimated for each failure mechanism. The %MTTF will obviously only be an approximation because the effects of the various failure mechanisms usually are strongly interdependent.

- *Failure characteristic.* Failure propagation may be divided into three classes:
  1. The failure propagation may be measured by one or several (condition monitoring) indicators. The failure is referred to as a gradual failure.
  2. The failure probability is age-dependent, that is there is a predictable wear-out limit. The failure is referred to as an *aging failure*.
  3. Complete randomness. The failure cannot be predicted by either condition monitoring indicators or by measuring the age of the item. The time-to-failure can only be described by an exponential distribution, and the failure is referred to as a sudden failure.
- *Maintenance task*. For each failure mechanism, an appropriate maintenance task may hopefully be found by the decision logic in Figure 9.5, which is described in Step 7. This field can therefore not be completed until Step 7 is performed.
- *Failure characteristic measure*. For gradual failures, the condition monitoring indicators are listed by name. Aging failures are described by an aging parameter, that is the shape parameter (α) in the Weibull distribution is recorded.
- *Recommended maintenance interval*. In this column, the interval between consecutive maintenance tasks is given. The length of the interval is determined in Step 8.

Step 7: Selection of Maintenance Task

This step is the most novel compared to other maintenance planning techniques. A decision logic is used to guide the analyst through a question and answer process. The input to the RCM decision logic is the dominant failure modes from the FMECA in Step 6. The main idea is for each dominant failure mode to decide whether a PM task is applicable and effective, or it is best to let the item deliberately run to failure and afterwards carry out a corrective maintenance task. There are generally three main reasons for doing a PM task:

(1) To prevent a failure
(2) To detect the onset of a failure
(3) To discover a hidden fault

The following basic maintenance tasks are considered:

(1) Scheduled on-condition task
(2) Scheduled overhaul
(3) Scheduled replacement
(4) Scheduled function test
(5) Run to failure

Scheduled on-condition task is a task to determine the condition of an item, for example by condition monitoring. There are three criteria that must be met for an on-condition task to be applicable.

(1) It must be possible to detect reduced failure resistance for a specific failure mode.
(2) It must be possible to define a potential failure condition that can be detected by an explicit task.
(3) There must be a reasonable consistent age interval between the time of potential failure (P) is detected and the time of functional failure (F).

The time interval from it is possible to reveal a potential failure (P) by the currently used monitoring technique until a functional failure (F) occurs is called the *PF interval*. The P–F interval can be regarded as the potential warning time in advance of a functional failure. The longer the P–F interval, the more time one has to make a good decision and plan tasks. P–F intervals are further discussed in Section 12.3.3.

Scheduled overhaul of an item may be performed at or before some specified age limit and is often called hard time maintenance. An overhaul task is considered applicable to an item only if the following criteria are met:

(a) There must be an identifiable age at which there is a rapid increase in the item's failure rate function.
(b) A large proportion of the items must survive to that age.
(c) It must be possible to restore the original failure resistance of the item by reworking it.

Scheduled replacement is replacement of an item (or one of its parts) at or before some specified age or time limit. A scheduled replacement task is applicable only under the following circumstances:

(a) The item must be subject to a critical failure.
(b) The item must be subject to a failure that has major potential consequences.
(c) There must be an identifiable age at which the item shows a rapid increase in the failure rate function.
(d) A large proportion of the items must survive to that age.

Scheduled function test is a scheduled failure-finding task or inspection of a hidden function to identify failures. Failure finding tasks are preventive only in the sense that they prevent surprises by revealing failures of hidden functions. A scheduled function test task is applicable to an item under the following conditions:

(a) The item must be subject to a functional failure that is not evident to the operating crew during the performance of normal duties. The task has to be based on information about the failure rate function, the likely consequences and costs of the failure, the PM task is supposed to prevent the cost and risk of the PM task, and so on.

(b) The item must be one for which no other type of task is applicable and effective.

Run to failure is a deliberate decision to run to failure because the other tasks are not possible or the economics are less favorable.

PM will not prevent all failures. Consequently, if there is a clear identifiable failure mode that cannot be adequately addressed by an applicable and effective PM task that will reduce the probability of failure to an acceptable level, then there is need to redesign or modify the item. If the consequences of failures are related to safety or the environment, redesign will normally be mandatory. For operational and economic consequences of failure this may be desirable, but a cost–benefit assessment has to be performed. The criteria given for using the various tasks should only be considered as guidelines for selecting an appropriate task. A task might be found appropriate even if some of the criteria are not fulfilled.

A variety of different RCM decision logic diagrams are used in the main RCM references. Some of these are rather complicated. The decision logic diagram in Figure 9.5 is very simple and may be too simple for many applications, but the resulting maintenance tasks may – in many cases – be the same. It should be

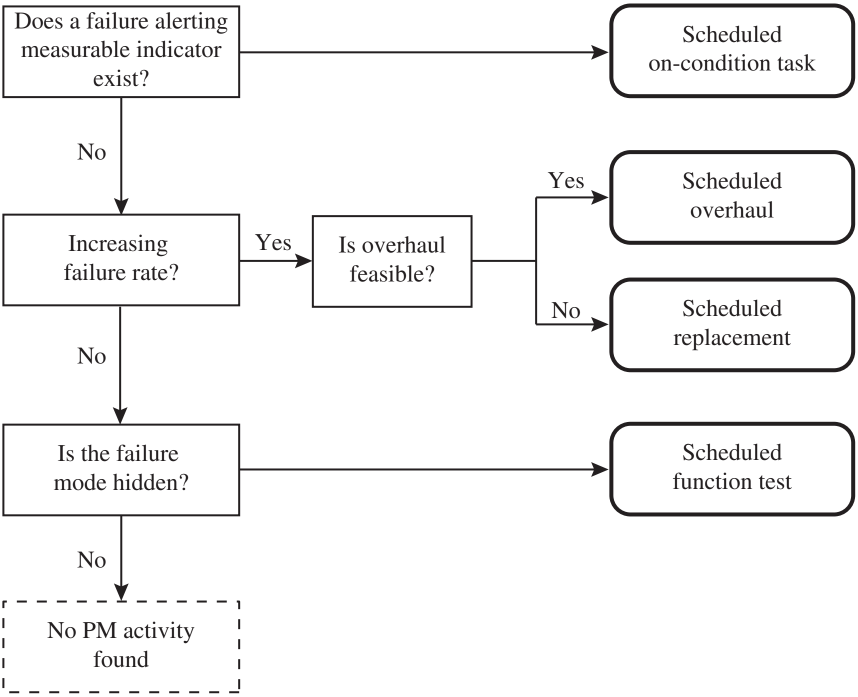


Figure 9.5 Maintenance task assignment/decision logic.

emphasized that such a logic can never cover all situations. In the case of a hidden function with aging failures, a combination of scheduled replacements and function tests is required.

Step 8: Determination of Maintenance Intervals

Some of the PM tasks are to be performed at regular intervals. To determine the optimal interval is a very difficult task that has to be based on information about the failure rate function, the likely consequences and costs of the failure the PM task is supposed to prevent, the cost and risk of the PM task, and so on. Several models are discussed in Chapter 12.

In practice, the various maintenance tasks have to be grouped into maintenance packages that are carried out at the same time, or in a specific sequence. The maintenance intervals can therefore not be optimized for each single item. The whole maintenance package has, at least to some degree, to be treated as an entity.

Step 9: Preventive Maintenance Comparison Analysis

Two overriding criteria for selecting maintenance tasks are used in RCM. Each task selected must meet two requirements:

(1) It must be applicable.
(2) It must be effective.

Applicability means that the task is applicable in relation to our reliability knowledge and in relation to the consequences of failure. If a task is found based on the preceding analysis, it should satisfy the applicability criterion. A PM task is applicable if it can eliminate a failure, or at least reduce the probability of the occurrence of failure to an acceptable level, or reduce the impact of the failure.

Cost-effectiveness means that the task does not cost more than the failure(s) it is going to prevent.

The effectiveness of a PM task is a measure of how well it accomplishes its purpose and if it is worth doing. Clearly, when evaluating the effectiveness of a task, we are balancing the cost of performing the maintenance with the cost of not performing it. The cost of a PM task may include the following:

(1) The risk/cost related to maintenance induced failures.
(2) The risk the maintenance personnel is exposed to during the task.
(3) The risk of increasing the likelihood of failure of another item while the one is out of service.
(4) The use and cost of physical resources.
(5) The unavailability of physical resources elsewhere while in use on this task.
(6) Production unavailability during maintenance.
(7) Unavailability of protective functions during maintenance of these.

In contrast, the cost of a failure may include the following:

(1) The consequences of the failure should it occur (loss of production, possible violation of laws or regulations, reduction in plant or personnel safety, or damage to other equipment).
(2) The consequences of not performing the PM task even if a failure does not occur (e.g. loss of warranty).
(3) Increased premiums for emergency repairs (such as overtime, expediting costs, or high replacement power cost).

Step 10: Treatment of Non-MSIs

In Step 4, critical items (MSIs) were selected for further analysis. A remaining question is what to do with the items that are not analyzed. For plants already having a maintenance program, a brief cost evaluation should be carried out. If the existing maintenance cost related to the non-MSIs is insignificant, it is reasonable to continue this program. See Paglia et al. (1991) for further discussion.

Step 11: Implementation

A necessary basis for implementing the result of the RCM analysis is that the organizational and technical maintenance support functions are available. A main issue is therefore to ensure that these support functions are available. Experience shows that many accidents occur either during maintenance or because of inadequate maintenance. When implementing a maintenance program, it is therefore of vital importance to consider the risk associated with the various maintenance tasks. For complicated maintenance operations, it may be relevant to perform a safe job analysis combined with a human HAZOP to reveal possible hazards and human errors related to the maintenance task (e.g. see Rausand and Haugen 2020).

Step 12: In-service Data Collection and Updating

The reliability data we have access to at the outset of the analysis may be scarce, or even second to none. In our opinion, one of the most significant advantages of RCM is that we systematically analyze and document the basis for our initial decisions, and, hence, can better utilize operating experience to adjust that decision as operating experience data become available. The full benefit of RCM is therefore only obtained when operation and maintenance experience is fed back into the analysis process.

The updating process should be concentrated on three major time perspectives:

(1) Short-term interval adjustments
(2) Medium-term task evaluation
(3) Long-term revision of the initial strategy

For each significant failure that occurs in the system, the failure characteristics should be compared with the FMECA. If the failure was not covered adequately in the FMECA, the relevant part of the RCM analysis should, if necessary, be revised.

The short-term update may be considered a revision of previous analysis results. The input to such an analysis is updated failure information and reliability estimates. This analysis should not require much resources as the framework for the analysis is already established. Only Steps 5–8 in the RCM process will be affected by short-term updates.

The medium-term update should carefully review the basis for the selection of maintenance tasks in Step 7. Analysis of maintenance experience may identify significant failure causes not considered in the initial analysis, requiring an updated FMECA in Step 6.

The long-term revision should consider all steps in the analysis. It is not sufficient to consider only the system being analyzed, it is required to consider the entire plant with its relations to the outside world, such as contractual considerations, new laws regulating environmental protection, and so on.

9.6 Total Productive Maintenance

TPM is an approach to maintenance management that was developed in Japan (Nakajima 1988) to support the implementation of just-in-time manufacturing and associated efforts to improve product quality. TPM activities focus on eliminating the *six major losses*:

Availability losses

(1) *Equipment failure (breakdown) losses*. Associated costs include downtime, labor, and spare part cost.
(2) *Setup and adjustment losses* that occur during product changeovers, shift change, or other changes in operating conditions.

Performance (speed) losses.

(3) *Idling and minor stoppages* that typically last up to 10 minutes. These include machine jams and other brief stoppages that are difficult to record, and consequently usually are hidden from efficiency reports. When combined, they can represent substantial equipment downtime.
(4) *Reduced speed losses* that occur when equipment must be slowed down to prevent quality defects or minor stoppages. In most cases, this loss is not recorded because the equipment continues to operate, albeit at a lower speed. Speed losses obviously have a negative effect on productivity and asset utilization.

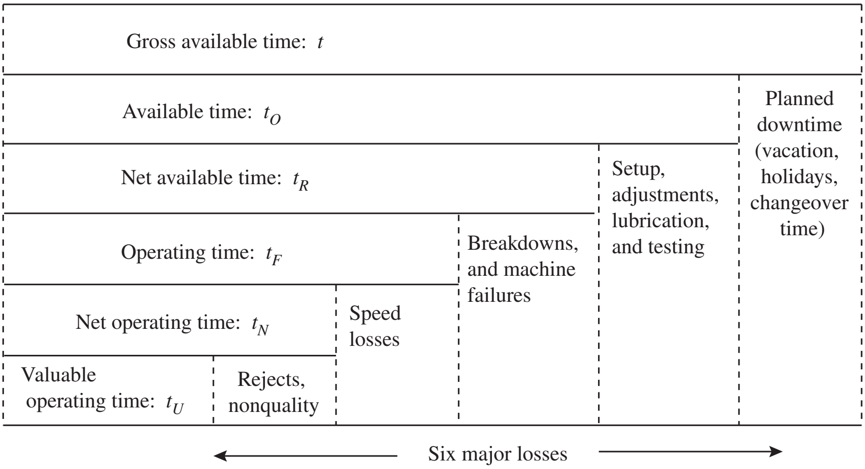

Gross available time: t
Available time: t_O
Net available time: t_R
Operating time: t_F
Net operating time: t_N
Valuable operating time: t_U
Rejects, nonquality
Speed losses
Breakdowns, and machine failures
Setup, adjustments, lubrication, and testing
Planned downtime (vacation, holidays, changeover time)
Six major losses

Figure 9.6 Time concepts used in Total productive maintenance.

Quality losses

(1) *Defects in process and reworking losses* that are caused by manufacture of defective or substandard products, that must be reworked or scrapped. These losses include the labor and material costs (if scrapped) associated with off-specification production.

(2) *Yield losses* reflect the wasted raw materials associated with the quantity of rejects and scrap that result from start-ups, changeovers, equipment limitations, poor product design, and so on. It excludes the category 5 defect losses that result during normal production.

The six major losses determine the *overall equipment effectiveness* (OEE), which is a multiplicative combination of equipment availability losses (1 and 2), equipment performance losses (3 and 4), and quality losses (5 and 6). The time concepts used in TPM are illustrated in Figure 9.6. The factors used to determine the OEE are:

Operational availability	$A_O =$	t_F/t_R
Performance rate	$R_P =$	t_N/t_F
Quality rate	$R_Q =$	t_U/t_F

The quality rate may alternatively be measured as

$$\text{Quality rate} = R_Q = \frac{\text{No. of processed products} - \text{No. of rejected products}}{\text{No. of processed products}}$$

The OEE is defined as

$$\text{OEE} = A_O R_P R_Q \tag{9.2}$$

The OEE is used as an indicator of how well machines, production lines, and processes are performing in terms of availability, performance, and quality. An OEE $\geq 85\%$ is considered to be "world class."

TPM has been described as a partnership approach to maintenance. Under TPM, small groups or teams create a cooperative relationship between maintenance and production. Production workers become involved in performing maintenance work allowing them to play a role in equipment monitoring and upkeep. This raises the skill of production workers and allows them to be more effective in maintaining the equipment in good condition. Team-based activities play an important role in TPM. Team-based activities involve groups from maintenance, production, and engineering. The technical skill of engineers and the experience of maintenance workers and equipment operators are communicated through these teams. The objective of the team-based activities is to improve equipment performance through better communication of current and potential equipment problems. Maintainability improvement and maintenance prevention are two team-based TPM activities. TPM has several benefits. The efforts of maintenance improvement teams should result in improved equipment availability and reduced maintenance costs. Maintainability improvement should result in increased maintenance efficiency and reduced repair time. TPM resembles total quality management (TQM) in several aspects, such as (i) total commitment to the program from upper level management is required, (ii) employees must be empowered to initiate corrective tasks, and (iii) a long range outlook must be accepted as TPM may take a year or more to implement and is an ongoing process.

9.7 Problems

9.1 Discuss the maintainability of your bicycle and suggest improvements of its maintainability.

9.2 Give a practical example where an as-bad-as-old repair action may be a realistic assumption.

9.3 Assume that you are driving your car when it suddenly fails and you have to leave it where it stopped. Explain what is meant by MDT in this case and list the main elements adding up to the MDT.

9.4 List and discuss the main differences between RCM and TPM. Are the two approaches competitors, can they be combined, or do they serve two totally different purposes? Justify your answers. (Further information about RCM and TPM may be found by searching the Internet.)

References

DIN 31051 (2012). *Fundamentals of maintenance*, *German standard*. Berlin: German Institute for Standardization.

Ebeling, C.E. (2009). *An Introduction to Reliability and Maintainability Engineering*, 2e. Waveland Press.

EN 13306 (2017). *Maintenance–maintenance terminology*, *European standard*. European Committee for Standardization (CEN).

Fedele, L. (2011). *Methodologies and Techniques for Advanced Maintenance*. London: Springer.

IEC 60300-3-11 (2009). *Dependability management–application guide–Part 3-11: reliability-centred maintenance*, *International standard*. Geneva: International Electrotechnical Commission.

IEC 60300-3-14 (2004). *Dependability management–application guide–Part 3-14: maintenance and maintenance support*, *International standard*. Geneva: International Electrotechnical Commission.

IEC 60706 (2006). *Maintainability of equipment*, *International standard [Series of several standards]*. Geneva: International Electrotechnical Commission.

IEC TR 62278-3 (2010). *Railway applications–specification and demonstration of reliability, availability, maintainability and safety (RAMS)–Part 3: guide to the application of IEC62278 for rolling stock RAM*, *International standard*. Geneva: International Electrotechnical Commission.

ISO 17359 (2018). *Condition monitoring and diagnosis of machines–general guidelines*, *International standard*. Geneva: International Organization for Standardization.

ISO 55000 (2014). *Asset management–overview, principles and terminology*, *International standard*. Geneva: International Organization for Standardization.

MIL-HDBK-338B (1998). *Electronic Reliability Design Handbook*, *Military handbook*. Washington, DC: U.S. Department of Defense.

MIL-HDBK-470A (1997). *Designing and developing maintainable products and systems*, *Military handbook*. Washington, DC.

Nakajima, S. (1988). *Introduction to TPM: Total Productivity Maintenance*, 11e. Cambridge, MA: Productivity Press.

NASA (2008). *RCM guide: reliability-centered maintenance guide for facilities and collateral equipment*, *Guideline*. Washington, DC: U.S. National Aeronautics and Space Administration.

Nowlan, F.S. and Heap, H.F. (1978). Reliability-Centered Maintenance. *Tech. Rep. A066-579*. San Francisco, CA: United Airlines.

Paglia, A., Barnard, D., and Sonnett, D.A. (1991). A case study of the RCM project at V.C. Summer nuclear generating station. *4th International Power Generation Exhibition and Conference*, volume 5, Tampa, FL. pp. 1003–1013.

Rausand, M. and Haugen, S. (2020). *Risk Assessment; Theory, Methods, and Applications*, 2e. Hoboken, NJ: Wiley.

SAE JA1010 (2011). *Maintainability program standard*, *Standard*. Warrendale, PA: SAE International.

SAE JA1012 (2011). *Guide to the reliability-centered maintenance RCM standard*, *Guideline*. Warrendale, PA: SAE International.

Smith, D.J. (2013). *Reliability, Maintainability and Risk: Practical Methods for Engineers*, 8e. Oxford: Butterworth Heinemann.

10

Counting Processes

10.1 Introduction

This chapter studies the reliability of a single and repairable item as a function of time. The goal is to determine relevant reliability metrics, such as the item's availability, the mean number of failures during a specified time interval, the mean time to the first item failure, and the mean time between item failures. For this purpose, the item is studied by using stochastic processes.

A *stochastic process* $\{X(t), t \in \Theta\}$ is a collection of random variables. The set Θ is called the *index set* of the process. For each *index* t in $\Theta, X(t)$ is a random variable. The index t is here interpreted as time, and $X(t)$ is called the *state* of the process at time t. When the index set Θ is countable, the process is a discrete-time stochastic process. When Θ is a continuum, we have a continuous-time stochastic process. The presentation of the various processes in this book is brief and limited, and focuses on results that can be applied in practice instead of mathematical rigor. The reader should consult a textbook on stochastic processes for more details. Good treatments of stochastic processes are given by, for example, Ross (1996), Cocozza-Thivent (1997), and Cha and Finkelstein (2018).

10.1.1 Counting Processes

Consider a repairable item that is put into operation at time $t = 0$. The first item failure occurs at time S_1, which is a random variable. When the item has failed, it is replaced or restored to a functioning state. The repair time is assumed to be so short that it may be neglected. The second failure occurs at time S_2, and so on. In this way, a sequence of failure times $S_1, S_2, \dots$ is obtained. Let T_i be the time between failure $i - 1$ and failure i for $i = 1, 2, \dots$, where S_0 is taken to be 0. T_i is called the *interoccurrence time* i for $i = 1, 2, \dots$. T_i is also called the *time between failures*, and the *interarrival time*.

System Reliability Theory: Models, Statistical Methods, and Applications, Third Edition.
Marvin Rausand, Anne Barros, and Arnljot Høyland.

Companion Website: www.wiley.com/go/SystemReliabilityTheory3e

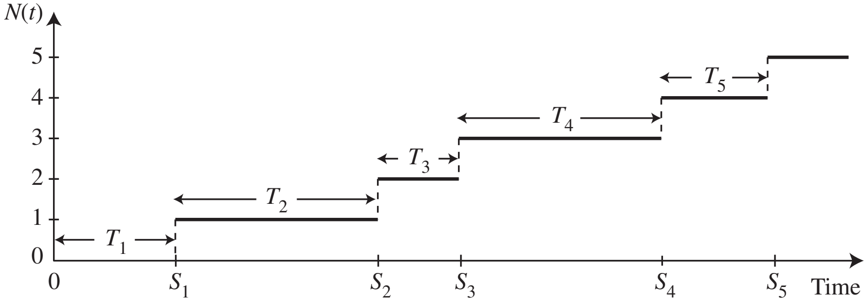


Figure 10.1 Relation between the number of events $N(t)$, the interoccurrence times (T_i), and the calendar times (S_i)

Throughout this chapter, t denotes a specified point of time, irrespective whether t is *calendar time* (a realization of S_i) or *local time* (a realization of an interoccurrence time T_i). We hope that this convention does not confuse the reader. The time concepts are illustrated in Figure 10.1.

The sequence of interoccurrence times, $T_1, T_2, \ldots$ is generally not independent and identically distributed – unless the item is replaced upon failure, or restored to an as-good-as-new condition, and the operating context remains constant during the whole period.

A counting process is a special type of a stochastic process, and is defined as:

Definition 10.1 (Counting process)
A stochastic process $\{N(t), t \geq 0\}$ that satisfies:

(1) $N(t) \geq 0$
(2) $N(t)$ is integer valued.
(3) If $s < t$, then $N(s) \leq N(t)$
(4) For $s < t$, $[N(t) - N(s)]$ represents the number of failures that have occurred in the interval $(s, t]$.

□

Definition 10.1 is adapted from Ross (1996). A counting process $\{N(t), t \geq 0\}$ may alternatively be represented by the sequence of failure (calendar) times $S_1, S_2, \ldots$, or by the sequence of interoccurrence times $T_1, T_2, \ldots$. The three representations contain the same information about the counting process. Main features of counting processes are illustrated in Examples 10.1 and 10.2. The two examples also introduce some new concepts.

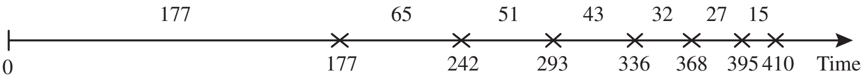


Figure 10.2 The dataset in Example 10.1.

Example 10.1 (Sad versus happy items)

The following failure times (calendar time in days) come from Ascher and Feingold (1984). The dataset is recorded from time $t = 0$ until seven failures have occurred during a total time of 410 (days). The data represent a single item, and the repair times are assumed to be negligible. This means that the item is assumed to be functioning again almost immediately after a failure is encountered.

Number of failures $N(t)$	Calendar time S_j	Interoccurrence time T_j
0	0	0
1	177	177
2	242	65
3	293	51
4	336	43
5	368	32
6	395	27
7	410	15

The dataset is shown in Figure 10.2. The interoccurrence times are seen to become shorter with time. The item seems to be deteriorating, and failures tend to become more frequent. An item with this property is called a *sad* item by Ascher and Feingold (1984). An item with the opposite property, where failures become less frequent with operating time, is called a *happy* item.

The number of failures $N(t)$ is shown as a function of (calendar) time t in Figure 10.3. Observe that $N(t)$ by definition is constant between failures and jumps (a height of 1 item) at the failure times S_i for $i = 1, 2, \ldots$. It is thus sufficient to plot the jumping points $(S_i, N(S_i))$ for $i = 1, 2, \ldots$. The plot is called an $N(t)$ plot, or a Nelson–Aalen plot (see Chapter 14).

Observe that $N(t)$ as a function of t tends to be convex when the item is *sad*. In the same way, $N(t)$ tends to be concave when the item is *happy*.[1] If $N(t)$ is

[1] Observe that we are using the terms *convex* and *concave* in a rather inaccurate way here. What we mean is that the observed points $\{t_i, N(t_i)\}$ for $i = 1, 2, \ldots$ approximately follow a convex/concave curve.

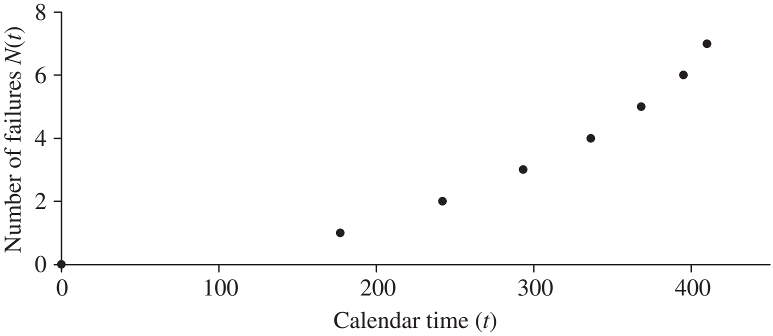


Figure 10.3 Number of failures $N(t)$ as a function of time for the data in Example 10.1.

(approximately) linear, the item is steady, that is, the interoccurrence times have the same expected length. In Figure 10.2, $N(t)$ is clearly seen to be convex, and the item is *sad*. □

Example 10.2 (Compressor failure data)
Failure time data for a specific compressor at a Norwegian process plant was collected as part of a student thesis at NTNU. All compressor failures in the time period 1968–1989 were recorded. In this period, a total of 321 failures occurred. 90 of these failures were critical failures and 231 failures were noncritical. In this context, a critical failure is a failure causing compressor downtime. Noncritical failures may be corrected without having to stop the compressor. The majority of the noncritical failures were instrument failures, and failures related to the seal oil system and the lubrication oil system.

As above, let $N(t)$ be the number of compressor failures in the time interval $(0, t]$. From a production point of view, the critical failures are the most important, because they lead to process shutdown. The operating times (in days) at which the 90 critical failures occurred are listed in Table 10.1. Here, the time t denotes the *operating* time, which means that the downtimes caused by compressor failures and process shutdowns are not included. An $N(t)$ plot of the 90 critical failures is shown in Figure 10.4.

In this case, the $N(t)$ plot is slightly concave and indicates a *happy* item. The time between critical failures seems to increase with the time in operation. Also observe that several failures have occurred within short intervals. This indicates that the failures may be dependent, or that the maintenance personnel have not been able to correct the failures properly at the first attempt. □

An analysis of life data from a repairable item should always be started by establishing an $N(t)$ plot. If $N(t)$ as a function of the time t is nonlinear, methods based

Table 10.1 Failure times (operating days) in chronological order.

1.0	4.0	4.5	92.0	252.0	277.0
277.5	284.5	374.0	440.0	444.0	475.0
536.0	568.0	744.0	884.0	904.0	1017.5
1288.0	1337.0	1338.0	1351.0	1393.0	1412.0
1413.0	1414.0	1546.0	1546.5	1575.0	1576.0
1666.0	1752.0	1884.0	1884.2	1884.4	1884.6
1884.8	1887.0	1894.0	1907.0	1939.0	1998.0
2178.0	2179.0	2188.5	2195.5	2826.0	2847.0
2914.0	3156.0	3156.5	3159.0	3211.0	3268.0
3276.0	3277.0	3321.0	3566.5	3573.0	3594.0
3640.0	3663.0	3740.0	3806.0	3806.5	3809.0
3886.0	3886.5	3892.0	3962.0	4004.0	4187.0
4191.0	4719.0	4843.0	4942.0	4946.0	5084.0
5084.5	5355.0	5503.0	5545.0	5545.2	5545.5
5671.0	5939.0	6077.0	6206.0	6206.5	6305.0

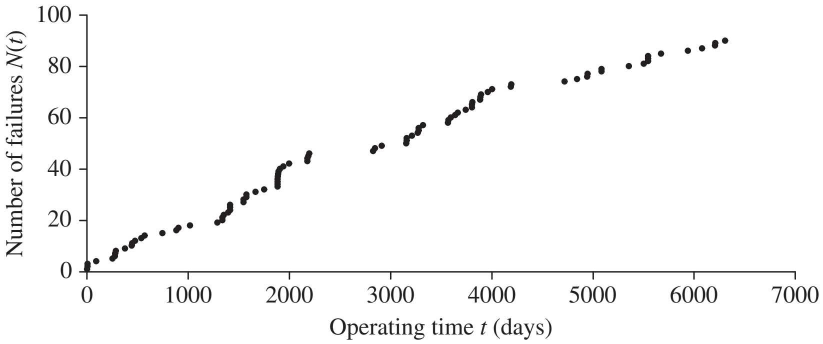


Figure 10.4 Number of critical compressor failures $N(t)$ as a function of time (days), (totaling 90 failures).

on the assumption of independent and identically distributed times between failures are obviously not appropriate. It is, however, not certain that such methods are appropriate even if the $N(t)$ plot is very close to a straight line. The interoccurrence times may be strongly correlated. Methods to check whether the interoccurrence times are correlated or not, are discussed, e.g. by Ascher and Feingold (1984) and Bendell and Walls (1985). The $N(t)$ plot is further discussed in Section 10.4.

10.1.2 Basic Concepts

Throughout this section, we assume that the events that are counted are *failures*. In some of the applications later in this chapter we also study other types of events, such as repairs. Some of the concepts must be reformulated to be meaningful in these applications. We hope that this does not confuse the reader.

- *Independent increments.* A counting process $\{N(t), t \geq 0\}$ is said to have independent increments if for $0 < t_1 < \cdots < t_k$, $k = 2, 3, \ldots$ $[N(t_1) - N(0)], [N(t_2) - N(t_1)], \ldots, [N(t_k) - N(t_{k-1})]$ are all independent random variables. In that case the number of failures in an interval is not influenced by the number of failures in any strictly earlier intervals (i.e. with no overlap). This means that even if the item has experienced an unusual high number of failures in a certain time interval, this does not influence the distribution of future failures.
- *Stationary increments.* A counting process is said to have stationary increments if for any two disjoint time points $t > s \geq 0$ and any constant $c > 0$, the random variables $[N(t) - N(s)]$ and $[N(t + c) - N(s + c)]$ are identically distributed. This means that the distribution of the number of failures in a time interval depends only on the length of the interval, and not on the interval's distance from the origin.
- *Stationary process.* A counting process is said to be stationary (or homogeneous) if it has stationary increments.
- *Nonstationary process.* A counting process is said to be nonstationary (or nonhomogeneous) if it is neither stationary nor eventually becomes stationary.
- *Regular process.* A counting process is said to be regular (or orderly) if

  $$\Pr[N(t + \Delta t) - N(t) \geq 2] = o(\Delta t), \tag{10.1}$$

  when Δt is small, and $o(\Delta t)$ is a function of Δt with the property that $\lim_{\Delta t \to 0} o(\Delta t)/\Delta t = 0$. This means that the item will not have two or more simultaneous failures.
- *Rate of the process.* The rate of the counting process at time t is defined as:

  $$w(t) = W'(t) = \frac{d}{dt} E[N(t)], \tag{10.2}$$

  where $W(t) = E[N(t)]$ is the mean number of failures (events) in the interval $(0, t]$. Thus

  $$w(t) = W'(t) = \lim_{\Delta t \to 0} \frac{E[N(t + \Delta t) - N(t)]}{\Delta t}, \tag{10.3}$$

and when Δt is small,

$$w(t) \approx \frac{E[N(t+\Delta t) - N(t)]}{\Delta t}$$
$$= \frac{\text{Mean number of failures in } (t, t+\Delta t]}{\Delta t}$$

A natural estimator of $w(t)$ is

$$\hat{w}(t) = \frac{\text{Number of failures in } (t, t+\Delta t]}{\Delta t}, \tag{10.4}$$

for some suitable Δt. It follows that the rate $w(t)$ of the counting process, may be regarded as the mean number of failures (events) per time unit at time t. When we are dealing with a *regular* process, the probability of two or more failures in $(t, t+\Delta t]$ is negligible when Δt is small and we may assume that

$$N(t+\Delta t) - N(t) = 0 \quad \text{or} \quad 1.$$

The mean number of failures in $(t, t+\Delta t]$ is hence approximately equal to the probability of failure in $(t, t+\Delta t]$, and

$$w(t) \approx \frac{\text{Probability of failure in } (t, t+\Delta t]}{\Delta t}. \tag{10.5}$$

Hence, $w(t)\Delta t$ can be interpreted as the probability of failure in the time interval $(t, t+\Delta t]$. Some authors write (10.5) as

$$w(t) = \lim_{\Delta t \to 0} \frac{\Pr[N(t+\Delta t) - N(t) = 1]}{\Delta t},$$

as definition of the rate of the process. Observe also that

$$E[N(t_0)] = W(t_0) = \int_0^{t_0} w(t)\, dt. \tag{10.6}$$

- *Rate of occurrence of failures (ROCOF).* When the events of a counting process are failures, the rate $w(t)$ of the process is often called the *rate of occurrence of failures* (ROCOF).
- *Time between failures.* We have denoted the time T_i between failure $i-1$ and failure i, for $i = 1, 2, \ldots$, the interoccurrence times. For a general counting process, the interoccurrence times are neither identically distributed nor independent. Hence, the mean times between failures, $\text{MTBF}_i = E(T_i)$, are in general a function of i and $T_1, T_2, \ldots, T_{i-1}$.
- *Forward recurrence time.* The forward recurrence time $Y(t)$ is the time to the next failure measured from an arbitrary point of time t. Thus $Y(t) = S_{N(t)+1} - t$. The forward recurrence time is also called the *residual lifetime*, the *remaining lifetime*, or the *excess life*. The forward recurrence time is illustrated in Figure 10.5.

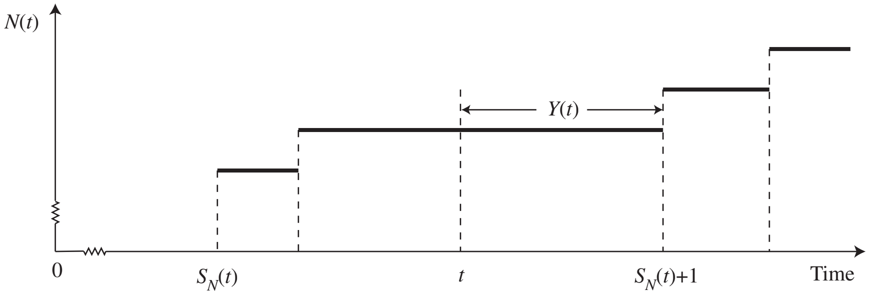


Figure 10.5 The forward recurrence time $Y(t)$.

Many of the results in this chapter are only valid for *nonlattice* distributions. A lattice distribution is defined as:

Definition 10.2 (Lattice distribution)
A nonnegative random variable is said to have a *lattice* (or periodic) distribution if there exists a number $d \geq 0$ such that

$$\sum_{n=0}^{\infty} \Pr(X = nd) = 1.$$

In words, X has a lattice distribution if X can only take on values that are integral multiples of some nonnegative number d. □

10.1.3 Martingale Theory

Martingale theory can be applied to counting processes to make a record of the *history* of the process. Let $\mathcal{H}_t$ denote the history of the process up to, but not including, time t. Usually, we think of $\mathcal{H}_t$ as $\{N(s), 0 \leq s < t\}$ which keeps record of all failures before time t. It could, however, contain more specific information about each failure.

A *conditional rate of failures* may be defined as

$$w_C(t \mid \mathcal{H}_t) = \lim_{\Delta t \to \infty} \frac{\Pr(N(t + \Delta t) - N(t) = 1 \mid \mathcal{H}_t)}{\Delta t}. \tag{10.7}$$

Thus, $w_C(t \mid \mathcal{H}_t)\Delta t$ is approximately the probability of failure in the interval $[t, t + \Delta t)$ conditional on the failure history up to, but not including time t. Observe that the rate of the process (ROCOF) defined in (10.2) is the corresponding unconditional rate of failures.

Usually, the process depends on the history through random variables and $w_C(t \mid \mathcal{H}_t)$ will consequently be *stochastic*. It should be observed that $w_C(t \mid \mathcal{H}_t)$ is stochastic only through the history: for a fixed history (that is for a given state just before time t), $w_C(t \mid \mathcal{H}_t)$ is not stochastic. To simplify the notation, we will in the following omit the explicit reference to the history $\mathcal{H}_t$ and let $w_C(t)$ be the conditional ROCOF.

The martingale approach for modeling counting processes require rather sophisticated mathematics. We will therefore avoid using this approach during the main part of the chapter, but will touch upon martingales in Section 10.5 where we discuss imperfect repair models. A brief, but clear introduction to martingales used in counting processes is given by Hokstad (1997). For a more rigorous treatment, see Andersen et al. (1993).

10.1.4 Four Types of Counting Processes

This chapter examines four types of counting processes.

(1) Homogeneous Poisson processes (HPP),
(2) Renewal processes
(3) Nonhomogeneous Poisson processes (NHPP)
(4) Imperfect repair processes

The HPP is introduced in Section 5.5. In the HPP model, all the interoccurrence times are independent and exponentially distributed with the same parameter (failure rate) λ.

The renewal process and the NHPP are generalizations of the HPP, both having the HPP as a special case. A renewal process is a counting process where the interoccurrence times are independent and identically distributed with an arbitrary time-to-failure distribution. Upon failure, the item is thus replaced or restored to an as-good-as-new condition. This is often called a *perfect repair*. Statistical analysis of observed interoccurrence times from a renewal process is discussed in detail in Chapter 14.

The NHPP differs from the HPP in that the ROCOF varies with time rather than being a constant. This implies that for an NHPP model, the interoccurrence times are neither independent nor identically distributed. The NHPP is often used to model repairable items that are subject to a *minimal repair* strategy, with negligible repair times. Minimal repair means that a failed item is restored just back to functioning state. After a minimal repair, the item continues as if nothing had happened. The likelihood of item failure is the same immediately before and after a failure. A minimal repair thus restores the item to an as-bad-as-old condition. The minimal repair strategy is discussed, for example, by Ascher and Feingold (1984) and Akersten (1991) who give a detailed list of relevant references on this subject.

The renewal process and the NHPP represent two extreme types of repair: replacement to an as-good-as-new condition and replacement to as-bad-as-old (minimal repair), respectively. Most repair actions are somewhere between these

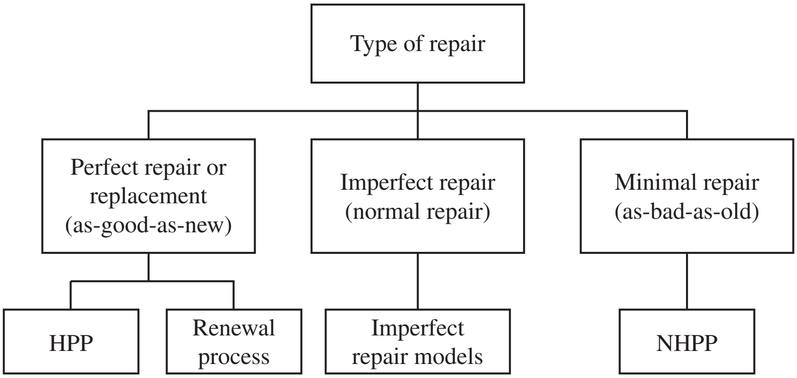

Type of repair
Perfect repair or replacement (as-good-as-new)
Imperfect repair (normal repair)
Minimal repair (as-bad-as-old)
HPP
Renewal process
Imperfect repair models
NHPP

Figure 10.6 Types of repair and stochastic point processes covered in this book.

extremes and are often called *imperfect repair*, or *normal repair*. A number of different models have been proposed for imperfect repair. A survey of some of these models is given in Section 10.5. The various types of repair and the models covered in this book are shown in Figure 10.6.

10.2 Homogeneous Poisson Processes

The HPP is introduced in Section 5.8.5. The HPP may be defined in a number of different ways. Three alternative definitions of the HPP are presented in the following, to illustrate different features of the HPP. The two first definitions are based on Ross (1996).

Definition 10.3 (Homogeneous Poisson process – 1)
The counting process $\{N(t), t \geq 0\}$ is said to be an HPP with rate λ, for $\lambda > 0$, if

(1) $N(0) = 0$
(2) The process has independent increments.
(3) The number of events in any interval of length t is Poisson distributed with mean λt. That is, for all $s, t > 0$,

$$\Pr(N(t+s) - N(s) = n) = \frac{(\lambda t)^n}{n!}\, e^{-\lambda t} \qquad \text{for} \qquad n = 0, 1, 2, \dots . \tag{10.8}$$

□

Observe that it follows from property 3 that an HPP has stationary increments and also that $E[N(t)] = \lambda t$, which explains why λ is called the rate of the process.

Definition 10.4 (Homogeneous Poisson process – 2)
The counting process $\{N(t), t \geq 0\}$ is said to be an HPP with rate λ, for $\lambda > 0$, if

(1) $N(0) = 0$
(2) The process has stationary and independent increments.
(3) $\Pr(N(\Delta t) = 1) = \lambda \Delta t + o(\Delta t)$
(4) $\Pr(N(\Delta t) \geq 2) = o(\Delta t)$

□

These two alternative definitions of the HPP are presented to clarify the analogy to the definition of the NHPP that is presented in Section 10.4. The third definition of the HPP is adapted from Cocozza-Thivent (1997).

Definition 10.5 (Homogeneous Poisson process – 3)
The counting process $\{N(t), t \geq 0\}$ is said to be an HPP with rate $\lambda > 0$, if $N(0) = 0$, and the interoccurrence times $T_1, T_2, \ldots$ are independent and exponentially distributed with parameter λ. □

10.2.1 Main Features of the HPP

The main features of the HPP can be easily deduced from the three alternative definitions:

(1) The HPP is a regular counting process with independent and stationary increments.
(2) The rate of occurrence of failures, ROCOF, of the HPP is constant and independent of time,

$$w(t) = \lambda \qquad \text{for} \qquad t \geq 0. \tag{10.9}$$

(3) The number of failures in the interval $(t, t + v]$ is Poisson distributed with mean λv,

$$\Pr[N(t + v) - N(t) = n] = \frac{(\lambda v)^n}{n!}\, e^{-\lambda v} \qquad \text{for} \quad t \geq 0, \qquad v > 0. \tag{10.10}$$

(4) The mean number of failures in the time interval $(t, t + v]$ is

$$W(t + v) - W(t) = E[N(t + v) - N(t)] = \lambda v. \tag{10.11}$$

Especially observe that $E[N(t)] = \lambda t$, and $\text{var}[N(t)] = \lambda t$.
(5) The interoccurrence times $T_1, T_2, \ldots$ are independent and identically distributed exponential random variables having mean $1/\lambda$
(6) The time of the nth failure $S_n = \sum_{i=1}^{n} T_i$ has a gamma distribution with parameters (n, λ). Its probability density function is

$$f_{S_n}(t) = \frac{\lambda}{(n-1)!}\, (\lambda t)^{n-1} e^{-\lambda t} \qquad \text{for} \qquad t \geq 0. \tag{10.12}$$

Further features of the HPP are presented and discussed, for example, by Ross (1996) and Ascher and Feingold (1984).

Remark 10.1 (Comparing definitions of the HPP)
Consider an HPP defined by Definition 10.5 where the interoccurrence times $T_1, T_2, \ldots$ are independent and exponentially distributed with parameter λ. The arrival time S_n is, according to (10.12), gamma distributed with parameters (n, λ). Because $N(t) = n$ if and only if $S_n \leq t < S_{n+1}$, and the interoccurrence time $T_{n+1} = S_{n+1} - S_n$, we use the law of total probability (see Section 6.2.4) to write

$$\begin{aligned}\Pr(N(t) = n) &= \Pr(S_n \leq t < S_{n+1}) \\ &= \int_0^t \Pr(T_{n+1} > t - s \mid S_n = s)\, f_{S_n}(s)\, ds \\ &= \int_0^t e^{-\lambda(t-s)}\ \frac{\lambda}{(n-1)!}\ (\lambda s)^{n-1}\ e^{-\lambda s}\, ds \\ &= \frac{(\lambda t)^n}{n!}\ e^{-\lambda t}. \end{aligned} \tag{10.13}$$

This shows that $\{N(t), t \geq 0\}$, is an HPP with mean λt according to Definition 10.5. □

10.2.2 Asymptotic Properties

The following asymptotic results apply:

$$\frac{N(t)}{t} \to \lambda \qquad \text{with probability 1, when} \qquad t \to \infty,$$

and

$$\frac{N(t) - \lambda t}{\sqrt{\lambda t}} \overset{\mathcal{L}}{\to} \mathcal{N}(0, 1),$$

such that

$$P\left(\frac{N(t) - \lambda t}{\sqrt{\lambda t}} \leq t\right) \approx \Phi(t) \qquad \text{when} \quad t \to \infty, \tag{10.14}$$

where $\Phi(t)$ is the distribution function of the standard normal distribution $\mathcal{N}(0, 1)$.

10.2.3 Estimate and Confidence Interval

General information about estimates and confidence intervals is found in Chapter 14. Readers who are not familiar with estimation theory may find it useful to consult that chapter. In this section, we suffice by summarizing some main formulas.

An obvious estimator for λ is

$$\hat{\lambda} = \frac{N(t)}{t}. \tag{10.15}$$

The estimator is unbiased, $E(\hat{\lambda}) = \lambda$, with variance, $\text{var}(\hat{\lambda}) = \lambda/t$.

A $(1 - \varepsilon)$ confidence interval for λ, when $N(t) = n$ events (failures) are observed during a time interval of length t, is given as (e.g. see Cocozza-Thivent 1997, p. 63):

$$\left(\frac{1}{2t}\, z_{1-\varepsilon/2,\ 2n},\ \frac{1}{2t}\, z_{\varepsilon/2,\ 2(n+1)}\right), \tag{10.16}$$

where $z_{\varepsilon,\nu}$ is the upper $100\varepsilon\%$ percentile of the chi-square (χ^2) distribution with ν degrees of freedom. Percentile values (e.g. 95%) of the χ^2 distribution with n degrees of freedom are found in R by the command `qchisq(0.95, df=n)`.

In some situations, it is of interest to give an upper $(1 - \varepsilon)$ confidence limit for λ. Such a limit is obtained through the one-sided confidence interval given by

$$\left(0,\ \frac{1}{2t}\, z_{\varepsilon,\ 2(n+1)}\right). \tag{10.17}$$

Observe that this interval is applicable even if no failures (i.e. $N(t) = 0$) are observed during the interval $(0, t)$.

10.2.4 Sum and Decomposition of HPPs

Let $\{N_1(t), t \geq 0\}$ and $\{N_2(t), t \geq 0\}$ be two independent HPPs with rates λ_1 and λ_2, respectively. Further, let $N(t) = N_1(t) + N_2(t)$. It is then easy to verify (see Problem 4) that $\{N(t), t \geq 0\}$ is an HPP with rate $\lambda = \lambda_1 + \lambda_2$.

Suppose that in an HPP $\{N(t), t \geq 0\}$, we can classify each failure as Type 1 and Type 2 that are occurring with probability p and $(1 - p)$, respectively. This is, for example, the case when we have a sequence of failures with two different failure modes (1 and 2), and p equals the relative number of failure mode 1. Then the number of events, $N_1(t)$ of Type 1, and $N_2(t)$ of Type 2 in the interval $(0, t]$ also generate HPPs, $\{N_1(t), t \geq 0\}$ and $\{N_2(t), t \geq 0\}$ with rates $p\lambda$ and $(1 - p)\lambda$, respectively. Furthermore, the two processes are independent. For a formal proof, see, for example, Ross (1996, p. 69). These results can be easily generalized to more than two cases.

Example 10.3 (Failures of a specific type)

Let $\{N(t), t \geq 0\}$ be an HPP with rate λ. Some failures develop into a consequence C, others do not. The failures developing into a consequence C are denoted a C-failure. C may, for example, be a specific failure mode. The probability that a failure develops into consequence C is denoted p and is constant for each failure. The failure consequences are further assumed to be independent of each other. Let

$N_C(t)$ be the number of C-failures in the time interval $(0, t]$. When $N(t)$ is equal to n, $N_C(t)$ is binomially distributed.

$$\Pr(N_C(t) = y \mid N(t) = n) = \binom{n}{y} p^y (1-p)^{n-y} \qquad \text{for} \qquad y = 0, 1, 2, \ldots, n.$$

The marginal distribution of $N_C(t)$ is

$$\begin{aligned}\Pr(N_C(t) = y) &= \sum_{n=y}^{\infty} \binom{n}{y} p^y (1-p)^{n-y} \frac{(\lambda t)^n}{n!} e^{-\lambda t} \\ &= \frac{p^y e^{-\lambda t}}{y!} (\lambda t)^y \sum_{n=y}^{\infty} \frac{[\lambda t(1-p)]^{n-y}}{(n-y)!} \\ &= \frac{(p\lambda t)^y e^{-\lambda t}}{y!} \sum_{x=0}^{\infty} \frac{[\lambda t(1-p)]^x}{x!} \\ &= \frac{(p\lambda t)^y e^{-\lambda t}}{y!} e^{\lambda t(1-p)} \\ &= \frac{(p\lambda t)^y}{y!} e^{-p\lambda t}, \end{aligned} \tag{10.18}$$

which shows that $\{N_C(t), t \geq 0\}$ is an HPP with rate $p\lambda$. The mean number of C-failures in the time interval $(0, t]$ is

$$E[N_C(t)] = p\lambda t.$$

□

10.2.5 Conditional Distribution of Failure Time

Suppose that exactly one (hidden) failure of an HPP with rate λ is known to have occurred sometime in the interval $(0, t_0]$. We want to determine the distribution of the time T_1 at which the failure occurred.

$$\begin{aligned}\Pr(T_1 \leq t \mid N(t_0) = 1) &= \frac{\Pr(T_1 \leq t \cap N(t_0) = 1)}{\Pr(N(t_0) = 1)} \\ &= \frac{\Pr(1 \text{ failure in } (0, t] \cap 0 \text{ failures in } (t, t_0])}{\Pr(N(t_0) = 1)} \\ &= \frac{\Pr(N(t) = 1) \Pr(N(t_0) - N(t) = 0)}{\Pr(N(t_0) = 1)} \\ &= \frac{\lambda t e^{-\lambda t}\ e^{-\lambda(t_0 - t)}}{\lambda t_0 e^{-\lambda t_0}} \\ &= \frac{t}{t_0} \qquad \text{for} \qquad 0 < t \leq t_0. \end{aligned} \tag{10.19}$$

When we know that exactly one failure (event) takes place in the time interval $(0, t_0]$, the time at which the failure occurs is *uniformly* distributed over $(0, t_0]$.

Hence, each interval of equal length in $(0, t_0]$ has the same probability of containing the failure. The expected time at which the failure occurs is

$$E(T_1 \mid N(t_0) = 1) = \frac{t_0}{2}. \tag{10.20}$$

This result is important for the analyses of safety-instrumented systems in Chapter 13.

10.2.6 Compound HPPs

Consider an HPP, $\{N(t), t \geq 0\}$, with rate λ. A random variable X_i is associated to failure event i, for $i = 1, 2, \ldots$. The variable X_i may, for example, be the consequence (economic loss) associated with failure i. The variables $X_1, X_2, \ldots$ are assumed to be independent with common distribution function

$$F_X(x) = \Pr(X \leq x).$$

The variables $X_1, X_2, \ldots$ are further assumed to be independent of $N(t)$. The cumulative consequence at time t is

$$Z(t) = \sum_{i=1}^{N(t)} X_i \qquad \text{for } t \geq 0. \tag{10.21}$$

The process $\{Z(t), t \geq 0\}$ is called a *Compound Poisson process*. Compound Poisson processes are discussed, for example, by Ross (1996). The same model is called a *cumulative damage model* by Barlow and Proschan (1975). To determine the mean value of $Z(t)$, Wald's equation is used (see box).[2]

Wald's equation

Let $X_1, X_2, X_3, \ldots$ be independent and identically distributed random variables with finite mean $E(X)$. Further let N be a stochastic integer variable such that the event $(N = n)$ is independent of $X_{n+1}, X_{n+2}, \ldots$ for all $n = 1, 2, \ldots$. Then

$$E\left(\sum_{i=1}^{N} X_i\right) = E(N)\, E(X). \tag{10.22}$$

A proof of Wald's equation may be found, for example, in Ross (1996). The variance of $\sum_{i=1}^{N} X_i$ is Ross (1996):

$$\text{var}\left(\sum_{i=1}^{N} X_i\right) = E(N)\text{var}(X_i) + [E(X_i)]^2\text{var}(N). \tag{10.23}$$

[2] Named after the Hungarian mathematician Abraham Wald (1902–1950).

Let $E(V_i) = \nu$ and $\text{var}(V_i) = \tau^2$. From (10.22) and (10.23), we get

$$E[Z(t)] = \nu\lambda t \qquad \text{and} \qquad \text{var}[Z(t)] = \lambda(\nu^2 + \tau^2)t.$$

Assume now that the consequences V_i are all positive, that is $\Pr(V_i > 0) = 1$ for all i, and that a total item failure occurs as soon as $Z(t) > c$ for some specified critical value c. Let T_c be the time to item failure. Observe that $T_c > t$ if and only if $Z(t) \le c$. Let $V_0 = 0$, then

$$\begin{aligned}\Pr(T_c > t) &= \Pr(Z(t) \le c) = \Pr\left(\sum_{i=0}^{N(t)} V_i \le c\right) \\ &= \sum_{n=0}^{\infty} \Pr\left(\sum_{i=0}^{n} V_i \le c \mid N(t) = n\right) \frac{(\lambda t)^n}{n!}\, e^{-\lambda t} \\ &= \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!}\, e^{-\lambda t}\, F_V^{(n)}(c),\end{aligned} \tag{10.24}$$

where $F_V^{(n)}(v)$ is the distribution function of $\sum_{i=0}^{n} V_i$, and the last equality is due to the fact that $N(t)$ is independent of $V_1, V_2, \ldots$.

The mean time to total item failure is thus

$$\begin{aligned}E(T_c) &= \int_0^\infty \Pr(T_c > t)\, dt \\ &= \sum_{n=0}^{\infty} \left(\int_0^\infty \frac{(\lambda t)^n}{n!} e^{-\lambda t}\, dt\right) F_V^{(n)}(c) \\ &= \frac{1}{\lambda} \sum_{n=0}^{\infty} F_V^{(n)}(c).\end{aligned} \tag{10.25}$$

Example 10.4 (Exponentially distributed consequences)
Consider a sequence of failure events that can be described as an HPP $\{N(t), t \ge t\}$ with rate λ. Failure i has consequence V_i, where $V_1, V_2, \ldots$ are independent and exponentially distributed with parameter ρ. The sum $\sum_{i=1}^{n} V_i$ is therefore gamma distributed with parameters (n, ρ) (see Section 5.4):

$$F_V^{(n)}(v) = 1 - \sum_{k=0}^{n-1} \frac{(\rho v)^k}{k!} e^{-\rho v} = \sum_{k=n}^{\infty} \frac{(\rho v)^k}{k!} e^{-\rho v}$$

Total item failure occurs as soon as $Z(t) = \sum_{i=1}^{N(t)} V_i > c$. The mean time to total item failure is given by (10.15) where

$$\begin{aligned}\sum_{n=0}^{\infty} F_V^{(n)}(c) &= \sum_{n=0}^{\infty} \sum_{k=n}^{\infty} \frac{(\rho c)^k}{k!} e^{-\rho c} = \sum_{k=0}^{\infty} \sum_{n=0}^{k} \frac{(\rho c)^k}{k!} e^{-\rho c} \\ &= \sum_{k=0}^{\infty} (1 + k) \frac{(\rho c)^k}{k!} e^{-\rho c} = 1 + \rho c\end{aligned}$$

Hence, when the consequences $V_1, V_2, \dots$ are exponentially distributed with parameter ρ, the mean time to total item failure is

$$E(T_c) = \frac{1 + \rho c}{\lambda}. \tag{10.26}$$

□

The distribution of the time T_c to total item failure is by Barlow and Proschan (1975, p. 94) shown to be an increasing failure rate average (IFRA) distribution for *any* distribution $F_V(v)$. (IFRA distributions are discussed in Section 5.7).

10.3 Renewal Processes

Renewal theory had its origin in the study of strategies for replacement of technical items, but later it was developed as a general theory within stochastic processes. As the name of the process indicates, it is used to model *renewals*, or replacement of items. This section gives a summary of some main aspects of renewal theory that are of particular interest in reliability analysis. This includes formulas for calculation of exact availability and mean number of failures within a given time interval. The latter can, for example, be used to determine optimal allocation of spare parts.

Example 10.5 (A renewal process)
An item is put into operation and is functioning at time $t = 0$. When the item fails at time T_1, it is replaced by a new item of the same type, or restored to an as-good-as-new state. When this item fails at time $T_1 + T_2$, it is again replaced, and so on. The replacement time is assumed to be negligible. The times-to-failure $T_1, T_2, \dots$ are assumed to be independent and identically distributed. The number of failures, and *renewals*, in a time interval $(0, t]$ is denoted $N(t)$. □

10.3.1 Basic Concepts

A *renewal process* is a counting process $\{N(t), t \geq 0\}$ with interoccurrence times $T_1, T_2, \dots$ that are independent and identically distributed with distribution function $F_T(t) = \Pr(T_i \leq t)$ for $t \geq 0$ and $i = 1, 2, \dots$.

The events that are observed are called *renewals*, and $F_T(t)$ is called the underlying distribution of the renewal process. We assume that $E(T_i) = \mu$ and $\text{var}(T_i) = \sigma^2 < \infty$ for $i = 1, 2, 3, \dots$. Observe that the HPP discussed in Section 10.2 is a renewal process where the underlying distribution is exponential with parameter λ. A renewal process may thus be considered as a generalization of the HPP.

The concepts that are introduced for a general counting process in Section 10.1.2 are also relevant for a renewal process, but the theory of renewal processes has been developed as a specific theory, and many of the concepts have therefore been given specific names. We therefore list the main concepts of renewal processes and introduce the necessary terminology.

(1) The time until the nth renewal (the nth arrival time), S_n

$$S_n = T_1 + T_2 + \cdots + T_n = \sum_{i=1}^{n} T_i. \tag{10.27}$$

(2) The number of renewals in the time interval $(0, t]$

$$N(t) = \max \{n;\ S_n \le t\}. \tag{10.28}$$

(3) The renewal function

$$W(t) = E[N(t)]. \tag{10.29}$$

Thus $W(t)$ is the mean number of renewals in the time interval $(0, t]$.

(4) The renewal density

$$w(t) = \frac{d}{dt}\, W(t). \tag{10.30}$$

Observe that the renewal density coincides with the rate of the process defined in (10.2), which is called the rate of occurrence of failures (ROCOF) when the renewals are failures. The mean number of renewals in the time interval $(t_1, t_2]$ is

$$W(t_2) - W(t_1) = \int_{t_1}^{t_2} w(t)\, dt. \tag{10.31}$$

The relation between the renewal periods T_i and the number of renewals $N(t)$, the renewal process, is illustrated in Figure 10.1. The properties of renewal processes are discussed in detail by Cox (1962), Ross (1996), Cocozza-Thivent (1997), and Cha and Finkelstein (2018).

10.3.2 The Distribution of S_n

To find the exact distribution of the time to the nth renewal S_n is often complicated. We outline an approach that may be used, at least in some cases. Let $F^{(n)}(t)$ be the distribution function of $S_n = \sum_{i=1}^{n} T_i$.

Because S_n may be written as $S_n = S_{n-1} + T_n$, and S_{n-1} and T_n are independent, the distribution function of S_n is the *convolution* of the distribution functions of S_{n-1} and T_n, respectively,[3]

$$F^{(n)}(t) = \int_0^t F^{(n-1)}(t-x)\, dF_T(x). \tag{10.32}$$

[3] More information about *convolution* may be found on https://en.wikipedia.org/wiki/Convolution.

The convolution of two (time-to-failure) distributions F and G is often denoted $F * G$, meaning that $F * G(t) = \int_0^t G(t-x)\, dF(x)$. Equation (10.32) can therefore be written $F^{(n)} = F_T * F^{(n-1)}$.

When $F_T(t)$ is absolutely continuous[4] with probability density function $f_T(t)$, the probability density function $f^{(n)}(t)$ of S_n may be found from

$$f^{(n)}(t) = \int_0^t f^{(n-1)}(t-x) f_T(x)\, dx. \tag{10.33}$$

By successive integration of (10.33) for $n = 2, 3, 4, \dots$, the probability density of S_n for a specified value of n can, in principle, be found.

It may also sometimes be relevant to use Laplace transforms to find the distribution of S_n. The Laplace transform of Eq. (10.33) is (see Appendix B),

$$f^{*(n)}(s) = [f_T^*(s)]^n. \tag{10.34}$$

The probability density function of S_n can now, at least in principle, be determined from the inverse Laplace transform of (10.34).

In practice, it is often time-consuming and complicated to find the exact distribution of S_n from (10.33) and (10.34). Often, an approximation to the distribution of S_n is sufficient.

From the strong law of large numbers, that is, with probability 1,

$$\frac{S_n}{n} \to \mu \qquad \text{as} \qquad n \to \infty. \tag{10.35}$$

According to the central limit theorem (see Eq. 6.39), $S_n = \sum_{i=1}^n T_i$ is asymptotically normally distributed

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

and

$$F^{(n)}(t) = \Pr(S_n \le t) \approx \Phi\left(\frac{t - n\mu}{\sigma\sqrt{n}}\right). \tag{10.36}$$

where $\Phi(\cdot)$ is the distribution function of the standard normal distribution $\mathcal{N}(0,1)$.

Example 10.6 (IFR interoccurrence times)

Consider a renewal process where the interoccurrence times have an increasing failure rate (IFR) distribution $F_T(t)$ (see Section 5.6) with mean time-to-failure μ. In this case, Barlow and Proschan (1965, p. 27) show that the survivor function, $R_T(t) = 1 - F_T(t)$ satisfies

$$R_T(t) \ge e^{-t/\mu} \qquad \text{when} \qquad t < \mu. \tag{10.37}$$

[4] For a definition of the term *absolutely continuous*, e.g. see https://en.wikipedia.org/wiki/Absolute_continuity.

The right-hand side of (10.37) is the survivor function of a random variable U_j with exponential distribution with failure rate $1/\mu$. Let us assume that we have n independent random variable $U_1, U_2, \dots, U_n$ with the same distribution. The distribution of $\sum_{j=1}^{n} U_j$ has then a gamma distribution with parameters $(n, 1/\mu)$ (see Section 5.4.2), and we therefore get

$$\begin{aligned} 1 - F^{(n)}(t) &= \Pr(S_n > t) = \Pr(T_1 + T_2 + \cdots + T_n > t) \\ &\geq \Pr(U_1 + U_2 + \cdots + U_n > t) = \sum_{j=0}^{n-1} \frac{(t/\mu)^j}{j!}\, e^{-t/\mu} \end{aligned}$$

Hence,

$$F^n(t) \leq 1 - \sum_{j=0}^{n-1} \frac{(t/\mu)^j}{j!}\, e^{-t/\mu} \qquad \text{for} \qquad t < \mu. \tag{10.38}$$

For a renewal (failure) process where the interoccurrence times have an IFR distribution with mean μ, Eq. (10.38) provides a conservative bound for the probability that the nth failure occurs before time t, when $t < \mu$. □

10.3.3 The Distribution of $N(t)$

From the strong law of large numbers, that is, with probability 1,

$$\frac{N(t)}{t} \rightarrow \frac{1}{\mu} \qquad \text{as} \qquad t \rightarrow \infty. \tag{10.39}$$

When t is large, $N(t) \approx t/\mu$. This means that $N(t)$ is approximately a linear function of t when t is large. In Figure 10.7, the number of renewals $N(t)$ is plotted as a function of t for a simulated renewal process where the underlying distribution is Weibull with parameters $\lambda = 1$ and $\alpha = 3$.

From the definition of $N(t)$ and S_n, it follows that

$$\Pr(N(t) \geq n) = \Pr(S_n \leq t) = F^{(n)}(t),$$

and

$$\begin{aligned} \Pr(N(t) = n) &= \Pr(N(t) \geq n) - \Pr(N(t) \geq n + 1) \\ &= F^{(n)}(t) - F^{(n+1)}(t). \end{aligned} \tag{10.40}$$

For large values of n, we can apply (10.36) and obtain

$$\Pr(N(t) \leq n) \approx \Phi\left(\frac{(n+1)\mu - t}{\sigma}\right), \tag{10.41}$$

and

$$\Pr(N(t) = n) \approx \Phi\left(\frac{t - n\mu}{\sigma\sqrt{n}}\right) - \Phi\left(\frac{t - (n+1)\mu}{\sigma\sqrt{n+1}}\right), \tag{10.42}$$

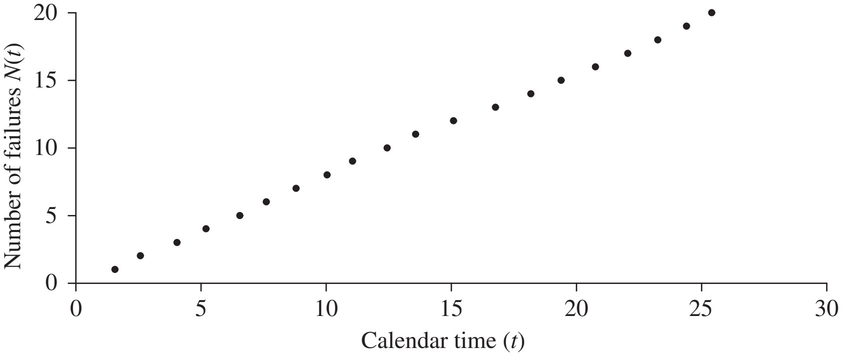


Figure 10.7 Number of renewals $N(t)$ as a function of t for a simulated renewal process where the underlying distribution is Weibull with parameters $\lambda = 1$ and $\alpha = 3$.

Takács (1956) derives the following alternative approximation formula which is valid when t is large

$$\Pr(N(t) \leq n) \approx \Phi\left(\frac{n - (t/\mu)}{\sigma\sqrt{t/\mu^3}}\right). \tag{10.43}$$

A proof of (10.43) is provided in Ross (1996, p. 109).

10.3.4 The Renewal Function

Because $N(t) \geq n$ if and only if $S_n \leq t$, we obtain (see Problem 10.5).

$$W(t) = E(N(t)) = \sum_{n=1}^{\infty} \Pr(N(t) \geq n) = \sum_{n=1}^{\infty} \Pr(S_n \leq t) = \sum_{n=1}^{\infty} F^{(n)}(t). \tag{10.44}$$

An integral equation for $W(t)$ may be obtained by combining (10.44) and (10.32):

$$\begin{aligned} W(t) &= F_T(t) + \sum_{r=2}^{\infty} F^{(r)}(t) = F_T(t) + \sum_{r=1}^{\infty} F^{(r+1)}(t) \\ &= F_T(t) + \sum_{r=1}^{\infty} \int_0^t F^{(r)}(t-x)\, dF_T(x) \\ &= F_T(t) + \int_0^t \sum_{r=1}^{\infty} F^{(r)}(t-x)\, dF_T(x) \\ &= F_T(t) + \int_0^t W(t-x)\, dF_T(x). \end{aligned} \tag{10.45}$$

This equation is known as the *fundamental renewal equation* and can sometimes be solved for $W(t)$.

Equation (10.44) can also be derived by a more direct argument. By conditioning on the time T_1 of the first renewal, we obtain

$$\begin{aligned} W(t) &= E[N(t)] = E[E(N(t) \mid T_1)] \\ &= \int_0^\infty E(N(t) \mid T_1 = x)\, dF_{T_1}(x), \end{aligned} \tag{10.46}$$

where

$$E(N(t) \mid T_1 = x) = \begin{cases} 0 & \text{when} \quad t < x \\ 1 + W(t-x) & \text{when} \quad t \geq x \end{cases}. \tag{10.47}$$

If the first renewal occurs at time x for $x \leq t$, the process starts over again from this point of time. The mean number of renewals in $(0, t]$ is thus 1 plus the mean number of renewals in $(x, t]$, which is $W(t - x)$.

Combining the two equations (10.46) and (10.47) yields

$$W(t) = \int_0^t [1 + W(t-x)]\, dF_T(x) = F_T(t) + \int_0^t W(t-x)\, dF_T(x),$$

and thereby an alternative derivation of (10.44) is provided.

The exact expression for the renewal function $W(t)$ is often difficult to determine from (10.44). Approximation formulas and bounds may therefore be useful.

Because $W(t)$ is the expected number of renewals in the interval $(0, t]$, the average length μ of each renewal is approximately $t/W(t)$. We should therefore expect that when $t \to \infty$, we get

$$\lim_{t\to\infty} \frac{W(t)}{t} = \frac{1}{\mu}. \tag{10.48}$$

This result is known as the *elementary renewal equation* and is valid for a general renewal process. A proof may, for example, be found in Ross (1996, p. 107).

When the renewals are item failures, the mean number of failures in $(0, t]$ is approximately

$$E[N(t)] = W(t) \approx \frac{t}{\mu} = \frac{t}{\text{MTBF}} \qquad \text{when } t \text{ is large.},$$

where $\mu = \text{MTBF}$ is the mean time between failures.

From the elementary renewal equation (10.48), the mean number of renewals in the interval $(0, t]$ is

$$W(t) \approx \frac{t}{\mu} \qquad \text{when } t \text{ is large.}$$

The mean number of renewals in the interval $(t, t + u]$ is

$$W(t+u) - W(t) \approx \frac{u}{\mu} \qquad \text{when } t \text{ is large, and} \qquad u > 0, \tag{10.49}$$

and the underlying distribution $F_T(t)$ is nonlattice. This result is known as *Blackwell's theorem*, and a proof may be found in Feller (1968).

Blackwell's theorem (10.49) has been generalized by Smith (1958), who shows that when the underlying distribution $F_T(t)$ is nonlattice, then

$$\lim_{t\to\infty} \int_0^t Q(t-x)\,dW(x) = \frac{1}{\mu}\int_0^\infty Q(u)\,du, \tag{10.50}$$

where Q(t) is a nonnegative, nonincreasing function which is Riemann integrable[5] over $(0, \infty)$. This result is known as the *key renewal equation.*

By introducing $Q(t) = \alpha^{-1}$ for $0 < t \le \alpha$ and $Q(t) = 0$ otherwise, in (10.50), we get Blackwell's theorem (10.49).

Let

$$F_e(t) = \frac{1}{\mu}\int_0^t [1 - F_T(u)]\,du, \tag{10.51}$$

where $F_e(t)$ is a distribution function with a special interpretation that is explained in Definition 10.6. By using $Q(t) = 1 - F_e(t)$ in (10.50) we get

$$\lim_{t\to\infty}\left(W(t) - \frac{t}{\mu}\right) = \frac{E(T_i^2)}{2\mu^2} - 1 = \frac{\sigma^2+\mu^2}{2\mu^2} - 1 = \frac{1}{2}\left(\frac{\sigma^2}{\mu^2} - 1\right),$$

if $E(T_i^2) = \sigma^2 + \mu^2 < \infty$. We may thus use the following approximation when t is large

$$W(t) \approx \frac{t}{\mu} + \frac{1}{2}\left(\frac{\sigma^2}{\mu^2} - 1\right). \tag{10.52}$$

Upper and lower bounds for the renewal function are supplied in Section 10.3.7.

10.3.5 The Renewal Density

When $F_T(t)$ has density $f_T(t)$, we may differentiate (10.45) and get

$$w(t) = \frac{d}{dt}W(t) = \frac{d}{dt}\sum_{n=1}^{\infty} F_T^{(n)}(t) = \sum_{n=1}^{\infty} f_T^{(n)}(t). \tag{10.53}$$

Equation (10.53) can sometimes be used to find the renewal density $w(t)$. Another approach is to differentiate (10.46) with respect to t

$$w(t) = f_T(t) + \int_0^t w(t-x) f_T(x)\,dx. \tag{10.54}$$

Yet another approach is to use Laplace transforms. From Appendix B, the Laplace transform of (10.54) is

$$w^*(s) = f_T^*(s) + w^*(s) f_T^*(s).$$

[5] For further information about Riemann integrable functions, e.g. see https://en.wikipedia.org/wiki/Riemann_integral. The Riemann integral is named after the German mathematician Georg Friedrich Bernhard Riemann (1826–1866).

Hence,

$$w^*(s) = \frac{f_T^*(s)}{1 - f_T^*(s)}. \tag{10.55}$$

Remark 10.2 According to (10.5), the probability of a failure (renewal) in a short interval $(t, t + \Delta t]$ is approximately $w(t)\Delta t$. Because the probability that the *first* failure occurs in $(t, t + \Delta t]$ is approximately $f_T(t)\Delta t$, we can use (10.54) to conclude that a "later" failure (i.e. not the first) occurs in $(t, t + \Delta t]$ with probability approximately equal to $\left(\int_0^t w(t-x)f_T(x)\,dx\right)\Delta t$. □

The exact expression for the renewal density $w(t)$ is often difficult to determine from (10.53) to (10.55). In the same way as for the renewal function, we therefore have to suffice with approximation formulas and bounds.

From (10.48), we should expect that

$$\lim_{t\to\infty} w(t) = \frac{1}{\mu}, \tag{10.56}$$

Smith (1958) shows that (10.56) is valid for a renewal process with underlying probability density function $f_T(t)$ when there exists a $p > 1$ such that $|f_T(t)|^p$ is Riemann integrable. The renewal density $w(t)$ therefore approaches the constant $1/\mu$ when t is large.

Consider a renewal process where the renewals are item failures. The interoccurrence times $T_1, T_2, \dots$ are then the times-to-failure, and $S_1, S_2, \dots$ are the times when the failures occur. Let $z(t)$ be the failure rate (force of mortality, FOM) function of the time to the first failure T_1. The conditional renewal density (ROCOF) $w_C(t)$ in the interval $(0, T_1)$ must be equal to $z(t)$. When the first failure has occurred, the item is renewed or replaced, and started up again with the same failure rate (FOM) as for the initial item. The conditional renewal rate (ROCOF) may then be expressed as

$$w_C(t) = z(t - S_{N(t-)}),$$

where $t - S_{N(t-)}$ is the time since the last failure, strictly before time t. The conditional ROCOF is illustrated in Figure 10.8 when the interoccurrence times are Weibull distributed with scale parameter $\lambda = 1$ and shape parameter $\alpha = 3$. The plot is based on simulated interoccurrence times from this distribution.

Example 10.7 (Gamma distributed renewal periods)
Consider a renewal process where the renewal periods $T_1, T_2, \dots$ are independent and gamma distributed with parameters $(2, \lambda)$, with probability density function

$$f_T(t) = \lambda^2 t\, e^{-\lambda t} \qquad \text{for} \qquad t > 0, \qquad \lambda > 0.$$

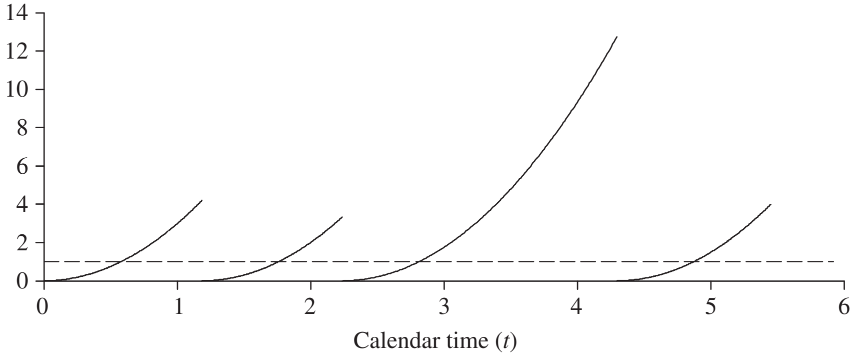


Figure 10.8 Illustration of the conditional ROCOF (fully drawn line) for simulated data from a Weibull distribution with parameters $\alpha = 3$ and $\lambda = 1$. The corresponding asymptotic renewal density is drawn by the dotted line.

The mean renewal period is $E(T_i) = \mu = 2/\lambda$, and the variance is $\text{var}(T_i) = \sigma^2 = 2/\lambda^2$. The time until the nth renewal, S_n, is gamma distributed (see Section 5.4.2) with probability density function

$$f^{(n)}(t) = \frac{\lambda}{(2n-1)!}\,(\lambda t)^{2n-1}\,e^{-\lambda t} \qquad \text{for} \quad t > 0.$$

The renewal density is according to (10.54)

$$\begin{aligned} w(t) &= \sum_{n=1}^{\infty} f^{(n)}(t) = \lambda e^{-\lambda t} \sum_{n=1}^{\infty} \frac{(\lambda t)^{2n-1}}{(2n-1)!} \\ &= \lambda e^{-\lambda t} \frac{e^{\lambda t} - e^{-\lambda t}}{2} = \frac{\lambda}{2}\,(1 - e^{-2\lambda t}) \end{aligned}$$

The renewal function is

$$W(t) = \int_0^t w(x)\,dx = \frac{\lambda}{2}\int_0^t (1 - e^{-2\lambda x})\,dx = \frac{\lambda t}{2} - \frac{1}{4}\,(1 - e^{-2\lambda t}). \tag{10.57}$$

□

The renewal density $w(t)$ and the renewal function $W(t)$ are illustrated in Figure 10.9 for $\lambda = 1$.

Observe that when $t \to \infty$, then

$$\begin{aligned} W(t) &\to \frac{\lambda t}{2} = \frac{t}{\mu} \\ w(t) &\to \frac{\lambda}{2} = \frac{1}{\mu} \end{aligned}$$

in accordance with (10.48) and (10.56), respectively. We may further use (10.52) to find a better approximation for the renewal function $W(t)$. From (10.56), we get the left-hand side of (10.52)

$$W(t) - \frac{t}{\mu} = W(t) - \frac{\lambda t}{2} \to -\frac{1}{4} \qquad \text{when} \qquad t \to \infty.$$

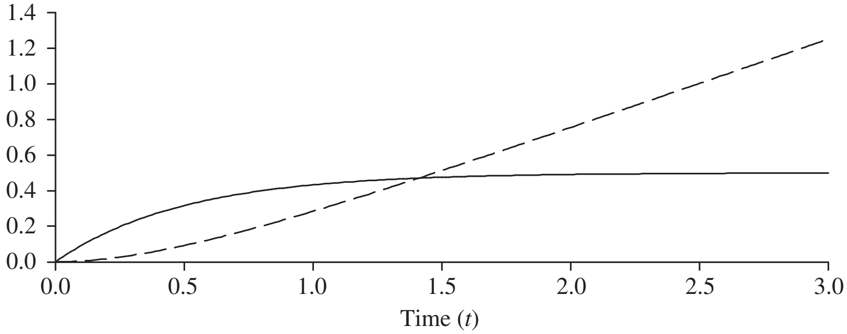


Figure 10.9 Renewal density $w(t)$ (fully drawn line) and renewal function $W(t)$ (dotted line) for Example 10.7, with ($\lambda = 1$).

The right-hand side of (10.52) is (with $\mu = 2/\mu$ and $\sigma^2 = 2/\lambda^2$)

$$\frac{t}{\mu} + \frac{1}{2}\left(\frac{\sigma^2}{2\mu^2} - 1\right) = \frac{t}{\mu} - \frac{1}{4}.$$

We can therefore use the approximation

$$W(t) \approx \frac{\lambda t}{2} - \frac{1}{4} \qquad \text{when } t \text{ is large.}$$

Example 10.8 (Weibull distributed renewal periods)
Consider a renewal process where the renewal periods $T_1, T_2, \ldots$ are independent and Weibull distributed with shape parameter α and scale parameter λ. In this case, the renewal function $W(t)$ cannot be deduced directly from (10.45). Smith and Leadbetter (1963) show that $W(t)$ can be expressed as an infinite, absolutely convergent series where the terms can be found by a simple recursive procedure. They show that $W(t)$ can be written

$$W(t) = \sum_{k=1}^{\infty} \frac{(-1)^{k-1} A_k (\lambda t)^{k\alpha}}{\Gamma(k\alpha + 1)}. \tag{10.58}$$

By introducing this expression for $W(t)$ in the fundamental renewal equation, the constants A_k; $k = 1, 2, \ldots$ can be determined. The calculation, which is quite comprehensive, leads to the following *recursion* formula:

$$\begin{aligned} A_1 &= \gamma_1 \\ A_2 &= \gamma_2 - \gamma_1 A_1 \\ A_3 &= \gamma_3 - \gamma_1 A_2 - \gamma_2 A_1 \\ &\vdots \end{aligned} \tag{10.59}$$

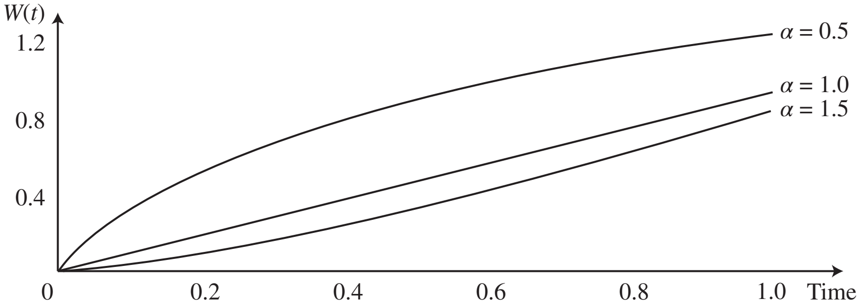


Figure 10.10 The renewal function for Weibull distributed renewal periods with $\lambda = 1$ and $\alpha = 0.5$, $\alpha = 1$, and $\alpha = 1.5$. Source: The figure is adapted from Smith and Leadbetter (1963).

$$A_n = \gamma_n - \sum_{j=1}^{n-1} \gamma_j A_{n-j}$$

$$\vdots$$

where

$$\gamma_n = \frac{\Gamma(n\alpha + 1)}{n!} \qquad \text{for} \qquad n = 1, 2, \dots .$$

For $\alpha = 1$, the Weibull distribution is an exponential distribution with parameter λ. In this case,

$$\gamma_n = \frac{\Gamma(n + 1)}{n!} = 1 \qquad \text{for } n = 1, 2, \dots .$$

This leads to

$$A_1 = 1$$

$$A_n = 0 \quad \text{for} \quad n \geq 2$$

The renewal function is thus according to (10.58)

$$W(t) = \frac{(-1)^0 A_1 \lambda t}{\Gamma(2)} = \lambda t.$$

The renewal function $W(t)$ is illustrated in Figure 10.10 for $\lambda = 1$ and three values of α. □

10.3.6 Age and Remaining Lifetime

The *age* $Z(t)$ of an item which is operating at time t is defined as

$$Z(t) = \begin{cases} t & \text{for} \quad N(t) = 0 \\ t - S_{N(t)} & \text{for} \quad N(t) > 0 \end{cases} . \tag{10.60}$$

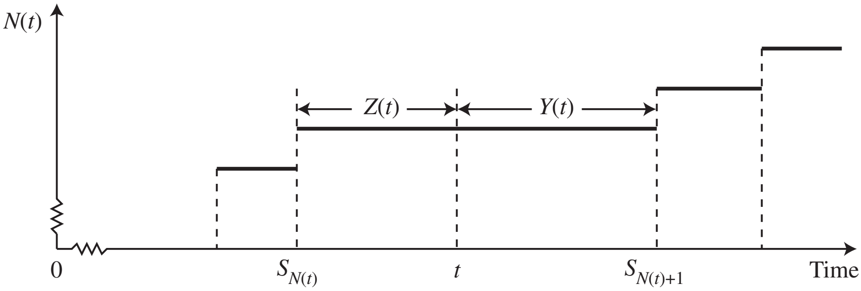


Figure 10.11 The age $Z(t)$ and the remaining lifetime $Y(t)$.

The *remaining lifetime* $Y(t)$ of an item that is in operation at time t is given as

$$Y(t) = S_{N(t)+1} - t. \tag{10.61}$$

The age $Z(t)$ and the remaining lifetime $Y(t)$ are illustrated in Figure 10.11. The remaining lifetime is also called the residual lifetime, the excess life, or the forward recurrence time (e.g. see Ross 1996).

Observe that $Y(t) > y$ means that there is no renewal in the time interval $(t, t+y]$.

Consider a renewal process where the renewals are item failures, and let T be the time from start-up to the first failure. The distribution of the remaining life $Y(t)$ of the item at time t is given by

$$\Pr(Y(t) > y) = \Pr(T > y + t \mid T > t) = \frac{\Pr(T > y + t)}{\Pr(T > t)},$$

and the mean remaining lifetime at time t is

$$E[Y(t)] = \frac{1}{\Pr(T > t)} \int_t^\infty \Pr(T > u)\, du.$$

See also Section 5.3.6, where $E[Y(t)]$ is called the mean residual lifetime (MRL) at time t. When T is exponentially distributed with failure rate λ, the mean remaining lifetime at time t is $1/\lambda$ which is an obvious result because of the memoryless property of the exponential distribution.

Limiting Distribution

Consider a renewal process with a nonlattice underlying distribution $F_T(t)$. We observe the process at time t. The time till the next failure is the remaining lifetime $Y(t)$. The limiting distribution of $Y(t)$ when $t \to \infty$ is (e.g. see Ross 1996, p. 116)

$$\lim_{t\to\infty} \Pr(Y(t) \le t) = F_e(t) = \frac{1}{\mu} \int_0^t [1 - F_T(u)]\, du, \tag{10.62}$$

which is the same distribution as we used in (10.52). The mean of the limiting distribution $F_e(t)$ of the remaining lifetime is

$$\begin{aligned} E(Y) &= \int_0^\infty \Pr(Y > y)\, dy = \int_0^\infty [1 - F_e(y)]\, dy \\ &= \frac{1}{\mu}\int_0^\infty \int_y^\infty \Pr(T > t)\, dt\, dy = \frac{1}{\mu}\int_0^\infty \int_0^t \Pr(T > t)\, dy\, dt \\ &= \frac{1}{\mu}\int_0^\infty t \Pr(T > t)\, dt = \frac{1}{2\mu}\int_0^\infty \Pr(T > \sqrt{x})\, dx \\ &= \frac{1}{2\mu}\int_0^\infty \Pr(T^2 > x)\, dx = \frac{E(T^2)}{2\mu} = \frac{\sigma^2 + \mu^2}{2\mu} \end{aligned}$$

where $E(T) = \mu$ and $\text{var}(T) = \sigma^2$, and we assume that $E(T^2) = \sigma^2 + \mu^2 < \infty$.

We have thus shown that the limiting mean remaining life is

$$\lim_{t\to\infty} E[Y(t)] = \frac{\sigma^2 + \mu^2}{2\mu}. \tag{10.63}$$

Example 10.9 (Example 10.7 (cont.))

Again, consider the renewal process in Example 10.7 where the underlying distribution is gamma distributed with parameters $(2, \lambda)$, with mean time between renewals $E(T_i) = \mu = 2/\lambda$ and variance $\text{var}(T_i) = 2/\lambda^2$. The mean remaining life of an item that is in operation at time t far from now is from (10.62)

$$E[Y(t)] \approx \frac{\sigma^2 + \mu^2}{2\mu} = \frac{3}{2\lambda} \qquad \text{when } t \text{ is large.}$$

□

The distribution of the age $Z(t)$ of an item that is in operation at time t can be derived by starting with

$$\begin{aligned} Z(t) > z &\iff \text{no renewals in } (t - z, t) \\ &\iff Y(t - z) > z \end{aligned}$$

Therefore,

$$\Pr(Z(t) > z) = \Pr(Y(t - z) > z).$$

When the underlying distribution $F_T(t)$ is nonlattice, we can show that the limiting distribution of the age $Z(t)$ when $t \to \infty$ is

$$\lim_{t\to\infty} \Pr(Z(t) \le t) = F_e(t) = \frac{1}{\mu}\int_0^t [1 - F_T(u)]\, du, \tag{10.64}$$

that is, the same distribution as (10.62). When $t \to \infty$, both the remaining lifetime $Y(t)$ and the age $Z(t)$ at time t have the same distribution. When t is large, then

$$E[Y(t)] \approx E[Z(t)] \approx \frac{\sigma^2 + \mu^2}{2\mu}. \tag{10.65}$$

Assume that a renewal process with a nonlattice underlying distribution has been "running" for a long time, and that the process is observed at a random time, which we denote $t = 0$. The time T_1 to the first renewal after time $t = 0$ is equal to the remaining lifetime of the item that is in operation at time $t = 0$. The distribution of T_1 is equal to (10.62) and the mean time to the first renewal is given by (10.63). Similarly, the age of the item that is in operation at time $t = 0$ has the same distribution and the same mean as the time to the first renewal. For a formal proof, see Ross (1996).

Remark 10.3 This result may seem a bit strange. When we observe a renewal process that has been "running" for a long time at a random time t, the length of the corresponding interoccurrence time is $S_{N(t)+1} - S_{N(t)}$, as illustrated in Figure 10.11, and the mean length of the interoccurrence time is μ. We obviously have that $S_{N(t)+1} - S_{N(t)} = Z(t) + Y(t)$, but $E[Z(t) + E(Y(t)] = (\sigma^2 + \mu^2)/\mu$ is greater than μ. This rather surprising result is known as the *inspection paradox*, and is further discussed by Ross (1996). □

If the underlying distribution function $F_T(t)$ is new better than used (NBU) or new worse than used (NWU) (see Section 5.6.3), bounds may be derived for the distribution of the remaining lifetime $Y(t)$ of the item that is in operation at time t. Barlow and Proschan (1975) show that the following apply:

$$\text{If} \quad F_T(t) \quad \text{is NBU, then} \quad \Pr(Y(t) > y) \le \Pr(T > y), \tag{10.66}$$

$$\text{If} \quad F_T(t) \quad \text{is NWU, then} \quad \Pr(Y(t) > y) \ge \Pr(T > y). \tag{10.67}$$

Intuitively, these results are obvious. If an item has an NBU life distribution, then a new item should have a higher probability of surviving the interval $(0, y]$ than a used item. The opposite should apply for an item with an NWU life distribution.

When the distributions of $Z(t)$ and $Y(t)$ are to be determined, the following lemma is useful:

Lemma 10.1 If

$$g(t) = h(t) + \int_0^t g(t-x)\, dF(x), \tag{10.68}$$

where the functions h and F are known and g is unknown, then

$$g(t) = h(t) + \int_0^t h(t-x)\, dW_F(x), \tag{10.69}$$

where

$$W_F(x) = \sum_{r=1}^{\infty} F^{(r)}(x).$$

Observe that Eq. (10.69) is a generalization of the fundamental renewal Eq. (10.45).

Example 10.10 Consider a renewal process with underlying distribution $F_T(t)$. The distribution of the remaining lifetime $Y(t)$ of an item that is in operation at time t can be given by (e.g. see Bon 1995, p. 129)

$$\Pr(Y(t) > y) = \Pr(T > y + t) + \int_0^t \Pr(T > y + t - u)\; dW_F(u). \tag{10.70}$$

By introducing the survivor function $R(t) = 1 - F_T(t)$, and assuming that the renewal density $w_F(t) = dW_F(t)/dt$ exists, (10.69) may be written

$$\Pr(Y(t) > y) = R(y + t) + \int_0^t R(y + t - u) w_F(u)\; du. \tag{10.71}$$

If the probability density function $f(t) = dF_T(t)/dt = -dR(t)/dt$ exists, we have from the definition of $f(t)$ that

$$R(t) - R(t + y) \approx f(t)y \qquad \text{when } y \text{ is small.}$$

Equation (10.70) may in this case be written

$$\begin{aligned}\Pr(Y(t) > y) &\approx R(t) - f(t)y + \int_0^t [R(t-u) - f(t-u)y]\; w_F(u)\; du \\ &= R(t) + \int_0^t R(t-u)\; w_F(u)\; du \\ &\quad - y\left(f(t) + \int_0^t f(t-u)\; w_F(u)\; du\right) \\ &= \Pr(Y(t) > 0) - w_F(t)y. \end{aligned} \tag{10.72}$$

The last line in (10.72) follows from Lemma 10.1. Because $\Pr(Y(t) > 0) = 1$, we have the following approximation

$$\Pr(Y(t) > y) \approx 1 - w_F(t)y \qquad \text{when } y \text{ is small.} \tag{10.73}$$

If we observe a renewal process at a random time t, the probability of having a failure (renewal) in a short interval of length y after time t is, from (10.73), approximately $w_F(t)y$, and it is hence relevant to call $w_F(t)$ the rate of occurrence of failures (ROCOF). □

10.3.7 Bounds for the Renewal Function

We will now establish some bound for the renewal function $W(t)$. For this purpose, consider a renewal process with interarrival times $T_1, T_2, \ldots$. We stop observing the process at the first renewal after time t, that is, at renewal $N(t) + 1$. Because the event $N(t) + 1 = n$ only depends on $T_1, T_2, \dots , T_n$, we can use Wald's equation to get

$$E(S_{N(t)+1}) = E\left(\sum_{i=1}^{N(t)+1} T_i \right) = E(T)\ E[N(t) + 1] = \mu\ [W(t) + 1]. \tag{10.74}$$

Because $S_{N(t)+1}$ is the first renewal after t, it can be expressed as

$$S_{N(t)+1} = t + Y(t).$$

The mean value is from (10.74)

$$\mu\ [W(t) + 1] = t + E[Y(t)],$$

such that

$$W(t) = \frac{t}{\mu} + \frac{E[Y(t)]}{\mu} - 1. \tag{10.75}$$

When t is large and the underlying distribution is nonlattice, we can use (10.63) to get

$$W(t) - \frac{t}{\mu} \ \to\ \frac{1}{2}\left(\frac{\sigma^2}{\mu^2} - 1\right) \qquad \text{when} \qquad t \to \infty, \tag{10.76}$$

which is the same result as we found in (10.52).

Lorden (1970) shows that the renewal function $W(t)$ of a *general* renewal process is bounded by

$$\frac{t}{\mu} - 1 \le W(t) \le \frac{t}{\mu} + \frac{\sigma^2}{\mu^2}. \tag{10.77}$$

For a proof, see Cocozza-Thivent (1997).

Several families of life distributions are introduced in Section 5.4. A distribution is said to be "new better than used in expectation" (NBUE) when the mean remaining lifetime of a used item is less, or equal to the mean life of a new item. In the same way, a distribution is said to be "new worse than used in expectation" (NWUE) when the mean remaining life of a used item is greater, or equal to the mean life of a new item.

For an NBUE distribution, $E[Y(t)] \le \mu$, and

$$W(t) = \frac{t + E[Y(t)]}{\mu} - 1 \le \frac{t}{\mu} \qquad \text{for} \qquad t \ge 0,$$

and

$$\frac{t}{\mu} - 1 \le W(t) \le \frac{t}{\mu}. \tag{10.78}$$

If we have an NWUE distribution, then $E[Y(t)] \ge \mu$, and

$$W(t) = \frac{t + E[Y(t)]}{\mu} - 1 \ge \frac{t}{\mu} \qquad \text{for} \qquad t \ge 0. \tag{10.79}$$

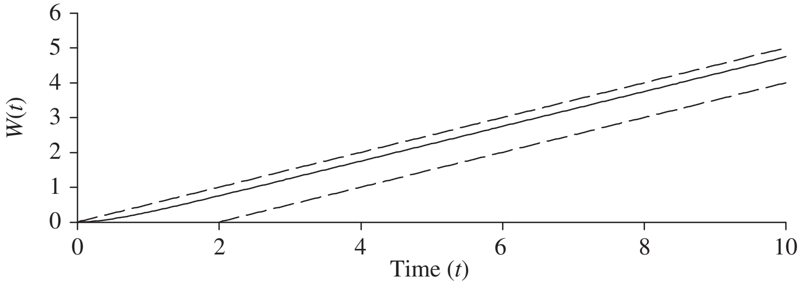


Figure 10.12 The renewal function $W(t)$ of a renewal process with underlying distribution that is gamma(2, λ), together with the bounds for $W(t)$, for $\lambda = 1$.

Further bounds for the renewal function are given by Dohi et al. (2002).

Example 10.11 (Example 10.5 (cont.))
Reconsider the renewal process where the underlying distribution has a gamma distribution with parameters (2, λ). This distribution has an IFR, and is therefore also NBUE. We can therefore apply the bounds in (10.78). In Figure 10.12 the renewal function (10.57)

$$W(t) = \frac{\lambda t}{2} - \frac{1}{4}(1 - e^{-2\lambda t}),$$

is plotted together with the bounds in (10.78)

$$\frac{\lambda t}{2} - 1 \le W(t) \le \frac{\lambda t}{2}.$$

□

10.3.8 Superimposed Renewal Processes

Consider a series structure of n independent items that are put into operation at time $t = 0$. All the n items are assumed to be new at time $t = 0$. When an item fails, it is replaced with a new item of the same type, or restored to an as-good-as-new condition. Each item therefore generates a renewal process. The n items are generally different, and the renewal processes therefore have different underlying distributions.

The process formed by the union of all the failures is called a *superimposed renewal process* (SRP). The n individual renewal processes and the SRP are illustrated in Figure 10.13.

In general, the SRP will *not* be a renewal process, but it has been shown, for example, by Drenick (1960), that superposition of an infinite number of independent *stationary* renewal processes is an HPP. Many items are composed of a large number of items in series and therefore, Drenick's result is often used as

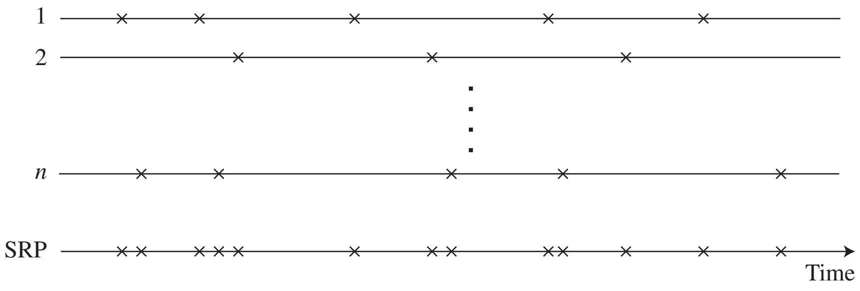


Figure 10.13 Superimposed renewal process.

a justification for assuming the time between item failures to be exponentially distributed.

Example 10.12 (Series structure)
Consider a series structure of two items. When an item fails, it is replaced or repaired to an as-good-as-new condition. Each item therefore generates an ordinary renewal process. The time required to replace or repair an item is considered to be negligible, and the items are assumed to fail and be repaired independent of each other. Both items are put into operation and are functioning at time $t = 0$. The series structure fails as soon as one of its items fails, and the structure failures produce a SRP. Times-to-failure for selected life distributions with IFRs for the two items and the series structure have been simulated on a computer and are shown in Figure 10.14. The conditional ROCOF (when the failure times are given) is also shown. Figure 10.14 further shows that the structure is not restored to an as-good-as-new state after each structure failure. The structure is subject to *imperfect repairs* (see Section 10.5) and the process of structure failures is not a renewal process because the times between structure failures do not have a common distribution. □

The SRP is further discussed, for example, by Cox and Isham (1980) and Ascher and Feingold (1984).

10.3.9 Renewal Reward Processes

Consider a renewal process $\{N(t), t \geq 0\}$, and let $(S_{i-1}, S_i]$ be the duration of the ith renewal cycle, with interoccurrence time $T_i = S_i - S_{i-1}$. Let V_i be a reward associated with renewal T_i, for $i = 1, 2, \ldots$. The rewards $V_1, V_2, \ldots$ are assumed to be independent random variables with the common distribution function $F_V(v)$, and with $E(T_i) < \infty$. This model is comparable with the compound Poisson process

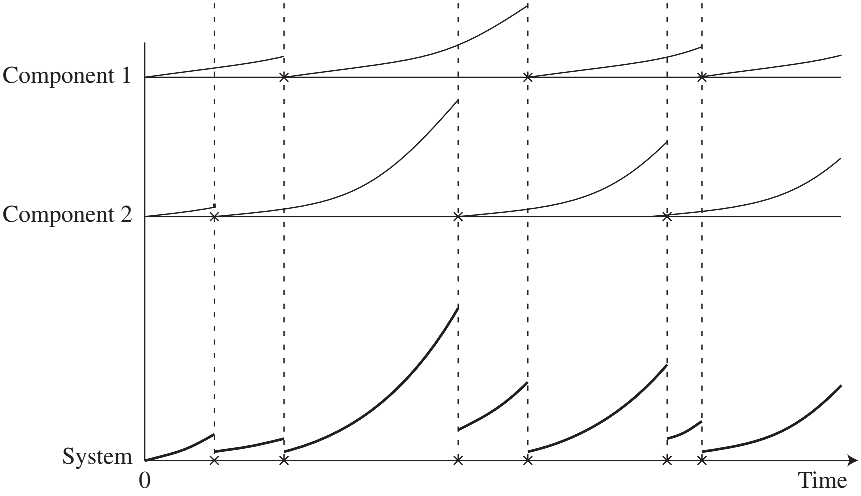


Figure 10.14 Superimposed renewal process. Conditional ROCOF $w_C(t)$ of a series structure of two items that are renewed upon failure.

that is described in Section 10.2.6. The accumulated reward in the time interval $(0, t]$ is

$$V(t) = \sum_{i=1}^{N(t)} V_i. \tag{10.80}$$

Let $E(T_i) = \mu_T$ and $E(V_i) = \mu_V$. According to Wald's equation (10.22), the mean accumulated reward is

$$E[V(t)] = \mu_V E[N(t)]. \tag{10.81}$$

According to the elementary renewal equation (10.48), when $t \to \infty$,

$$\frac{W(t)}{t} = \frac{E[N(t)]}{t} \to \frac{1}{\mu_T}.$$

Hence

$$\frac{E[V(t)]}{t} = \frac{\mu_V E[N(t)]}{t} \to \frac{\mu_V}{\mu_T}. \tag{10.82}$$

The same result is true even if the reward V_i is allowed to depend on the associated interoccurrence time T_i for $i = 1, 2, \ldots$. The pairs (T_i, V_i) for $i = 1, 2, \ldots$ are assumed to be independent and identically distributed (for proof, see Ross 1996). The reward V_i in renewal cycle i may, for example, be a function of the interoccurrence time T_i, for $i = 1, 2, \ldots$. When t is very large, then

$$V(t) \approx \mu_V \frac{t}{\mu_T},$$

which is an obvious result.

10.3.10 Delayed Renewal Processes

Sometimes, the first interoccurrence time T_1 has a distribution function $F_{T_1}(t)$ that is different from the distribution function $F_T(t)$ of the subsequent interoccurrence times. This may, for example, be the case for a failure process where the item at time $t = 0$ is not new. Such a renewal process is called a *delayed* renewal process, or a *modified* renewal process. To specify that the process is not delayed, we sometimes say that we have an *ordinary* renewal process.

Several of the results presented earlier in this section can be easily extended to delayed renewal processes.

The Distribution of *N*(*t*)

Analogous with (10.40) we get

$$\Pr(N(t) = n)^* = F_{T_1}^* * F_T^{*(n-1)} - F_{T_1}^* * F_T^{*(n)}. \tag{10.83}$$

The Distribution of S_n

The Laplace transform of the density of S_n is from (10.34)

$$f^{*(n)}(s) = f_{T_1}^*(s)[f_T^*(s)]^{n-1}. \tag{10.84}$$

The Renewal Function

The integral equation (10.45) for the renewal function $W(t)$ becomes

$$W(t) = F_{T_1}(t) + \int_0^t W(t-x)\, dF_T(x), \tag{10.85}$$

and the Laplace transform is

$$W^*(s) = \frac{f_{T_1}^*(s)}{s(1 - f_T^*(s))}. \tag{10.86}$$

The Renewal Density

Analogous with (10.54) we get

$$w(t) = f_{T_1}(t) + \int_0^t w(t-x) f_T(x)\, dx, \tag{10.87}$$

and the Laplace transform is

$$w^*(s) = \frac{f_{T_1}^*(s)}{1 - f_T^*(s)}. \tag{10.88}$$

All the limiting properties for ordinary renewal processes, when $t \to \infty$, will obviously also apply for delayed renewal processes.

For more detailed results see, for example, Cocozza-Thivent (1997). We will briefly discuss a special type of a delayed renewal process, the *stationary* renewal process.

Definition 10.6 (Stationary renewal process)
A stationary renewal process is a delayed renewal process where the first renewal period has distribution function

$$F_{T_1}(t) = F_e(t) = \frac{1}{\mu} \int_0^t [1 - F_T(x)]\, dx, \tag{10.89}$$

whereas the underlying distribution of the other renewal periods is $F_T(t)$. □

Remark 10.4

(1) Observe that $F_e(t)$ is the same distribution function we found in (10.63)
(2) When the probability density function $f_T(t)$ of $F_T(t)$ exists, the density of $F_e(t)$ is

$$f_e(t) = \frac{dF_e(t)}{dt} = \frac{1 - F_T(t)}{\mu} = \frac{R_T(t)}{\mu}.$$

(3) Cox (1962) shows that the stationary renewal process has a simple physical interpretation: Suppose a renewal process is started at time $t = -\infty$, but that the process is not observed before time $t = 0$. Then the first renewal period observed, T_1, is the remaining lifetime of the item in operation at time $t = 0$. According to (10.63), the distribution function of T_1 is $F_e(t)$. A stationary renewal process is called an *equilibrium renewal process* by Cox (1962). This is the reason why we use the subscript e in $F_e(t)$. Ascher and Feingold (1984) call the stationary renewal process a renewal process with *asynchronous sampling*, whereas an ordinary renewal process is called a renewal process with *synchronous sampling*.

□

Let $\{N_S(t), t \geq 0\}$ be a stationary renewal process, and let $Y_S(t)$ be the remaining life of an item at time t. The stationary renewal process has the following properties (Ross 1996):

$$W_S(t) = t/\mu, \tag{10.90}$$

$$\Pr(Y_S(t) \leq y = F_e(y), \qquad \text{for } t \geq 0 \tag{10.91}$$

$$\{N_S(t), t \geq 0\} \qquad \text{has stationary increments,} \tag{10.92}$$

where $F_e(y)$ is defined by Eq. (10.89).

Remark 10.5 An HPP is a stationary renewal process because of the memoryless property of exponential distribution. The HPP is seen to fulfill all the three properties (10.90), (10.91), and (10.92). □

Example 10.13 Reconsider the renewal process in Example 10.5 where the interoccurrence times are gamma distributed with parameters $(2, \lambda)$. The underlying distribution function is then

$$F_T(t) = 1 - e^{-\lambda t} - \lambda t\, e^{-\lambda t},$$

and the mean interoccurrence time is $E(T_i) = 2/\lambda$. Let us now assume that the process has been running for a long time and that when we start observing the process at time $t = 0$, it may be considered as a stationary renewal process.

According to (10.90), the renewal function for this stationary renewal process is $W_S(t) = \lambda t/2$, and the distribution of the remaining life, $Y_S(t)$ is (see Eq. (10.91)),

$$\begin{aligned}\Pr(Y_S(t) \le y) &= \frac{\lambda}{2} \int_0^y (e^{-\lambda u} + \lambda u\, e^{-\lambda u})\, du \\ &= 1 - \left(1 + \frac{\lambda y}{2}\right) e^{-\lambda y}\end{aligned}$$

The mean remaining lifetime of an item at time t is

$$E[Y_S(t)] = \int_0^\infty \Pr(Y_S(t) > y)\, dy = \int_0^\infty \left(1 + \frac{\lambda y}{2}\right) e^{-\lambda y}\, dy = \frac{3}{2\lambda}.$$

□

Delayed renewal processes are used in Section 10.3.11 to analyze alternating renewal processes.

10.3.11 Alternating Renewal Processes

Consider an item that is activated and is functioning at time $t = 0$. Whenever the item fails, it is repaired. Let $U_1, U_2, \ldots$ denote the successive times-to-failure (up-times) of the item. Let us assume that the times-to-failure are independent and identically distributed with distribution function $F_U(t) = \Pr(U_i \le t)$ and mean $E(U) = \text{MTTF}$. Likewise, assume the corresponding downtimes $D_1, D_2, \ldots$ to be independent and identically distributed with distribution function $F_D(d) = \Pr(D_i \le d)$ and mean $E(D) = \text{MDT}$. MDT is the total mean downtime following a failure, and will usually involve much more that the active repair time.[6]

[6] In the rest of this book, we are using T to denote time-to-failure. In this chapter, we have already used T to denote interoccurrence time (renewal period), and we will therefore use U to denote the time-to-failure (up-time) in this section. We hope that this does not confuse the reader.

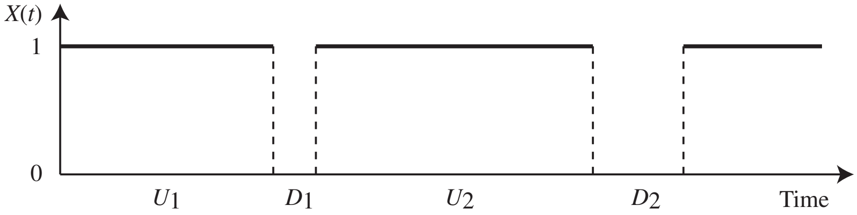


Figure 10.15 Alternating renewal process.

If we define the completed repairs to be the renewals, we obtain an ordinary renewal process with renewal periods (interoccurrence times) $T_i = U_i + D_i$ for $i = 1, 2, \ldots$. The mean time between renewals is $\mu_T = \text{MTTF} + \text{MDT}$. The resulting process is called an *alternating renewal process* and is shown in Figure 10.15.

The underlying distribution function, $F_T(t)$, is the convolution of the distribution functions $F_U(t)$ and $F_D(t)$,

$$F_T(t) = \Pr(T_i \le t) = \Pr(U_i + D_i \le t) = \int_0^t F_U(t - x)\, dF_D(x). \tag{10.93}$$

If instead, we let the renewals be the events when a "failure" occurs and start observing the item at a renewal, we get a *delayed* renewal process where the first renewal period T_1 is equal to U_1 whereas $T_i = D_{i-1} + U_i$ for $i = 2, 3, \ldots$.

In this case, the distribution function $F_{T_1}(t)$ of the first renewal period is given by

$$F_{T_1}(t) = \Pr(T_1 \le t) = \Pr(U_1 \le t) = F_U(t), \tag{10.94}$$

whereas the distribution function $F_T(t)$ of the other renewal periods is given by (10.83).

Example 10.14 Consider the alternating renewal process described above, and let the renewals be the completed repairs such that we have an ordinary renewal process. Let a reward V_i be associated with the ith interoccurrence time, and assume that this reward is defined such that we earn one unit per unit of time the item is functioning in the time period since the last failure. When the reward is measured in time units, then $E(V_i) = \mu_V = \text{MTTF}$. The average availability $A_{\text{av}}(0, t)$ of the item in the time interval $(0, t)$ has been defined as the mean fraction of time in the interval $(0, t)$ where the item is functioning. From (10.82), we therefore get

$$A_{\text{av}}(0, t) \to \frac{\mu_V}{\mu_T} = \frac{\text{MTTF}}{\text{MTTF} + \text{MDT}} \qquad \text{when} \qquad t \to \infty, \tag{10.95}$$

which is the same result we obtained in Section 6.5.1 based on heuristic arguments. □

Availability

The availability $A(t)$ of an item was defined as the probability that the item is functioning at time t, that is, $A(t) = \Pr(X(t) = 1)$, where $X(t)$ is the state variable of the item.

As above, consider an alternating renewal process where the renewals are completed repairs, and let $T = U_1 + D_1$. The availability of the item is then

$$A(t) = \Pr(X(t) = 1) = \int_0^\infty \Pr(X(t) = 1 \mid T = x)\, dF_T(x).$$

Because the item is assumed to be as-good-as-new at time $T = U_1 + D_1$, the process repeats itself from this point of time and

$$\Pr(X(t) = 1 \mid T = x) = \begin{cases} A(t-x) & \text{for } t > x \\ \Pr(U_1 > t \mid T = x) & \text{for } t \le x \end{cases}.$$

Therefore,

$$A(t) = \int_0^t A(t-x)\, dF_T(x) + \int_t^\infty \Pr(U_1 > t \mid T = x)\, dF_T(x),$$

but because $D_1 > 0$, then

$$\begin{aligned} \int_t^\infty \Pr(U_1 > t \mid U_1 + D_1 = x)\, dF_T(x) &= \int_0^\infty \Pr(U_1 > t \mid T = x)\, dF_T(x) \\ &= \Pr(U_1 > t) = 1 - F_U(t) \end{aligned}$$

Hence,

$$A(t) = 1 - F_U(t) + \int_0^t A(t-x)\, dF_T(x). \tag{10.96}$$

We apply Lemma 10.1 and get

$$A(t) = 1 - F_T(t) + \int_0^t [1 - F_T(t-x)]\, dW_{F_T}(x), \tag{10.97}$$

where

$$W_{F_T}(t) = \sum_{n=1}^\infty F_T^{(n)}(t),$$

is the renewal function for a renewal process with underlying distribution $F_T(t)$.

When $F_U(t)$ is a nonlattice distribution, the key renewal equation (10.50) can be used with $Q(t) = 1 - F_U(t)$ and we get

$$\int_0^t [1 - F_U(t-x)]\, dW_{F_T}(x) \to t \to \infty \frac{1}{E(T)} \int_0^\infty [1 - F_U(t)]\, dt = \frac{E(U)}{E(U) + E(D)}.$$

Because $F_T(t) \to 1$ when $t \to \infty$, we have thus shown that

$$A = \lim_{t\to\infty} A(t) = \frac{E(U)}{E(U) + E(D)} = \frac{\text{MTTF}}{\text{MTTF} + \text{MDT}}. \tag{10.98}$$

Observe that this is the same result as we got in (10.95) by using results from renewal reward processes.

Example 10.15 (Parallel structure)
Consider a parallel structure of n items that fail and are repaired independent of each other. Item i has a time-to-failure (up-time) U_i that is exponentially distributed with failure rate λ_i, and downtime D_i that is also exponentially distributed with (repair) rate μ_i, for $i = 1, 2, \ldots$. The parallel structure fails when all the n items are in a failed state at the same time. Because the items are assumed to be independent, a parallel structure failure must occur in the following way: Just prior to the last item failure, $(n-1)$ items must be in a failed state, and then the functioning item must fail.

Let us now assume that the parallel structure has been in operation for a long time, such that we can use limiting (average) availabilities. The probability that item i is in a failed state is then approximately:

$$\overline{A}_i \approx \frac{\text{MDT}}{\text{MTTF} + \text{MDT}} = \frac{1/\mu_i}{1/\lambda_i + 1/\mu_i} = \frac{\lambda_i}{\lambda_i + \mu_i}.$$

Similarly, the probability that item i is functioning is approximately:

$$A_i \approx \frac{\mu_i}{\lambda_i + \mu_i}.$$

The probability that a functioning item i will fail within a very short time interval of length Δt is approximately:

$$\Pr(\Delta t) \approx \lambda_i \, \Delta t.$$

The probability of parallel structure failure in the interval $(t, t + \Delta t)$, when t is large is,

$$\begin{aligned}\Pr[\text{Structure failure in } (t, t+\Delta t)] &= \sum_{i=1}^{n} \left[\frac{\mu_i}{\lambda_i + \mu_i} \prod_{j \neq i} \frac{\lambda_j}{\lambda_j + \mu_j} \right] \lambda_i \, \Delta t + o(\Delta t) \\ &= \sum_{i=1}^{n} \left[\frac{\lambda_i}{\lambda_i + \mu_i} \prod_{j \neq i} \frac{\lambda_j}{\lambda_j + \mu_j} \right] \mu_i \, \Delta t + o(\Delta t) \\ &= \prod_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \mu_j} \sum_{i=1}^{n} \mu_i \Delta t + o(\Delta t)\end{aligned}$$

Because Δt is assumed to be very small, no more than one structure failure will occur in the interval. We can therefore use Blackwell's theorem (10.49) to conclude that the above expression is just Δt times the reciprocal of the mean time between structure failures, MTBF_S, that is

$$\text{MTBF}_S = \left[\prod_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \mu_j} \sum_{i=1}^{n} \mu_i \right]^{-1}. \tag{10.99}$$

When the parallel structure is in a failed state, all the n items are in a failed state. Because the downtimes are assumed to be independent with rates μ_i for $i = 1, 2, \dots, n$, the downtime of the parallel structure will be exponentially distributed with rate $\sum_{i=1}^{n} \mu_i$, and the mean downtime of the parallel structure is

$$\mathrm{MDT}_S = \frac{1}{\sum_{i=1}^{n} \mu_i}.$$

The mean up-time, or the mean time-to-failure, MTTF_S of the parallel structure is equal to $\mathrm{MTBF}_S - \mathrm{MDT}_S$

$$\begin{aligned}\mathrm{MTTF}_S &= \left[\prod_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \mu_j} \sum_{i=1}^{n} \mu_i\right]^{-1} - \frac{1}{\sum_{i=1}^{n} \mu_i} \\ &= \frac{1 - \prod_{j=1}^{n} \lambda_j/(\lambda_j + \mu_j)}{\prod_{j=1}^{n} \lambda_j/(\lambda_j + \mu_j)\ \sum_{i=1}^{n} \mu_i}.\end{aligned} \tag{10.100}$$

To check that the above calculations are correct, we may calculate the average unavailability

$$\overline{A}_S = \frac{\mathrm{MDT}_S}{\mathrm{MTTF}_S + \mathrm{MDT}_S} = \prod_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \mu_j}.$$

(Example 10.15 is adapted from example 3.5(B) in Ross (1996)). □

Mean Number of Failures/Repairs

First, let the renewals be the events where a repair is completed. Then we have an ordinary renewal process with renewal periods $T_1, T_2, \dots$ which are independent and identically distributed with distribution function (10.93).

Assume that U_i and D_i both are continuously distributed with densities $f_U(t)$ and $f_D(t)$, respectively. The probability density function of the T_i's is then

$$f_T(t) = \int_0^t f_U(t - x) f_D(x)\, dx. \tag{10.101}$$

According to Appendix B, the Laplace transform of (10.101) is

$$f_T^*(s) = f_U^*(s) f_D^*(s).$$

Let $W_1(t)$ be the renewal function, that is, the mean number of completed repairs in the time interval $(0, t]$. According to (10.86)

$$W_1^*(s) = \frac{f_U^*(s) f_D^*(s)}{s[1 - f_U^*(s) f_D^*(s)]}. \tag{10.102}$$

In this case, both the U_i's and the D_i's are assumed to be continuously distributed, but this turns out *not* to be essential. Equation (10.102) is also valid for discrete distributions, or for a mixture of discrete and continuous distributions. In this case, we may use that

$$f_U^*(s) = E(e^{-sU_i})$$
$$f_D^*(s) = E(e^{-sD_i})$$

The mean number of completed repairs in $(0, t]$ can now, at least in principle, be determined for any choice of life- and repair time distributions.

Next, let the renewals be the events where a failure occurs. In this case, we get a delayed renewal process. The renewal periods $T_1, T_2, \dots$ are independent and $F_{T_1}(t)$ is given by (10.94), whereas the distribution of $T_2, T_3, \dots$ is given by (10.93).

Let $W_2(t)$ be the renewal function, that is, the mean number of failures in $(0, t]$ under these conditions. According to (10.86), the Laplace transform is

$$W_2^*(s) = \frac{f_U^*(s)}{s(1 - f_U^*(s)f_D^*(s))}, \tag{10.103}$$

which, at least in principle, can be inverted to obtain $W_2(t)$.

Availability at a Given Point of Time

By taking Laplace transforms of (10.97), we get

$$A^*(s) = \frac{1}{s} - F_U^*(s) + \left(\frac{1}{s} - F_U^*(s)\right) w_{F_T}^*(s).$$

Because

$$F^*(s) = \frac{1}{s} f^*(s),$$

then

$$A^*(s) = \frac{1}{s}[1 - f_U^*(s)][1 + w_{F_T}^*(s)].$$

For an ordinary renewal process (i.e. the renewals are the events where a repair is completed), then

$$w_{F_T}^*(s) = sW_1^*(s).$$

Hence,

$$A^*(s) = \frac{1}{s}[1 - f_U^*(s)]\left(1 + \frac{f_U^*(s)f_D^*(s)}{1 - f_U^*(s)f_D^*(s)}\right),$$

that is,

$$A^*(s) = \frac{1 - f_U^*(s)}{s(1 - f_U^*(s)f_D^*(s))}. \tag{10.104}$$

The availability $A(t)$ can in principle be determined from (10.104) for any choice of life and downtime distributions.

Example 10.16 (Exponential time-to-failure and exponential downtime) Consider an alternating renewal process where the item up-times $U_1, U_2, \dots$ are independent and exponentially distributed with failure rate λ. The corresponding downtimes are also assumed to be independent and exponentially distributed with rate $\mu = 1/\text{MDT}$.

Then

$$f_U(t) = \lambda e^{-\lambda t} \qquad \text{for} \qquad t > 0$$
$$f_U^*(s) = \frac{\lambda}{\lambda + s}$$

and

$$f_D(t) = \mu e^{-\mu t} \qquad \text{for} \qquad t > 0$$
$$f_D^*(s) = \frac{\mu}{\mu + s}$$

The availability $A(t)$ is then obtained from (10.104)

$$\begin{aligned} A^*(s) &= \frac{1 - \lambda/(\lambda + s)}{s[1 - (\lambda/(\lambda + s))(\mu/(\mu + s)]} \\ &= \frac{\mu}{\lambda + \mu}\frac{1}{s} + \frac{\lambda}{\lambda + \mu}\frac{1}{s + (\lambda + \mu)}. \end{aligned} \tag{10.105}$$

Equation (10.105) can be inverted (see Appendix B) and we get

$$A(t) = \frac{\mu}{\lambda + \mu} + \frac{\lambda}{\lambda + \mu}\, e^{-(\lambda + \mu)t}. \tag{10.106}$$

The availability $A(t)$ is shown in Figure 10.16.

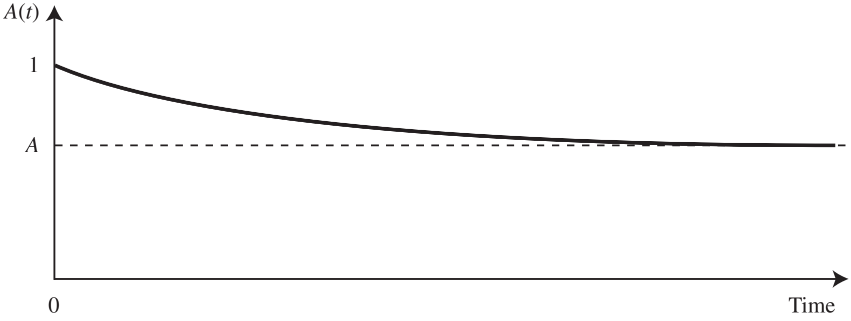


Figure 10.16 Availability of an item with exponential up- and downtimes.

The limiting availability is

$$A = \lim_{t\to\infty} A(t) = \frac{\mu}{\lambda+\mu} = \frac{1/\lambda}{1/\lambda + 1/\mu} = \frac{\text{MTTF}}{\text{MTTF} + \text{MDT}}.$$

By inserting $f_U^*(s)$ and $f_D^*(s)$ into (10.103), we get the Laplace transform of the mean number of renewals $W(t)$,

$$\begin{aligned} W^*(s) &= \frac{[\lambda/(\lambda+s)]\,[\mu/(\mu+s)]}{s(1-[\lambda/(\lambda+s)]\,[\mu/(\mu+s)])} \\ &= \frac{\lambda\mu}{\lambda+\mu}\,\frac{1}{s^2} - \frac{\lambda\mu}{(\lambda+\mu)^2}\,\frac{1}{s} + \frac{\lambda\mu}{(\lambda+\mu)^2}\,\frac{1}{s+(\lambda+\mu)} \end{aligned}$$

By inverting this expression, we get the mean number of completed repairs in the time interval $(0, t]$

$$W(t) = \frac{\lambda\mu}{\lambda+\mu} t - \frac{\lambda\mu}{(\lambda+\mu)^2} + \frac{\lambda\mu}{(\lambda+\mu)^2}\, e^{-(\lambda+\mu)t}. \tag{10.107}$$

□

Example 10.17 (Exponential time-to-failure and constant downtime)
Consider an alternating renewal process where the item up-times $U_1, U_2, \ldots$ are independent and exponentially distributed with failure rate λ. The downtimes are assumed to be constant and equal to τ with probability 1: $\Pr(D_i = \tau) = 1$ for $i = 1, 2, \ldots$.

The corresponding Laplace transforms are

$$f_U^*(s) = \frac{\lambda}{\lambda+s}$$

$$f_D^*(s) = E(e^{-sD}) = e^{-s\tau}\Pr(D=\tau) = e^{-s\tau}$$

Hence, the Laplace transform of the availability (10.104) becomes

$$\begin{aligned} A^*(s) &= \frac{1-\lambda/(\lambda+s)}{s[1-(\lambda/(\lambda+s))\,e^{-s\tau}]} = \frac{1}{s+\lambda-\lambda e^{-s\tau}} \\ &= \frac{1}{\lambda+s}\left[\frac{1}{1-(\lambda/(\lambda+s))\,e^{-s\tau}}\right] = \frac{1}{\lambda+s}\sum_{\nu=0}^{\infty}\left(\frac{\lambda}{\lambda+s}\right)^{\nu} e^{-s\nu\tau} \\ &= \frac{1}{\lambda}\sum_{\nu=0}^{\infty}\left(\frac{\lambda}{\lambda+s}\right)^{\nu+1} e^{-s\nu\tau}. \end{aligned} \tag{10.108}$$

The availability then becomes

$$A(t) = \mathcal{L}^{-1}(A^*(s)) = \sum_{\nu=0}^{\infty}\frac{1}{\lambda}\mathcal{L}^{-1}\left[\left(\frac{\lambda}{\lambda+s}\right)^{\nu+1} e^{-s\nu\tau}\right].$$

According to Appendix B

$$\mathcal{L}^{-1}\left[\left(\frac{\lambda}{\lambda+s}\right)^{\nu+1}\right] = \frac{\lambda^{\nu+1}}{\nu!} t^{\nu} e^{-\lambda t} = f(t),$$

$$\mathcal{L}^{-1}(e^{-s\nu\tau}) = \delta(t-\nu\tau),$$

where $\delta(t)$ is the Dirac delta-function. Thus

$$\begin{aligned}\mathcal{L}^{-1}\left[\left(\frac{\lambda}{\lambda+s}\right)^{\nu+1}e^{-s\nu\tau}\right] &= \mathcal{L}^{-1}\left[\left(\frac{\lambda}{\lambda+s}\right)^{\nu+1}\right] * \mathcal{L}^{-1}(e^{-s\nu\tau})\\ &= \int_0^\infty \delta(t-\nu\tau-x)f(x)\,dx = f(t-\nu\tau)u(t-\nu\tau)\end{aligned}$$

where

$$u(t-\nu\tau)=\begin{cases}1 & \text{if } t\ge\nu\tau\\ 0 & \text{if } t<\nu\tau\end{cases}.$$

Hence the availability is

$$A(t)=\sum_{\nu=0}^{\infty}\frac{\lambda^\nu}{\nu!}(t-\nu\tau)^\nu e^{-\lambda(t-\nu\tau)}u(t-\nu\tau). \tag{10.109}$$

The availability $A(t)$ is illustrated in Figure 10.17.

The limiting availability is then according to (10.98).

$$A=\lim_{t\to\infty}A(t)=\frac{\text{MTTF}}{\text{MTTF}+\text{MDT}}=\frac{1/\lambda}{(1/\lambda)+\tau}=\frac{1}{1+\lambda\tau}. \tag{10.110}$$

The Laplace transform for the renewal density is

$$\begin{aligned}w^*(s) &= \frac{f_T^*(s)f_D^*(s)}{1-f_T^*(s)f_D^*(s)}=\frac{\lambda e^{-s\tau}/(\lambda+s)}{1-\lambda e^{-s\tau}/(\lambda+s)}\\ &= \frac{1}{\lambda+s-\lambda e^{-s\tau}}\ \lambda e^{-s\tau}=\lambda A^*(s)\ e^{-s\tau}\end{aligned}$$

where $A^*(s)$ is given by (10.108).

Then the renewal density becomes

$$w(t)=\lambda\mathcal{L}^{-1}(A^*(s)e^{-s\tau})=\lambda\int_0^\infty \delta(t-\tau-x)A(x)\,dx,$$

that is,

$$w(t)=\begin{cases}\lambda A(t-\tau) & \text{if } t\ge\tau\\ 0 & \text{if } t<\tau\end{cases}. \tag{10.111}$$

Hence, the mean number of completed repairs in the time interval $(0,t]$ for $t>\tau$ is

$$W(t)=\int_0^t w(u)\,du=\lambda\int_\tau^t A(u-\tau)\,du=\lambda\int_0^{t-\tau}A(u)\,du. \tag{10.112}$$

□

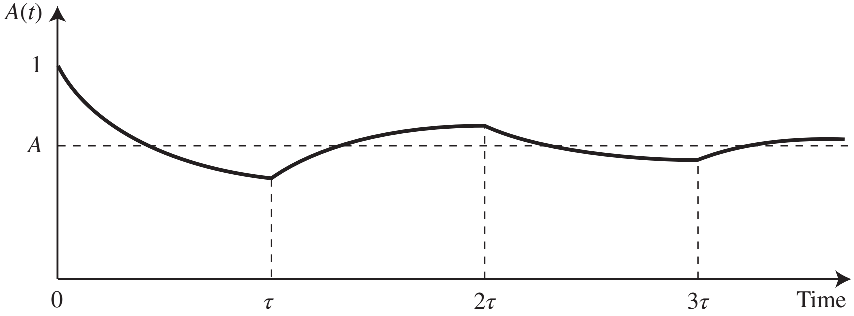


Figure 10.17 The availability of an item with exponential uptimes and constant downtime (τ).

10.4 Nonhomogeneous Poisson Processes

In this section, the HPP is generalized by allowing the rate of the process to be a function of time, in which case the counting process is called a nonhomogeneous Poisson process (NHPP).

10.4.1 Introduction and Definitions

An NHPP is defined as:

Definition 10.7 (Nonhomogeneous Poisson process)
A counting process $\{N(t), t \geq 0\}$ is a nonhomogeneous (or nonstationary) Poisson process (NHPP) with rate function $w(t)$ for $t \geq 0$, if

(1) $N(0) = 0$.
(2) $\{N(t), t \geq 0\}$ has independent increments.
(3) $\Pr(N(t + \Delta t) - N(t) \geq 2) = o(\Delta t)$, which means that the item will not experience more than one failure at the same time.
(4) $\Pr(N(t + \Delta t) - N(t) = 1) = w(t)\Delta t + o(\Delta t)$.

□

The basic "parameter" of the NHPP is the ROCOF function $w(t)$. This function is also called the *peril rate* of the NHPP. The *cumulative rate* of the process is

$$W(t) = \int_0^t w(u)\, du. \tag{10.113}$$

This definition also covers the situation in which the rate is a function of some observed explanatory variable that is a function of time t. Observe that the NHPP model does not require stationary increments. This means that failures may be more likely to occur at certain times than others, and hence the interoccurrence times are generally neither independent nor identically distributed. Consequently, statistical techniques based on the assumption of independent and identically distributed variables cannot be applied to an NHPP.

The NHPP is often used to model trends in the interoccurrence times, such as, improving (*happy*) or deteriorating (*sad*) items. It seems intuitive that a happy item has a decreasing ROCOF function, whereas a sad item has an increasing ROCOF function. Several studies of failure data from practical items have concluded that the NHPP was an adequate model, and that the items that were studied approximately satisfied the properties of the NHPP listed in Definition 10.7.

Due to the assumption of independent increments, the number of failures in a specified interval $(t_1, t_2]$ is independent of the failures and interoccurrence times prior to t_1. When a failure has occurred at time t_1, the conditional ROCOF $w_C(t \mid \mathcal{H}_t)$ in the next interval will be $w(t)$ and independent of the history $\mathcal{H}_{t_1}$ up to time t_1. In the case when no failure has occurred before t_1, $w(t) = z(t)$ (i.e. the failure rate function (FOM) for $t < t_1$). A practical implication of this assumption is that the conditional (ROCOF), $w_C(t)$, is the same just before a failure and immediately after the corresponding repair. This assumption is called *minimal repair*. When replacing failed parts that may have been in operation for a long time, with new ones, an NHPP clearly is not a realistic model. For the NHPP to be realistic, the parts put into service should be identical to the old ones, and hence should be aged outside the item under identical conditions for the same period of time.

Consider an item consisting of a large number of components. Suppose that a critical component fails and causes an item failure and that this component is immediately replaced by a component of the same type, thus causing a negligible item downtime. Because only a small fraction of the item is replaced, it seems natural to assume that the items's reliability after the repair essentially is the same as immediately before the failure. In other words, the assumption of *minimal repair* is a realistic approximation. When an NHPP is used to model a repairable item, the item is treated as a *black box* in that no concern is made about how the item "looks inside."

A car is a typical example of a repairable item. Usually the operating time of a car is expressed in terms of the mileage indicated on the speedometer. Repair actions will usually not imply any extra mileage. The repair "time" is thus negligible. Many repairs are accomplished by adjustments, or replacement of single components. The minimal repair assumption is therefore often applicable and the NHPP may be accepted as a realistic model, at least as a first order approximation.

Consider an NHPP with ROCOF $w(t)$, and suppose that failures occur at times $S_1, S_2, \ldots$. An illustration of $w(t)$ is shown in Figure 10.18.

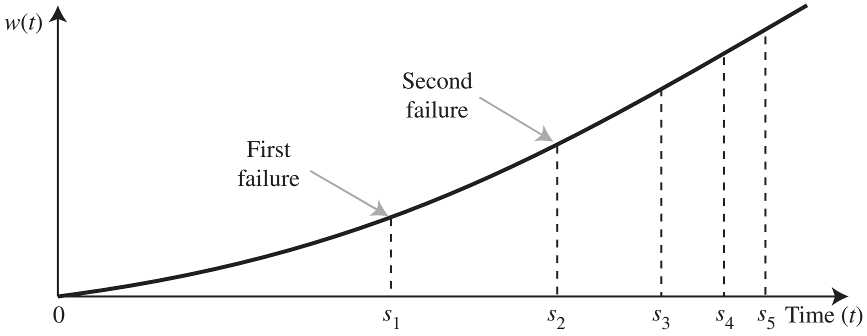


Figure 10.18 The ROCOF $w(t)$ of an NHPP and random failure times.

10.4.2 Some Results

From the definition of the NHPP it follows (e.g. see Ross 1996) that the number of failures in the interval $(0, t]$ is Poisson distributed

$$\Pr(N(t) = n) = \frac{[W(t)]^n}{n!}\, e^{-W(t)} \qquad \text{for} \qquad n = 0, 1, 2, \ldots . \tag{10.114}$$

The mean number of failures in $(0, t]$ is therefore

$$E[N(t)] = W(t),$$

and the variance is $\text{var}[N(t)] = W(t)$. The cumulative rate $W(t)$ of the process (10.106) is therefore the mean number of failures in the interval $(0, t]$, and is sometimes called the *mean value function* of the process. When n is large, $\Pr(N(t) \le n)$ may be determined by normal approximation

$$\begin{aligned}\Pr(N(t) \le n) &= \Pr\left(\frac{N(t) - W(t)}{\sqrt{W(t)}} \le \frac{n - W(t)}{\sqrt{W(t)}}\right)\\ &= \Phi\left(\frac{n - W(t)}{\sqrt{W(t)}}\right).\end{aligned} \tag{10.115}$$

From (10.114) it follows that the number of failures in the interval $(v, t + v]$ is Poisson distributed

$$\Pr(N(t + v) - N(v) = n) = \frac{[W(t + v) - W(v)]^n}{n!}\, e^{-[W(t+v)-W(v)]}$$
$$\text{for} \qquad n = 0, 1, 2, \ldots$$

and that the mean number of failures in the interval $(v, t + v]$ is

$$E[N(t + v) - N(v)] = W(t + v) - W(v) = \int_v^{t+v} w(u)\, du. \tag{10.116}$$

The probability of no failure in the interval (t_1, t_2) is

$$\Pr[N(t_2) - N(t_1) = 0] = e^{-\int_{t_1}^{t_2} w(t)\, dt}.$$

Let S_n be the time until failure n for $n = 0, 1, 2, \ldots$, where $S_0 = 0$. The distribution of S_n is given by:

$$\Pr(S_n > t) = \Pr(N(t) \le n - 1) = \sum_{k=0}^{n-1} \frac{W(t)^k}{k!}\, e^{-W(t)}. \tag{10.117}$$

When $W(t)$ is small, this probability may be determined from standard tables of the Poisson distribution. When $W(t)$ is large, the probability may be determined by normal approximation, see Eq. (10.115)

$$\begin{aligned}\Pr(S_n > t) &= \Pr(N(t) \le n - 1)\\ &\approx \Phi\left(\frac{n - 1 - W(t)}{\sqrt{W(t)}}\right).\end{aligned} \tag{10.118}$$

Time to First Failure

Let T_1 be the time from $t = 0$ until the first failure. The survivor function of T_1 is

$$R_1(t) = \Pr(T_1 > t) = \Pr(N(t) = 0) = e^{-W(t)} = e^{-\int_0^t w(t)\, dt}. \tag{10.119}$$

Hence, the failure rate (FOM) function $z_{T_1}(t)$ of the first interoccurrence time T_1 is equal to the ROCOF $w(t)$ of the process. Observe the different meanings of the two expressions. $z_{T_1}(t)\Delta t$ approximates the (conditional) probability that the *first* failure occurs in $(t, t + \Delta t]$, whereas $w(t)\Delta t$ approximates the (unconditional) probability that a failure, not necessarily the first, occurs in $(t, t + \Delta t]$.

A consequence of (10.119) is that the distribution of the first interoccurrence time, that is, the time from $t = 0$ until the item's first failure, determines the ROCOF of the entire process. Thompson (1981) claims that this is a nonintuitive fact that casts doubt on the NHPP as a realistic model for repairable items. Use of an NHPP model implies that if we are able to estimate the failure rate (FOM) function of the time to the *first* failure, such as for a specific type of cars, we at the same time have an estimate of the ROCOF of the entire life of the car.

Time Between Failures

Assume that the process is observed at time t_0, and let $Y(t_0)$ be the time until the next failure. In Section 10.3.5, $Y(t_0)$ is called the remaining lifetime, or the forward recurrence time. By using (10.114), the distribution of $Y(t_0)$ can be expressed as

$$\begin{aligned}\Pr(Y(t_0) > t) &= \Pr(N(t + t_0) - N(t_0) = 0) = e^{-[W(t+t_0)-W(t_0)]}\\ &= e^{-\int_{t_0}^{t+t_0} w(u)\, du} = e^{-\int_0^t w(u+t_0)\, du}.\end{aligned} \tag{10.120}$$

Observe that this result is independent of whether t_0 is a failure time or an arbitrary point in time. Assume that t_0 is the time, S_{n-1}, of failure $n-1$. In this case, $Y(t_0)$ is the time between failure $n-1$ and failure n (i.e. the nth interoccurrence time $T_n = S_n - S_{n-1}$). The failure rate (FOM) function of the nth interoccurrence time T_n is from (10.120)

$$z_{t_0}(t) = w(t + t_0) \qquad \text{for} \qquad t \geq 0. \tag{10.121}$$

Observe that this is a conditional failure rate, given that $S_{n-1} = t_0$. The mean time between failure $n-1$ (at time t_0) and failure n, MTBF$_n$ is

$$\text{MTBF}_n = E(T_n) = \int_0^\infty \Pr(Y_{t_0} > t)\, dt = \int_0^\infty e^{-\int_0^t w(u+t_0)\, du}\, dt. \tag{10.122}$$

Example 10.18 Consider an NHPP with ROCOF $w(t) = 2\lambda^2 t$, for $\lambda > 0$ and $t \geq 0$. The mean number of failures in the interval $(0, t)$ is $W(t) = E[N(t)] = \int_0^t w(u)\, du = (\lambda t)^2$. The distribution of the time to the first failure, T_1, is given by the survivor function

$$R_1(t) = e^{-W(t)} = e^{-(\lambda t)^2} \qquad \text{for} \qquad t \geq 0,$$

that is, a Weibull distribution with scale parameter λ and shape parameter $\alpha = 2$. If we observe the process at time t_0, the distribution of the time $Y(t_0)$ till the next failure is from (10.120)

$$\Pr(Y(t_0) > t) = e^{-\int_0^t w(u+t_0)\, du} = e^{-\lambda^2(t^2 + 2t_0 t)}.$$

If t_0 is the time of failure $n-1$, the time to the next failure, $Y(t_0)$ is the nth interoccurrence time T_n and the failure rate (FOM) function of T_n is

$$z_{t_0}(t) = 2\lambda^2(t + t_0),$$

which is linearly increasing with the time t_0 of failure $n-1$. Observe again that this is a conditional rate, given that failure $n-1$ occurred at time $S_{n-1} = t_0$. The mean time between failure $n-1$ and failure n is

$$\text{MTBF}_n = \int_0^\infty e^{-\lambda^2(t^2 + 2t_0 t)}\, dt.$$

□

Relation to the Homogeneous Poisson Process

Let $\{N(t), t \geq 0\}$ be an NHPP with ROCOF $w(t) > 0$ such that the inverse $W^{-1}(t)$ of the cumulative rate $W(t)$ exists, and let $S_1, S_2, \ldots$ be the times when the failures occur.

Consider the time-transformed occurrence times $W(S_1), W(S_2), \ldots,$ and let $\{N^*(t), t \geq 0\}$ be the associated counting process. The distribution of the (transformed) time $W(S_1)$ till the first failure is from (10.121)

$$\Pr[W(S_1) > t] = \Pr[S_1 > W^{-1}(t)] = e^{-W[W^{-1}(t)]} = e^{-t},$$

that is, an exponential distribution with parameter 1.

The new counting process is defined by

$$N(t) = N^*[W(t)] \qquad \text{for} \qquad t \geq t,$$

hence,

$$N^*(t) = N[W^{-1}(t)] \qquad \text{for } t \geq 0,$$

and we get from (10.116)

$$\begin{aligned}\Pr[N^*(t) = n] &= \Pr(N[W^{-1}(t)] = n) \\ &= \frac{(W[W^{-1}(t)])^n}{n!} e^{-W[W^{-1}(t)]} = \frac{1^n}{n!}\ e^{-t}\end{aligned}$$

that is, the Poisson distribution with rate 1. We have thereby shown that an NHPP with cumulative – and invertible – rate $W(t)$ can be transformed into an HPP with rate 1, by time-transforming the failure occurrence times $S_1, S_2, \dots$ to $W(S_1), W(S_2), \dots$.

10.4.3 Parametric NHPP Models

Several parametric models have been established to describe the ROCOF of an NHPP. Among these are:

(1) The power law model
(2) The linear model
(3) The log-linear model

All the three models may be written in the common form (see Atwood 1992)

$$w(t) = \lambda_0\ g(t; \vartheta), \tag{10.123}$$

where λ_0 is a common multiplier, and $g(t; \vartheta)$ determines the shape of the ROCOF $w(t)$. The three models may be parameterized in various ways. This section presents the parameterization of Crowder et al. (1991), although the parametrization of Atwood (1992) may be more logical.

The Power Law Model

For the power law model, the ROCOF is

$$w(t) = \lambda\beta t^{\beta-1} \qquad \text{for} \qquad \lambda > 0, \qquad \beta > 0, \qquad \text{and} \qquad t \geq 0. \tag{10.124}$$

This NHPP is sometimes referred to as a *Weibull process* because the ROCOF has the same functional form as the failure rate (FOM) function of the Weibull distribution. Also observe that the first arrival time T_1 of this process is Weibull distributed with shape parameter β and scale parameter λ. According to Ascher and Feingold (1984), one should avoid the name Weibull process in this situation because it gives the wrong impression that the Weibull distribution can be used to model trend in interoccurrence times of a repairable item. Hence, such a notation may lead to confusion.

A repairable item modeled by the Power law model is seen to be improving (happy) if $0 < \beta < 1$, and deteriorating (sad) if $\beta > 1$. If $\beta = 1$, the model reduces to an HPP. The case $\beta = 2$ is seen to give a linearly increasing ROCOF. This model is studied in Example 10.18.

Assume that we have observed an NHPP in a time interval $(0, t_0]$ and that failures have occurred at times $s_1, s_2, \dots, s_n$. Maximum likelihood estimates $\hat{\beta}$ and $\hat{\lambda}$ of β and λ, respectively, are given by

$$\hat{\beta} = \frac{n}{n \ln t_0 - \sum_{i=1}^{n} \ln s_i}, \tag{10.125}$$

and

$$\hat{\lambda} = \frac{n}{t_0^{\hat{\beta}}}. \tag{10.126}$$

The estimates are further discussed by Crowder et al. (1991) and Cocozza-Thivent (1997). A $(1 - \varepsilon)$ confidence interval for β is given by Cocozza-Thivent (1997)

$$\left(\frac{\hat{\beta}}{2n}\, z_{(1-\varepsilon/2),2n}, \frac{\hat{\beta}}{2n}\, z_{(1+\varepsilon/2),2n} \right), \tag{10.127}$$

where $z_{\varepsilon,\nu}$ is the upper $100\varepsilon\%$ percentile of the χ^2 distribution with ν degrees of freedom.

The Linear Model

For the linear model, the ROCOF is

$$w(t) = \lambda(1 + \alpha t) \qquad \text{for} \qquad \lambda > 0 \qquad \text{and} \qquad t \geq 0. \tag{10.128}$$

The linear model is discussed by Vesely (1991) and Atwood (1992). A repairable item modeled by the linear model is deteriorating if $\alpha > 0$, and improving when $\alpha < 0$. When $\alpha < 0$, then $w(t)$ will sooner or later become less than zero. The model should only be used in time intervals where $w(t) > 0$.

The Log-Linear Model

For the log-linear model, which is also called the *Cox–Lewis* model, the ROCOF is

$$w(t) = e^{\alpha + \beta t} \qquad \text{for} - \infty < \alpha, \beta < \infty \qquad \text{and} \qquad t \geq 0. \tag{10.129}$$

A repairable item modeled by the log-linear model is improving (happy) if $\beta < 0$, and deteriorating (sad) if $\beta > 0$. When $\beta = 0$, the log-linear model reduces to an HPP.

The log-linear model was proposed by Cox and Lewis (1966) who used the model to investigate trends in the interoccurrence times between failures in air conditioning equipment in aircrafts. The first arrival time T_1 has failure rate (FOM) function $z(t) = e^{\alpha+\beta t}$ and hence has a truncated Gumbel distribution of the smallest extreme.

Assume that we have observed an NHPP in a time interval $(0, t_0]$ and that failures occurred at times $s_1, s_2, \dots, s_n$. Maximum likelihood estimates $\hat{\alpha}$ and $\hat{\beta}$ of α and β, respectively, are found by solving

$$\sum_{i=1}^{n} s_i + \frac{n}{\beta} - \frac{nt_0}{(1 - e^{-\beta t_0})} = 0, \tag{10.130}$$

to give $\hat{\beta}$, and then taking

$$\hat{\alpha} = \ln\left(\frac{n\hat{\beta}}{e^{\hat{\beta} t_0} - 1}\right), \tag{10.131}$$

The estimates are further discussed by Crowder et al. (1991).

10.4.4 Statistical Tests of Trend

The simple graph in Figure 10.3 clearly indicates an increasing rate of failures, that is, a deteriorating or *sad* item. The next step in an analysis of the data may be to perform a *statistical test* to find out whether the observed trend is *statistically significant* or just accidental. A number of tests have been developed for this purpose, that is for testing the null hypothesis

H_0: "No trend" (or more precisely that the interoccurrence times are independent and identically exponentially distributed, that is, an HPP)
against the alternative hypothesis
H_1: "Monotonic trend" (i.e. The process is an NHPP that is either *sad* or *happy*)

Among these are two nonparametric tests that we will discuss:

(1) The Laplace test
(2) The Military Handbook (MIL HDBK) test

These two tests are discussed in detail by Ascher and Feingold (1984) and Crowder et al. (1991). It can be shown that the Laplace test is optimal when the true failure mechanism is that of a log-linear NHPP model (Cox and Lewis 1966), whereas the Military Handbook test is optimal when the true failure mechanism is that of a power law NHPP model (Bain et al. 1985).

The Laplace Test

The test statistic for the case where the item is observed until n failures have occurred is

$$U = \frac{\frac{1}{n-1}\sum_{j=1}^{n-1} S_j - (S_n/2)}{S_n/\sqrt{12(n-1)}}, \tag{10.132}$$

where $S_1, S_2, \ldots$ are the failure times. For the case where the item is observed until time t_0, the test statistic is

$$U = \frac{\frac{1}{n}\sum_{j=1}^{n} S_j - (t_0/2)}{t_0/\sqrt{12n}}. \tag{10.133}$$

In both cases, the test statistic U is approximately standard normally $\mathcal{N}(0,1)$ distributed when the null hypothesis H_0 is true. The value of U is seen to indicate the direction of the trend, with $U < 0$ for a *happy* item and $U > 0$ for a *sad* item. Optimal properties of the Laplace test have, for example, been investigated by Gaudoin (1992).

Military Handbook Test

The test statistic of the so-called Military Handbook test (MIL-HDBK-189C 2011) for the case where the item is observed until n failures have occurred is

$$Z = 2\sum_{i=1}^{n-1} \ln \frac{S_n}{S_i}. \tag{10.134}$$

For the case where the item is observed until time t_0, the test statistic is

$$Z = 2\sum_{i=1}^{n} \ln \frac{t_0}{S_i}. \tag{10.135}$$

The asymptotic distribution of Z is in the two cases a χ^2 distribution with $2(n-1)$ and $2n$ degrees of freedom, respectively.

The hypothesis of no trend (H_0) is rejected for *small* or *large* values of Z. Low values of Z correspond to deteriorating items, whereas large values of Z correspond to improving items.

10.5 Imperfect Repair Processes

The Sections 10.3 and 10.4 deal with two main categories of models that can be used to describe the occurrence of failures of repairable items; renewal processes and NHPPs – where the HPP is a special case of both models. When using a renewal process, the repair action is considered to be *perfect*, meaning the item is as-good-as-new after the repair action is completed. When using an NHPP, the repair action is *minimal*, meaning that the reliability of the item is the same immediately after the repair action as it was immediately before the failure occurred. In this case, we say the item is as-bad-as-old after the repair action. The renewal process and the NHPP may thus be considered as two extreme cases. Items subject to normal repair will be somewhere between these two extremes. Several models have been suggested for the *normal*, or *imperfect* repair situation, a repair that is somewhere between a minimal repair and a renewal.

This section considers an item that is put into operation at time t. The initial failure rate (FOM) function of the item is $z(t)$, and the conditional ROCOF of the item is $w_C(t)$. The conditional ROCOF is defined by (10.7).

When the item fails, a repair action is initiated. The repair action brings the item back to a functioning state and may involve a repair, or a replacement of the component that produced the item failure. The repair action may also involve maintenance and upgrading of the rest of the item, and even replacement of the whole item. The time required to perform the repair action is considered to be negligible. Preventive maintenance, except for preventive maintenance carried out during a repair action, is disregarded.

A high number of models have been suggested for modeling imperfect repair processes. Most of the models may be classified in two main groups: (i) models where the repair actions reduce the rate of failures (ROCOF) and (ii) models where the repair actions reduce the (virtual) age of the item. A survey of available models are provided, e.g. by Pham and Wang (1996), Hokstad (1997), and Akersten (1998).

10.5.1 Brown and Proschan's model

One of the best known imperfect repair models is described by Brown and Proschan (1983). Brown and Proschan's model is based on the following repair policy: A item is put into operation at time $t = 0$. Each time the item fails, a repair action is initiated, that with probability p is a *perfect* repair that will bring the item back to an as-good-as-new condition. With probability $1 - p$, the repair action is a *minimal* repair, leaving the item in an as-bad-as-old condition. The renewal process and the NHPP are seen to be special cases of Brown and Proschan's model, when $p = 1$ and $p = 0$, respectively. Brown and Proschan's model may therefore be regarded as a mixture of the renewal process and the NHPP. Observe that the probability p of a perfect repair is independent of the time elapsed since the previous failure and also of the age of the item. Let us, as an example, assume that $p = 0.02$. This means that we for most failures will make do with a minimal repair, and on the average renew (or, replace) the item once for every 50 failures. This may be a realistic model, but the problem is that the renewals come at random, meaning that we have the same probability of renewing a rather new

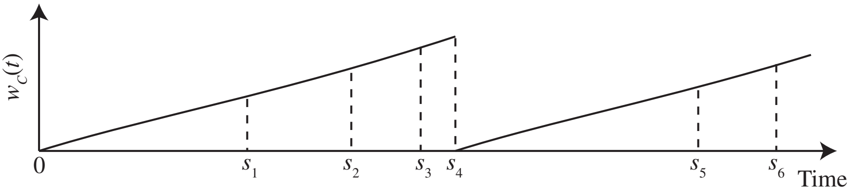


Figure 10.19 An illustration of a possible shape of the conditional ROCOF of Brown and Proschan's imperfect repair model.

item as an old item. Figure 10.19 illustrates a possible shape of the conditional ROCOF.

Datasets available for repairable items are usually limited to the times between failures, $T_1, T_2, \ldots$. Detailed repair modes associated to each failure are in general not recorded. Based on this "masked" data set, Lim (1998) provides a procedure for estimating p and the other parameters of Brown and Proschan's model.

Brown and Proschan's model is extended by Block et al. (1985) to age-dependent repair, that is, when the item fails at time t, a perfect repair is performed with probability $p(t)$ and a minimal repair is performed with probability $1 - p(t)$. Let Y_1 be the time from $t = 0$ until the first perfect repair. When a perfect repair is carried out, the process starts over again, and we get a sequence of times between perfect repairs $Y_1, Y_2, \ldots$ that will form a renewal process. Let $F(t)$ be the distribution of the time to the first failure T_1, and let $f(t)$ and $R(t) = 1 - F(t)$ be the corresponding probability density function and the survivor function, respectively. The failure rate (FOM) function of T_1 is then $z(t) = f(t)/R(t)$, and we know from Chapter 5 that the distribution function may be written as

$$F(t) = 1 - e^{-\int_0^t z(x)\,dx} = 1 - e^{-\int_0^t [f(x)/R(x)]\,dx}.$$

The distribution of Y_i is given by Block et al. (1985)

$$F_p(t) = \Pr(Y_i \leq t) = 1 - e^{-\int_0^t [p(x)f(x)/R(x)]\,dx} = 1 - e^{-\int_0^t z_p(x)\,dx}. \tag{10.136}$$

Hence, the time between renewals, Y has failure rate (FOM) function

$$z_p(t) = \frac{p(t)f(t)}{R(t)} = p(t)z(t). \tag{10.137}$$

Block et al. (1985) supply an explicit formula for the renewal function and discuss the properties of $F_p(t)$.

10.5.2 Failure Rate Reduction Models

Several models have been suggested where each repair action results in a reduction of the conditional ROCOF. The reduction may be a fixed reduction, a certain percentage of the actual value of the rate of failures, or a function of the history of the process. Models representing the first two types were proposed by Chan and Shaw (1993). Let $z(t)$ be the failure rate (FOM) function of the time to the first failure. If all repairs were minimal repairs, the ROCOF of the process would be $w_1(t) = z(t)$. Consider the failure at time S_i, and let S_{i-} be the time immediately before time S_i. In the same way, let S_{i+} be the time immediately after time S_i. The models suggested by Chan and Shaw (1993) may then be expressed by the conditional ROCOF as

$$w_C(S_{i+}) = w_C(S_{i-}) - \Delta \qquad \text{for a fixed reduction } \Delta \tag{10.138}$$
$$w_C(S_{i+}) = w_C(S_{i-})(1-\rho) \qquad \text{for a proportional reduction} \qquad 0 \le \rho \le 1.$$

Between two failures, the conditional ROCOF is assumed to be vertically parallel to the initial ROCOF, $w_1(t)$. The parameter ρ in 10.138 is an index representing the efficiency of the repair action. When $\rho = 0$, we have minimal repair, and the NHPP is therefore a special case of Chan and Shaw's proportional reduction model. When $\rho = 1$, the repair action brings the conditional ROCOF down to zero, but does not represent a renewal process because the interoccurrence times are not identically distributed, except for the special case when $w_1(t)$ is a linear function. The conditional ROCOF of Chan and Shaw's proportional reduction model is illustrated in Figure 10.20 for some possible failure times and with $\rho = 0.30$.

Chan and Shaw's model 10.138 is generalized by Doyen and Gaudoin (2002) and Doyen and Gaudoin (2011). They propose a set of models where the proportionality factor ρ depends on the history of the process. In their models, the conditional ROCOF is expressed as

$$w_C(S_{i+}) = w_C(S_{i-}) - \varphi(i, S_1, S_2, \dots, S_i), \tag{10.139}$$

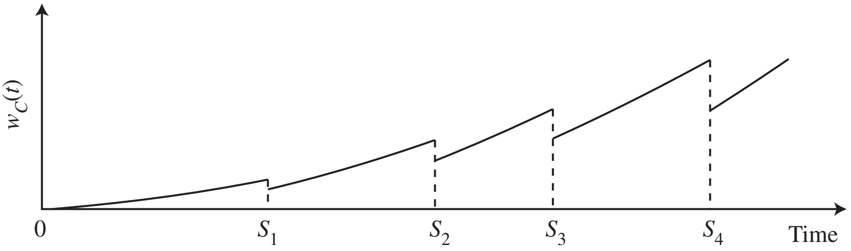


Figure 10.20 The conditional ROCOF of Chan and Shaw's proportional reduction model for some possible failure times ($\rho = 0.30$).

where $\varphi(i, S_1, S_2, \dots, S_i)$ is the reduction of the conditional ROCOF resulting from the repair action. Between two failures, they assume that the conditional ROCOF is vertically parallel to the initial ROCOF $w_1(t)$. These assumptions lead to the conditional ROCOF

$$w_C(t) = w_1(t) - \sum_{i=1}^{N(t)} \varphi(i, S_1, S_2, \dots, S_i). \tag{10.140}$$

When we, as in Chan and Shaw's model (10.138) assume a proportional reduction after each repair action, the conditional ROCOF in the interval $(0, S_1)$ becomes $w_C(t) = w_1(t)$. In the interval $[S_1, S_2)$, the conditional ROCOF is $w_C(t) = w_1(t) - \rho\, w_1(S_1)$. In the third interval $[S_2, S_3)$, the conditional ROCOF is

$$\begin{aligned} w_C(t) &= w_1(t) - \rho\, w_1(S_1) - \rho\, (w_1(S_2) - \rho\, w_1(S_1)) \\ &= w_1(t) - \rho\, [(1-\rho)^0 w_1(S_2) + (1-\rho)^1 w_1(S_1)] \end{aligned}$$

By continuing this derivation, we can show that the conditional ROCOF of Chan and Shaw's proportional reduction model (10.138) may be written as

$$w_C(t) = w_1(t) - \rho \sum_{i=0}^{N(t)} (1-\rho)^i\ w_1(S_{N(t)-i}). \tag{10.141}$$

This model is called *arithmetic reduction of intensity* with infinite memory (ARI$_\infty$) by Doyen and Gaudoin (2011).

In (10.138), the reduction is proportional to the conditional ROCOF just before time t. Another approach is to assume that a repair action can only reduce a proportion of the wear that has accumulated since the previous repair action. This can be formulated as:

$$w_C(S_{i+}) = w_C(S_{i-}) - \rho[w_C(S_{i-}) - w_C(S_{i-1+})]. \tag{10.142}$$

The conditional ROCOF of this model is

$$w_C(t) = w_1(t) - \rho\, w_1(S_{N(t)}). \tag{10.143}$$

This model is called arithmetic reduction of intensity with memory one (ARI$_1$) by Doyen and Gaudoin (2011). If $\rho = 0$, the item is as-bad-as-old after the repair action, and the NHPP is thus a special case of the ARI$_1$ model. If $\rho = 1$, the conditional ROCOF is brought down to zero by the repair action, but the process is not a renewal process, because the interoccurrence times are not identically distributed. For the ARI$_1$ model, there exists a deterministic function $w_{\min}(t)$ that is always smaller than the conditional ROCOF such that there is a nonzero probability that the ROCOF will be excessively close to $w_{\min}(t)$.

$$w_{\min}(t) = (1-\rho)\ w_1(t).$$

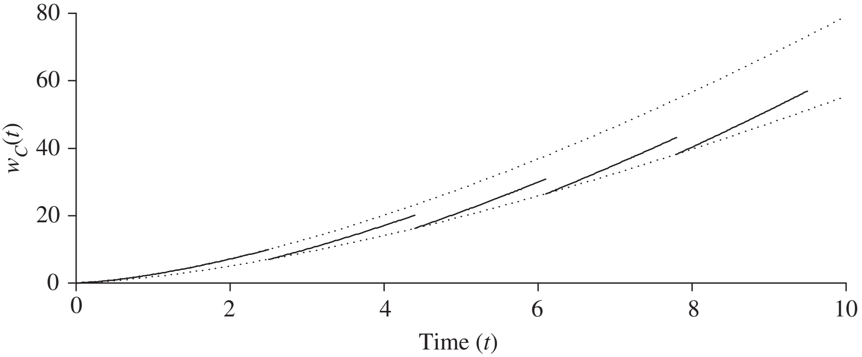


Figure 10.21 The ARI_1 model for some possible failure times. The "underlying" ROCOF $w_1(t)$ is a power law model with shape parameter $\beta = 2.5$, and the parameter $\rho = 0.30$. The upper dotted curve is $w_1(t)$, and the lower dotted curve is the minimal wear intensity $(1 - \rho)w_1(t)$.

This intensity is a minimal wear intensity, that is to say a maximal lower boundary for the conditional ROCOF. The ARI_1 model is illustrated in Figure 10.21 for some possible failure times.

The two models ARI_∞ and ARI_1 may be considered as two extreme cases. To illustrate the difference, we may consider the conditional ROCOF as an index representing the wear of the item. By the ARI_∞ model, every repair action will reduce, by a specified percentage ρ, the total accumulated wear of the item since the item was installed. By the ARI_1 model, the repair action will only reduce, by a percentage ρ, the wear that has been accumulated since the previous repair action. This is why Doyen and Gaudoin (2002) say that the ARI_∞ has infinite memory, whereas the ARI_1 has memory one (one period).

Doyen and Gaudoin (2011) also introduce a larger class of models in which only the first m terms of the sum in (10.143) are considered. They call this model the arithmetic reduction of intensity model of memory m (ARI_m), and the corresponding conditional ROCOF is

$$w_C(t) = w_1(t) - \rho \sum_{i=0}^{\min\{m-1,N(t)\}} (1-\rho)^i \, w_C(S_{N(t)-i}). \tag{10.144}$$

The ARI_m model has a minimal wear intensity:

$$w_{\min}(t) = (1-\beta)^m w_1(t).$$

In all these models, we observe that the parameter ρ may be regarded as an index of the efficiency of the repair action.

- $0 < \rho < 1$. The repair action is efficient.
- $\rho = 1$. Optimal repair. The conditional ROCOF is put back to zero (but the repair effect is different from the as-good-as-new situation.

- $\rho = 0$. The repair action has no effect on the wear of the item. The item state after the repair action is as-bad-as-old.
- $\rho < 0$. The repair action is harmful to the item, and will introduce extra problems.

10.5.3 Age Reduction Models

Malik (1979) proposes a model where each repair action reduces the *age* of the item. The age of the item is hence considered as a virtual concept.

To establish a model, assume that an item is put into operation at time $t = 0$. The initial ROCOF $w_1(t)$ is equal to the failure rate (FOM) function $z(t)$ of the interval until the first item failure. $w_1(t)$ is then the ROCOF of an item where all repairs are minimal repairs. The first failure occurs at time S_1, and the conditional ROCOF just after the repair action is completed is

$$w_C(S_{1+}) = w_1(S_1 - \vartheta),$$

where $S_1 - \vartheta$ is the new virtual age of the item. After the next failure, the conditional ROCOF is $w_C(S_{2+}) = w_1(S_2 - 2\vartheta)$, and so on. The conditional ROCOF at time t is

$$w_C(t) = w_1(t - N(t)\vartheta).$$

Next, let ϑ be a function of the history such that

$$w_C(t) = w_1\left(t - \sum_{i=1}^{N(t)} \vartheta(i, S_1, S_2, \dots S_i)\right). \tag{10.145}$$

Between two consecutive failures, assume that the conditional ROCOF is horizontally parallel with the initial ROCOF $w_1(t)$.

Doyen and Gaudoin (2002) propose an age reduction model where the repair action reduces the virtual age of the item with an amount proportional to its age just before the repair action. Let ρ be the percentage of reduction of the virtual age. In the interval $(0, S_1)$ the conditional ROCOF is $w_C(t) = w_1(t)$. Just after the first failure (when the repair is completed), the virtual age is $S_1 - \rho S_1$, and in the interval (S_1, S_2) the conditional ROCOF is $w_C(t) = w_1(t - \rho S_1)$. Just before the second failure at time S_2, the virtual age is $S_2 - \rho S_1$, and just after the second failure the virtual age is $S_2 - \rho S_1 - \rho(S_2 - \rho S_1)$. In the interval (S_2, S_3) the conditional ROCOF is $w_C(t) = w_1[t - \rho S_1 - \rho(S_2 - \rho S_1)]$ which may be written as $w_C(t) = w_1(t - \rho(1-\rho)^0 S_2 - \rho(1-\rho)^1 S_1)$. By continuing this argument, it is easy to realize that the conditional ROCOF of this age reduction model is

$$w_C(t) = w_1\left(t - \rho \sum_{i=0}^{N(t)} (1-\rho)^i S_{N(t)-i}\right). \tag{10.146}$$

Doyen and Gaudoin (2002) call this model arithmetic reduction of age with infinite memory (ARA$_\infty$). Observe that when $\rho = 0$, we get $w_C(t) = w_1(t)$ and have an NHPP. When $\rho = 1$, we get $w_C(t) = w_1(t - S_{N(t)})$ which represents that the repair action leaves the item in an as-good-as-new condition. The NHPP and the renewal process are therefore special cases of the ARA$_\infty$ model.

Malik (1979) introduces a model in which the repair action at time S_i reduces the last operating time from $S_i - S_{i-1}$ to $\rho(S_i - S_{i-1})$ where as before, $0 \le \rho \le 1$. Using this model, Shin et al. (1996) develop an optimal maintenance policy and derive estimates for the various parameters. The corresponding conditional ROCOF is

$$w_C(t) = w_1(t - \rho S_{N(t)}).$$

The minimal wear intensity is equal to $w_1((1-\rho)t)$. This model is by Doyen and Gaudoin (2002) called arithmetic reduction of age with memory one (ARA$_1$).

In analogy with the failure rate reduction models, we may define a model called arithmetic reduction of age with memory m by

$$w(t) = w_1\left(t - \rho \sum_{i=0}^{\min\{m-1, N(t)\}} (1-\rho)^i S_{N(t)-i}\right).$$

The minimal wear intensity is

$$w_{\min}(t) = w_1((1-\beta)^m t).$$

10.5.4 Trend Renewal Process

Let $S_1, S_2, \ldots$ be the failure times of an NHPP with ROCOF $w(t)$, and let $W(t)$ be the mean number of failures in the interval $(0, t]$. Section 10.4.2 shows that the time-transformed process with occurrence times $W(S_1), W(S_2), \ldots$ is an HPP with rate 1. In the transformed process, the mean time between failures (and renewals) will then be 1. Lindqvist (1998) generalizes this model, by replacing the HPP with rate 1 with a renewal process with underlying distribution $F(\cdot)$ with mean 1. He called the resulting process a trend-renewal process, TRP(F, w). To specify the process, we need to specify the rate $w(t)$ of the initial NHPP and the distribution $F(t)$.

If we have a TRP(F, w) with failure times $S_1, S_2, \ldots$, the time-transformed process with occurrence times $W(S_1), W(S_2), \ldots$ is a renewal process with underlying distribution $F(t)$. The transformation is illustrated in Figure 10.22. The requirement that $F(t)$ has mean value 1 is made for convenience. The scale is then taken care of by the rate $w(t)$.

Lindqvist (1998) shows that the conditional ROCOF of the TRP(F, w) is

$$w_C^{\text{TRP}}(t) = z[W(t) - W(S_{N(t-)})]w(t), \tag{10.147}$$

where $z(t)$ is the failure rate (FOM) function of the distribution $F(t)$. The conditional ROCOF of the TRP(F, w) is hence a product of a factor, $w(t)$, that depends

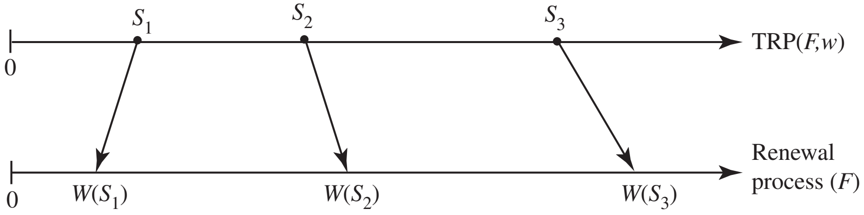


Figure 10.22 Illustration of the transformation of a TRP(F, w) to a renewal process.

on the age t of the item, and a factor that depends on the (transformed) time from the previous failure. When both the failure rate (FOM) function $z(t)$ and the initial ROCOF $w(t)$ are increasing functions, then the conditional ROCOF (10.147) at time t after a failure at time s_0 is

$$z[W(t + s_0) - W(s_0)]w(t + s_0).$$

To check the properties of the trend renewal process (TRP), we may look at some special cases:

- If $z(t) = \lambda$, and $w(t) = \beta$ are both constant, the conditional ROCOF is also constant, $w_C(t) = \lambda\beta$. Hence the HPP is a special case of the TRP.
- If $z(t) = \lambda$ is constant, the conditional ROCOF is $w_C(t) = \lambda w(t)$, and the NHPP is hence a special case of the TRP.
- If $z(0) = 0$, the conditional ROCOF is equal to 0 just after each failure, that is, $w_C(S_{N(t_+)}) = 0$.
- If $w(t) = \beta$ is constant, we have an ordinary renewal process, $w_C(t) = z(t - S_{N(t_-)})$.
- If $z(0) > 0$, The conditional ROCOF just after a failure is $z(0)w(S_{N(t_+)})$ and is increasing with t when $w(t)$ is an increasing function.
- If $z(t)$ is the failure rate (FOM) function of a Weibull distribution with shape parameter α and $w(t)$ is a power law (Weibull) process with shape parameter β, the conditional ROCOF will have a Weibull form with shape parameter $\alpha\beta - 1$.

Example 10.19 Consider a trend renewal process with initial ROCOF $w(t) = 2\theta^2 t$, that is, a linearly increasing ROCOF, and a distribution $F(t)$ with failure rate (FOM) function $z(t) = 2.5\ \lambda^{2.5}\ t^{1.5}$, that is, a Weibull distribution with shape parameter $\alpha = 2.5$ and scale parameter λ. For the mean value of $F(t)$ to be equal to 1, the scale parameter must be $\lambda \approx 0.88725$. The conditional ROCOF in the interval until the first failure is from (10.147)

$$w_C(t) = 5\ \lambda^{2.5}\ \theta^5 t^4 \qquad \text{for} \qquad 0 \le t < S_1.$$

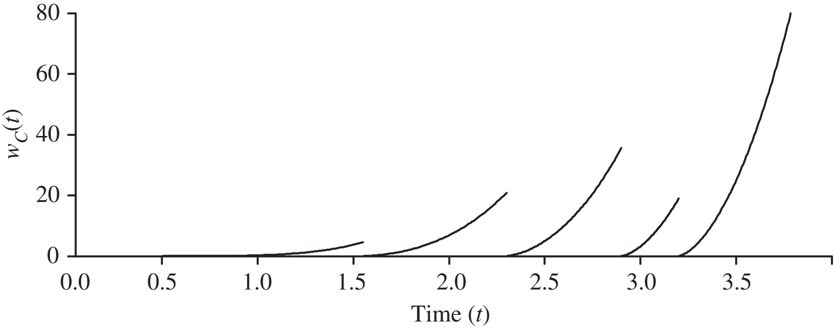


Figure 10.23 Illustration of the conditional ROCOF $w_C(t)$ in Example 10.19 for some possible failure times.

Just after the first failure, $w_C(S_{1+}) = 0$. Generally, we can find $w_C(t)$ from (10.147). Between failure n and failure $n + 1$, the conditional ROCOF is

$$w_C(t) = 5\ \lambda^{2.5}\ \theta^5(t^2 - S_n^2)^{1.5}t \qquad \text{for} \qquad S_n \le t < S_{n+1}.$$

The conditional ROCOF $w_C(t)$ is illustrated for some possible failure times $S_1, S_2, \dots$ in Figure 10.23. □

The trend renewal process is further studied by Lindqvist (1998) and Elvebakk (1999), who also provides estimates for the parameters of the model.

10.6 Model Selection

A simple framework for model selection for a repairable item is shown in Figure 10.24. Figure 10.24 is inspired by a figure in Ascher and Feingold (1984), but new aspects have been added.

We will illustrate the model selection framework by a simple example. In offshore and onshore reliability data (OREDA), failure data from 449 pumps were collected from 61 different installations. A total of 524 critical failures were recorded, that is, on the average 1.17 failures per pump. To get adequate results, we have to merge failure data from several valves. It is important that the data that are merged are homogeneous, meaning that the valves are of the same type and that the operating contexts are comparable. Because there are very few data from each valve, this analysis will have to be qualitative. The total data set should be split into homogeneous subsets and each subset has to be analyzed separately.

We now continue with a subset of the data that is deemed to be homogeneous. The next step is to check whether or not there is a trend in the ROCOF. This may

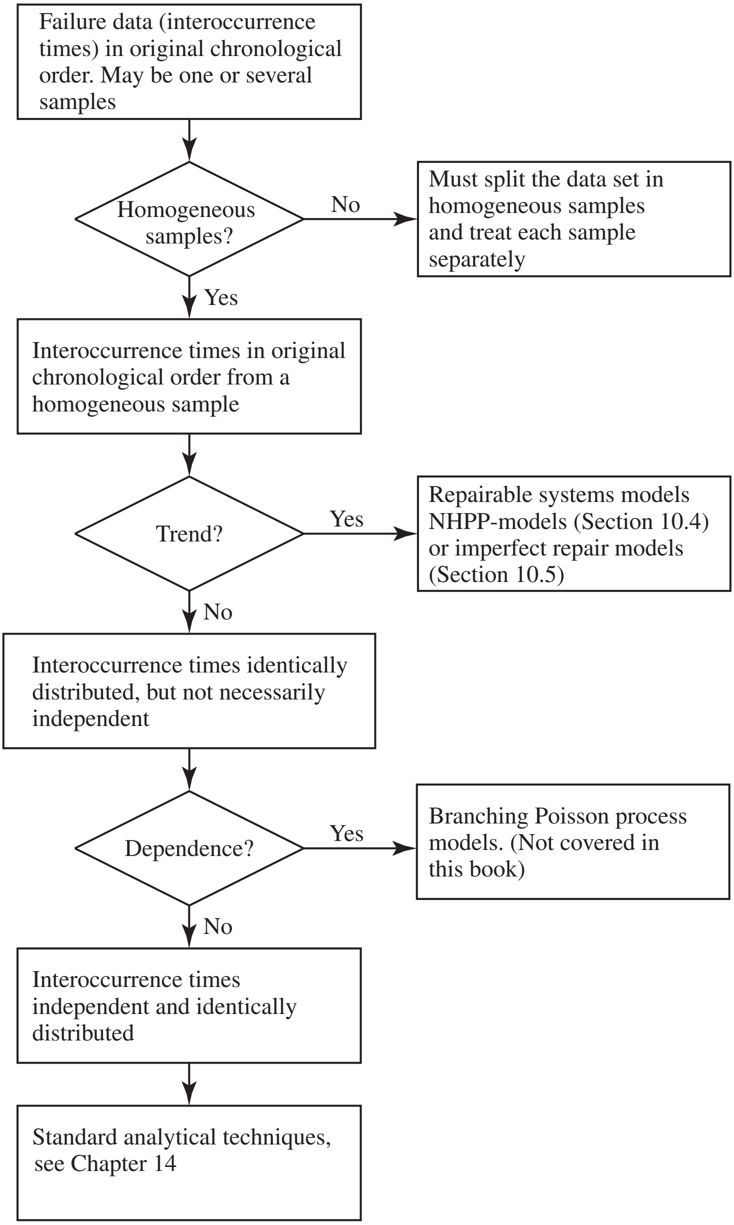

Failure data (interoccurrence times) in original chronological order. May be one or several samples
Homogeneous samples?
No
Must split the data set in homogeneous samples and treat each sample separately
Yes
Interoccurrence times in original chronological order from a homogeneous sample
Trend?
Yes
Repairable systems models NHPP-models (Section 10.4) or imperfect repair models (Section 10.5)
No
Interoccurrence times identically distributed, but not necessarily independent
Dependence?
Yes
Branching Poisson process models. (Not covered in this book)
No
Interoccurrence times independent and identically distributed
Standard analytical techniques, see Chapter 14

Figure 10.24 Model selection framework.

be done by establishing a Nelson–Aalen plot as described in Chapter 14. If the plot is approximately linear, we conclude that the ROCOF is close to constant. If the plot is convex (concave), we conclude that the ROCOF is increasing (decreasing). The ROCOF may also be increasing in one part of the life length and decreasing in another part.

If we conclude that the ROCOF is increasing (or decreasing), we may use either a NHPP or one of the imperfect repair models described in Section 10.5. Which model to use must (usually) be decided by a qualitative analysis of the repair actions, whether it is a minimal repair or and age, or failure rate, reduction repair. In some cases, we may have close to minimal repairs during a period followed by a major overhaul. In the Norwegian offshore sector, such overhauls are often carried our during annual revision stops. When we have decided a model, we may use the methods described in this chapter to analyze the data. More detailed analyzes are described, for example, in Crowder et al. (1991).

If no trend in the ROCOF is detected, we conclude that the intervals between failures are identically distributed, but not necessarily independent. The next step is then to check whether or not the data may be considered as independent. Several plotting techniques and formal tests are available, but these methods are not covered in this book. An introduction to such methods may, for example, be found in Crowder et al. (1991).

If we can conclude that the intervals between failures are independent and identically distributed, we have a renewal process, and we can use the methods described in Chapter 14 to analyze the data.

If the intervals are dependent, we have to use methods that are not described in this book. Please consult, for example, Crowder et al. (1991) for relevant approaches.

10.7 Problems

10.1 Consider an HPP $\{N(t), t \geq 0\}$ and let $t, s \geq 0$. Determine

$$E[N(t)\, N(t+s)].$$

10.2 Consider an HPP $\{N(t), t \geq 0\}$ with rate $\lambda > 0$. Verify that

$$\Pr(N(t) = k \mid N(s) = n) = \binom{n}{k} \left(\frac{t}{s}\right)^k \left(1 - \frac{t}{s}\right)^{n-k}$$
$$\text{for} \quad 0 < t < s \quad \text{and} \quad 0 \leq k \leq n.$$

10.3 Let T_1 be the time to the first occurrence of an HPP $\{N(t), t \geq 0\}$ with rate λ.

(a) Show that

$$\Pr(T_1 \leq s \mid N(t) = 1) = \frac{s}{t} \qquad \text{for} \qquad s \leq t.$$

(b) Determine $E(T_1)$ and $\text{SD}(T_1)$.

10.4 Let $\{N_1(t); t \geq 0\}$ and $\{N_2(t); t \geq 0\}$ be two independent HPPs with rates λ_1 and λ_2, respectively. Let $N(t) = N_1(t) + N_2(t)$ and show that $\{N(t); t \geq 0\}$ is an HPP with rate $\lambda_1 + \lambda_2$.

10.5 Let $\{N(t), t \geq 0\}$ be a counting process, with possible values $0, 1, 2, 3, \ldots$. Show that the mean value of $N(t)$ can be written

$$E[N(t)] = \sum_{n=1}^{\infty} \Pr(N(t) \geq n) = \sum_{n=0}^{\infty} \Pr(N(t) > n). \tag{10.148}$$

10.6 Let $S_1, S_2, \ldots$ be the occurrence times of an HPP $\{N(t), t \geq 0\}$ with rate λ. Assume that $N(t) = n$. Show that the random variables $S_1, S_2, \ldots, S_n$ have the joint probability density function

$$f_{S_1,\ldots,S_n|N(t)=n}(s_1, \ldots, s_n) = \frac{n!}{t^n} \qquad \text{for} \qquad 0 < s_1 < \cdots < s_n \leq t.$$

10.7 Consider a renewal process $\{N(t), t \geq 0\}$. Is it true that:

(a) $N(t) < r$ if and only if $S_r > t$?

(b) $N(t) \leq r$ if and only if $S_r \geq t$?

(c) $N(t) > r$ if and only if $S_r < t$?

10.8 Consider an NHPP with rate

$$w(t) = \lambda \left(\frac{t+1}{t} \right) \qquad \text{for} \qquad t \geq 0.$$

(a) Make a sketch of $w(t)$ as a function of t.

(b) Make a sketch of the cumulative ROCOF, $W(t)$, as a function of t.

10.9 Consider an NHPP $\{N(t), t \geq 0\}$ with rate:

$$w(t) = \begin{cases} 6 - 2t \text{ for } 0 \leq t \leq 2 \\ 2 \text{ for } 2 < t \leq 20 \\ -18 + t \text{ for } t > 20 \end{cases}.$$

(a) Make a sketch of $w(t)$ as a function of t.

(b) Make a sketch of the corresponding cumulative ROCOF, $W(t)$, as a function of t.

(c) Estimate the number of failures/events in the interval $(0, 12)$

10.10 Section 10.3.8 claims that the superposition of independent renewal processes is generally *not* a renewal process. Explain why the superposition of independent HPPs is a renewal process. What is the renewal density of this superimposed process?

10.11 Atwood (1992) applies the following parametrization of the power law model, the linear model and the log-linear model:

$$\begin{aligned}
w(t) &= \lambda_0(t/t_0)^{\beta} && \text{(power law model)}\\
w(t) &= \lambda_0[1 + \beta(t - t_0)] && \text{(linear model)}\\
w(t) &= \lambda_0\; e^{\beta(t-t_0)} && \text{(log-linear model).}
\end{aligned}$$

(a) Discuss the meaning of t_0 in these models.

(b) Show that Atwood's parameterization is compatible with the parameterization used in Section 10.4.4.

(c) Show that $w(t) = \lambda_0$ when $t = t_0$ for all the three models.

(d) Show that $w(t)$ is increasing if $\beta > 0$, is constant if $\beta = 0$, and decreasing if $\beta < 0$, for all the three models.

10.12 Use the MIL-HDBK test described in Section 10.4.4 to check if the "increasing trend" of the data in Example 10.1 is significant (use 5% level).

10.13 The objective of this problem is to study different counting processes and to create a procedure to asses the performance of a maintained system when the repair duration is negligible compared to the item's lifetime.

(a) Assume that the item is repaired after each failure to an as-good-as-new state and that its failure rate is constant ($\lambda = 5 \times 10^{-4}$) per hour between two failures.

i. What kind of counting process is it? What are the MTTF and the MTBF of the item?

ii. Consider the data set in Table 10.2, which can be downloaded from the `book companion site`. Each column provides a sequence of failure times for one item (in hours). All the items are identical and are operated in the same conditions. S_1 is the first failure time for each item, S_2 the second failure time, and so on. Make a plot of $N_1(t), \ldots, N_5(t)$ as a function of time t in the same figure, where the y-axis is the number of failures (from 0 to 10) and the x-axis is the time (failure times).

Table 10.2 Data set for Problem 13.

	Item 1	Item 2	Item 3	Item 4	Item 5
S_1	2099	2504	4081	1015	382
S_2	5352	3060	5210	3686	1621
S_3	8116	3626	6722	4535	1629
S_4	9085	5559	15584	5279	6726
S_5	10581	6691	17759	5860	8356
S_6	12672	11848	21397	7454	12832
S_7	13042	17688	21858	12412	12910
S_8	14114	18955	24192	15361	23659
S_9	15310	19454	25468	15542	24169
S_{10}	15483	19590	29063	19305	24572

iii. Make a plot of $E[N(t)]$, by using all the failure times in Table 10.2 and by using the exact formula from Section 10.2.1. Plot both of them in the same figure and comment if the number of failure times is sufficient to provide an accurate estimate of $E[N(t)]$.

iv. Determine the probability $P_t^k = \Pr(N(t) = k)$ for $k = 1, 2, \dots, 5$, that an item will experience k failures in $(0, t)$ for $k = 0, 1, 2$. Make a plot for each k in the same figure, give the times upon which the probabilities P_t^k cross each others and give an explanation for such "crossing" times.

v. Let $t = \text{MTBF}$. Determine the probabilities $P^k_{\text{MTBF}} = \Pr(N(\text{MTBF}) = k)$ to have k failures at the time equal to MTBF for different values of k. Chose k such that you obtain intuitively expected results.

(b) Assume that the item is repaired to an as-good-as-new state after each failure but that its failure rate is no more constant between two failures. Further, assume that the time-to-failure is Weibull distributed with shape parameter 4 and scale parameter 500.

i. What kind of counting process is this?

ii. Which procedure can you apply to get an empirical expression of $E[N(t)]$?

iii. Find the approximated value of $E[N(t)]$ when t is high.

10.14 Consider an item that is repaired to an as-good-as-new state after each failure and with a failure rate that is constant ($\lambda = 5 \times 10^{-4}\ \text{h}^{-1}$) between two failures. The mean downtime (MDT) is six hours. By using theoretical formulas, determine the average availability of the item and the average number of hours per year that it is out of operation. Do you need any further assumptions to use these formulas? If "yes," record the assumptions made.

References

Akersten, P.A. (1991). Repairable systems reliability. Studied by TTT-plotting techniques. PhD thesis. Linköping, Sweden: Linköping University.

Akersten, P.A. (1998). Imperfect repair models. In: *Proceedings of the European Conference on Safety and Reliability–ESREL'98* (ed. S. Lydersen, G.K. Hansen, and H.A. Sandtorv), 369–372. Rotterdam, The Netherlands: A. A. Balkema.

Andersen, P.K., Borgan, Ø., Gill, R.D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer.

Ascher, H. and Feingold, H. (1984). *Repairable Systems Reliability; Modeling, Inference, Misconceptions, and Their Causes*. New York: Marcel Dekker.

Atwood, C.L. (1992). Parametric estimation of time-dependent failure rates for probabilistic risk assessment. *Reliability Engineering and System Safety* 37 (3): 181–194.

Bain, L.J., Engelhardt, M., and Wright, F.T. (1985). Test for an increasing trend in the intensity of a Poisson process. *Journal of the American Statistical Association* 80: 419–422.

Barlow, R.E. and Proschan, F. (1965). *Mathematical Theory of Reliability*. New York: Wiley.

Barlow, R.E. and Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing, Probability Models*. New York: Holt, Rinehart, and Winston.

Bendell, A. and Walls, L.A. (1985). Exploring reliability data. *Quality and Reliability Engineering* 1: 37–51.

Block, H.W., Borges, W.S., and Savits, T.H. (1985). Age-dependent minimal repair. *Journal of Applied Probability* 22 (2): 370–385.

Bon, J.L. (1995). *Fiabilité des Systèmes, Méthodes Mathématiques*. Paris: Masson.

Brown, M. and Proschan, F. (1983). Imperfect repair. *Journal of Applied Probability* 20 (4): 851–859.

Cha, J.H. and Finkelstein, M. (2018). *Point Processes for Reliability Analysis: Shocks and Repairable Systems*. Springer.

Chan, J.K. and Shaw, L. (1993). Modeling repairable systems with failure rates that depend on age and maintenance. *IEEE Transactions on Reliability* 42 (4): 566–571.

Cocozza-Thivent, C. (1997). *Processus Stochastiques et Fiabilité des Systèmes*. Paris: Springer (in French).

Cox, D.R. (1962). *Renewal Theory*. Methuen, London: Methuen & Co.

Cox, D.R. and Isham, V. (1980). *Point Processes*. London: Chapman and Hall.

Cox, D.R. and Lewis, P.A. (1966). *The Statistical Analysis of Series of Events*. Methuen, London: Methuen & Co.

Crowder, M.J., Kimber, A.C., Sweeting, T.J., and Smith, R.L. (1991). *Statistical Analysis of Reliability Data*. Boca Raton, FL: CRC Press / Chapman and Hall.

Dohi, T., Kaio, N., and Osaki, S. (2002). Renewal processes and their computational aspects. In: *Stochastic Models in Reliability and Maintenance* (ed. T. Dohi), 1–30. Berlin: Springer-Verlag.

Doyen, L. and Gaudoin, O. (2002). Models for assessing maintenance efficiency. *Proceedings from ESREL Conference*, Trondheim, Norway.

Doyen, L. and Gaudoin, O. (2011). Modeling and assessment of aging and efficiency of corrective and planned preventive maintenance. *IEEE Transactions on Reliability* 60 (4): 759–769.

Drenick, R.F. (1960). The failure law of complex equipment. *Journal of the Society of Industrial Applied Mathematics* 8 (4): 680–690.

Elvebakk, G. (1999). Analysis of repairable systems data: statistical inference for a class of models involving renewals, heterogenity, and time trends. PhD thesis. Trondheim, Norway: Norwegian University of Science and Technology.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, vol. 1. New York: Wiley.

Gaudoin, O. (1992). Optimal properties of the Laplace trend test for software-reliability models. *IEEE Transactions on Reliability* 41 (4): 525–532.

Hokstad, P.R. (1997). The failure intensity process and the formulation of reliability and maintenance models. *Reliability Engineering and System Safety* 58: 69–82.

Lim, T.J. (1998). Estimating system reliability with fully masked data under Brown-Proschan imperfect repair model. *Reliability Engineering and System Safety* 59: 277–289.

Lindqvist, B.H. (1998). Statistical and probabilistic models in reliability. In: *Statistical Modeling and Analysis of Repairable Systems* (ed. D.C. Ionesco and N. Limnios), 3–25. Boston, MA: Birkhauser.

Lorden, G. (1970). On excess over the boundary. *Annals of Mathematical Statistics* 41: 520–527.

Malik, M.A.K. (1979). Reliable preventive maintenance policy. *AIII Transactions* 11: 221–228.

MIL-HDBK-189C (2011). Reliability growth management, *Military handbook*. Washington, DC: U.S. Department of Defense.

Pham, H. and Wang, H. (1996). Imperfect maintenanece. *European Journal of Operational Research* 94 (3): 425–438.

Ross, S.M. (1996). *Stochastic Processes*. New York: Wiley.

Shin, I., Lim, T.J., and Lie, C.H. (1996). Estimating parameters of intensity function and maintenance effect for repairable unit. *Reliability Engineering and System Safety* 54: 1–10.

Smith, W.L. (1958). Renewal theory and its ramifications. *Journal of the Royal Statistical Society* 20: 243–302.

Smith, W.L. and Leadbetter, M.R. (1963). On the renewal function of the Weibull distribution. *Technometrics* 5: 393–396.

Takács, L. (1956). On a probability problem arising in the theory of counters. *Proceedings of the Cambridge Philosophical Society* 32 (3): 488–498.

Thompson, W.A.J. (1981). On the foundations of reliability. *Technometrics* 23: 1–13.

Vesely, W.E. (1991). Incorporating aging effects into probabilistic risk analysis using a Taylor expansion approach. *Reliability Engineering and System Safety* 32 (3): 315–337.

11

Markov Analysis

11.1 Introduction

The models in the preceding chapters are all based on the assumption that the components and the systems can be in one out of two possible states: a *functioning state* or a *failed state*. We have also seen that the models are rather static and not well suited for analysis of repairable systems.

Stochastic processes are introduced in Chapter 10. This chapter introduces a specific type of stochastic processes, called Markov[1] chains, to model systems with several states and *transitions* between the states. A Markov chain is a stochastic process $\{X(t), t \geq 0\}$ having the Markov property. (The Markov property is defined in Section 11.1.1.) The random variable $X(t)$ is the *state* of the process at *time t*. The collection of all possible states is called the *state space*, and is denoted $\mathcal{X}$. The state space $\mathcal{X}$ is either finite or countable infinite. In most applications, the state space is *finite*, and the states correspond to real states of a system (see Example 11.1). Unless stated otherwise, $\mathcal{X}$ is taken to be $\{0, 1, 2, \dots, r\}$, such that $\mathcal{X}$ contains $r + 1$ different states. The time may be discrete, taking values in $\{0, 1, 2, \dots\}$, or continuous. When the time is discrete, we have a *discrete-time Markov chain*, and when the time is continuous, we have a *continuous-time Markov chain*. Many authors use the term *Markov process* for a continuous-time Markov chain. This term is also used in the current book. When the time is discrete, we denote the time by n and the discrete-time Markov chain by $\{X_n, n = 0, 1, 2, \dots\}$.

The theoretical basis for Markov chains is presented briefly in this book, and it is recommended to consult a textbook on stochastic processes for more details. An excellent introduction to Markov chains may be found in, for example, Ross (1996). Continuous-time Markov chains and their application in reliability engineering is treated by Cocozza-Thivent (1997), Pukite and Pukite (1998), and Trivedi and Bobbio (2017).

[1] Named after the Russian mathematician Andrei A. Markov (1856–1922).

System Reliability Theory: Models, Statistical Methods, and Applications, Third Edition.
Marvin Rausand, Anne Barros, and Arnljot Høyland.

Companion Website: www.wiley.com/go/SystemReliabilityTheory3e

The main focus in this book is on continuous-time Markov chains and how these chains can be used to model the reliability of a system. This chapter starts by defining the Markov property and continuous-time Markov chains. A set of linear, first-order differential equations, called the *Kolmogorov equations*, are established to determine the probability distribution $\boldsymbol{P}(t) = [P_0(t), P_1(t), \dots , P_r(t)]$ of the chain at time t, where $P_i(t)$ is the probability that the chain (the system) is in state i at time t. We then show that $\boldsymbol{P}(t)$, under specific conditions, will approach a limit $\boldsymbol{P}$ when $t \to \infty$. This limit is called the *steady state* distribution of the chain (the system). Several system performance metrics such as state visit frequency, system availability, and mean time to first system failure, are introduced. The steady state distribution and system performance metrics are then determined for some simple systems, such as series and parallel systems, systems with dependent components, and various types of standby systems. The time-dependent solution of the Kolmogorov equations is briefly discussed. The chapter ends with a brief discussion and an introduction to semi-Markov, multiphase, and piecewise deterministic Markov processes (PDMPs). They are generalizations of the continuous-time Markov chains and may be used to model many maintained systems.

Example 11.1 (States of a parallel structure)
Consider a parallel structure of two components. Each component is assumed to have two states, a functioning state (1), and a failed state (0). Because each of the components has two possible states, the parallel structure has $2^2 = 4$ possible states. These states are listed in Table 11.1. The state space is therefore $\mathcal{X} = \{0, 1, 2, 3\}$. The structure is fully functioning when the state is 3, and failed when the state is 0. In states 1 and 2, the system is operating with only one component in function. □

When the structure has n components, and each component has two states (functioning, and failed), the structure has at most 2^n different states. In some

Table 11.1 Possible states of a structure of two components.

State	Component 1	Component 2
3	Functioning	Functioning
2	Functioning	Failed
1	Failed	Functioning
0	Failed	Failed

applications, we may introduce more than two states for each component. A pump may, for example, have three states: operating, standby, or failed. A producing item may, for example, operate with 100% capacity, 80% capacity, and so on. In other applications, it is important to distinguish the various failure modes of an item, and we may define the various failure modes as states. For a complicated structure, the number of states may become overwhelming, and it may be necessary to simplify the system model, and separately consider modules of the structure.

11.1.1 Markov Property

Consider a chain that is started at time 0. If $X(s) = i$, the chain is said to be in state i at time s. The conditional probability that the chain is in state j at time $t + s$, when it was in state i at time s is

$$\Pr(X(t+s) = j \mid X(s) = i, X(u) = x(u), \quad 0 \le u < s).$$

Definition 11.1 (Markov property)
A continuous time Markov chain $\{X(t), t \ge 0\}$ is said to have the Markov property when

$$\begin{aligned}\Pr(X(t+s) = j \mid X(s) = i, X(u) = x(u),\ 0 \le u < s)\\ = \Pr(X(t+s) = j \mid X(s) = i) \ \text{ for all possible } \ x(u), \quad 0 \le u < s.\end{aligned} \tag{11.1}$$

□

In other words, when the *present* state of the chain is known, the future development of the chain is independent of anything that has happened in the past. A chain satisfying the Markov property (11.1) is a *continuous-time Markov chain*, but will in the following be called a *Markov process*.

Further, assume that the Markov process for all i, j in $\mathcal{X}$ fulfills

$$\Pr(X(t+s) = j \mid X(s) = i) = \Pr(X(t) = j \mid X(0) = i) \qquad \text{for all } s, t \ge 0,$$

which says that the probability of a transition from state i to state j does not depend on the global time, and only depends on the time interval available for the transition. A process with this property is known as a process with *stationary transition probabilities*, or a *time-homogeneous* process.

From now on, we only consider Markov processes (i.e. chains fulfilling the Markov property) that have stationary transition probabilities. A consequence of this assumption is that a Markov process cannot be used to model a system where the transition probabilities are influenced by long-term trends and/or seasonal variations. To use a Markov model, we have to assume that the environmental and operational conditions for the system are relatively stable as a function of time.

11.2 Markov Processes

Consider a Markov process $\{X(t), t \geq 0\}$ with state space $\mathcal{X} = \{0, 1, 2, \dots, r\}$ and stationary transition probabilities. The transition probabilities of the Markov process

$$P_{ij}(t) = \Pr(X(t) = j \mid X(0) = i) \qquad \text{for all} \qquad i,j \in \mathcal{X},$$

may be arranged as a matrix

$$\mathbb{P}(t) = \begin{pmatrix} P_{00}(t) & P_{01}(t) & \cdots & P_{0r}(t) \\ P_{10}(t) & P_{11}(t) & \cdots & P_{1r}(t) \\ \vdots & \vdots & \ddots & \vdots \\ P_{r0}(t) & P_{r1}(t) & \cdots & P_{rr}(t) \end{pmatrix}. \tag{11.2}$$

Because all entries in $\mathbb{P}(t)$ are probabilities,

$$0 \leq P_{ij}(t) \leq 1 \qquad \text{for all } \ t \geq 0, \ \ i,j \in \mathcal{X}.$$

When a process is in state i at time 0, it must either be in state i at time t or have made a transition to a different state. This means that

$$\sum_{j=0}^{r} P_{ij}(t) = 1 \qquad \text{for all } \ i \in \mathcal{X}. \tag{11.3}$$

The sum of each *row* in the matrix $\mathbb{P}(t)$ is therefore equal to 1. Observe that the entries in row i represent the transitions out of state i (for $j \neq i$), and that the entries in *column* j represent the transition into state j (for $i \neq j$).

Let $0 = S_0 \leq S_1 \leq S_2 \leq \dots$ be the times at which transitions occur, and let $T_i = S_{i+1} - S_i$ be the ith interoccurrence time, or *sojourn time*, for $i = 1, 2, \dots$. The sojourn time in state i is hence the length of time of a visit to state i. Assume that the transition takes place immediately before time S_i such that the trajectory of the process is continuous from the right. A possible trajectory of a process is illustrated in Figure 11.1.

Consider a Markov process that enters state i at time 0, such that $X(0) = i$. Let $\widetilde{T}_i$ be a generic sojourn time in state i. Observe that T_i denotes the ith sojourn time, whereas $\widetilde{T}_i$ is the time spent during a visit to *state i*. We want to find the probability $\Pr(\widetilde{T}_i > t)$. Assume now that the process is still in state i at time s, that is, $\widetilde{T}_i > s$, and that we are interested in finding the probability that it will remain in state i for t time units more. Hence, we want to find $\Pr(\widetilde{T}_i > t + s \mid \widetilde{T}_i > s)$. Because the process has the Markov property, the probability for the process to stay t time units

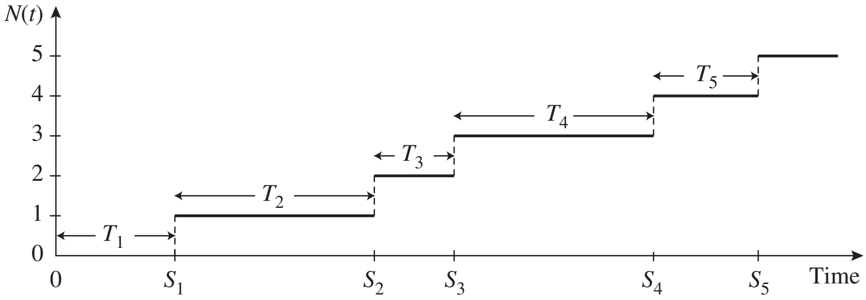


Figure 11.1 Trajectory of a Markov process.

more is determined only by the current state i. The fact that the process has been staying there for s time units is therefore irrelevant. Thus,

$$\Pr(\widetilde{T}_i > t + s \mid \widetilde{T}_i > s) = \Pr(\widetilde{T}_i > t) \qquad \text{for } s, t \geq 0.$$

Hence, the random variable $\widetilde{T}_i$ is memoryless and must be exponentially distributed.

The sojourn times $T_1, T_2, \ldots$ must therefore also be independent and exponentially distributed. The independence follows from the Markov property. See Ross (1996) for a more detailed discussion.

Let $X_n = X(S_n)$. The process $\{X_n, n = 1, 2, \ldots\}$ is called the *skeleton* of the (continuous-time) Markov process. Transitions of the skeleton occur at discrete times $n = 1, 2, \ldots$. The skeleton may be imagined as a process where all the sojourn times are deterministic and of equal length. It is straightforward to show that the skeleton of a (continuous-time) Markov process is a discrete-time Markov chain, see Ross (1996). The skeleton is also called the *embedded* Markov chain.

A Markov process may now be constructed as a stochastic process having the properties that each time it enters a state i (see Ross 1996):

(1) The amount of time the process spends in state i before making a transition into a different state is exponentially distributed with rate, say α_i.
(2) When the process leaves state i, it will next enter state j with some probability P_{ij}, where $\sum_{\substack{j=0 \\ j\neq i}}^{r} P_{ij} = 1$.

The mean sojourn time in state i is therefore

$$E(\widetilde{T}_i) = \frac{1}{\alpha_i}.$$

If $\alpha_i = \infty$, state i is called an *instantaneous* state, because the mean sojourn time in such a state is zero. When the Markov process enters such a state, the state is instantaneously left. In this book, we assume that the Markov process has no instantaneous states, and that $0 \leq \alpha_i < \infty$ for all i. If $\alpha_i = 0$, then state i is called *absorbing* because once entered it is never left. Sections 11.2 and 11.3 assume that there are no absorbing states. Absorbing states are further discussed in Section 11.4.

A Markov process may hence be seen as a stochastic process that moves from state to state in accordance with a discrete-time Markov chain. The amount of time it spends in each state – before going to the next state – is exponentially distributed. The amount of time the process spends in state i, and the next state visited, are independent random variables.

Let a_{ij} be defined by

$$a_{ij} = \alpha_i P_{ij} \qquad \text{for all } \ i \neq j. \tag{11.4}$$

Because α_i is the rate at which the process leaves state i and P_{ij} is the probability that it goes to state j, it follows that a_{ij} is the rate when in state i the process makes a transition into state j. We call a_{ij} the *transition rate* from i to j.

Because $\sum_{j\neq i} P_{ij} = 1$, it follows from (11.4) that

$$\alpha_i = \sum_{\substack{j=0 \\ j\neq i}}^{r} a_{ij}. \tag{11.5}$$

Because the sojourn times are exponentially distributed,

$$\begin{aligned} &\Pr(\widetilde{T}_i > t) = e^{-\alpha_i t} \\ &\Pr(T_{ij} \leq t) = 1 - e^{-a_{ij} t} \qquad \text{for } \ i \neq j, \end{aligned}$$

where T_{ij} is the time the chain spends in state i before entering into state j. We therefore have that (Remember that $e^x = \sum_{k=0}^{\infty} x^k/k!$)

$$\lim_{\Delta t\to 0} \frac{1 - P_{ii}(\Delta t)}{\Delta t} = \lim_{\Delta t\to 0} \frac{\Pr(\widetilde{T}_i < \Delta t)}{\Delta t} = \alpha_i \tag{11.6}$$

$$\lim_{\Delta t\to 0} \frac{P_{ij}(\Delta t)}{\Delta t} = \lim_{\Delta t\to 0} \frac{\Pr(T_{ij} < \Delta t)}{\Delta t} = a_{ij} \qquad \text{for } \ i \neq j. \tag{11.7}$$

For proof, see Ross (1996).

Because we, from (11.4) and (11.5), can deduce α_i and P_{ij} when we know a_{ij} for all i,j in $\mathcal{X}$, we may equally well define a continuous-time Markov chain by specifying; (i) The state space $\mathcal{X}$, and (ii) the transition rates a_{ij} for all $i \neq j$ in $\mathcal{X}$. This second definition is often more natural, and is our main approach.

The transition rates a_{ij} may be arranged as a matrix,

$$\mathbb{A} = \begin{pmatrix} a_{00} & a_{01} & \cdots & a_{0r} \\ a_{10} & a_{11} & \cdots & a_{1r} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r0} & a_{r1} & \cdots & a_{rr} \end{pmatrix}, \tag{11.8}$$

where the following notation for the diagonal elements is introduced:

$$a_{ii} = -\alpha_i = -\sum_{\substack{j=0 \\ j\neq i}}^{r} a_{ij}. \tag{11.9}$$

We call $\mathbb{A}$ the *transition rate matrix* of the Markov chain. Some authors refer to the matrix $\mathbb{A}$ as the *infinitesimal generator* of the chain.

Observe that the entries of row i are the transition rates out of state i (for $j \neq i$). We call them *departure rates* from state i. According to (11.5) $-a_{ii} = \alpha_i$ is the sum of the departure rates from state i, and hence the *total departure rate* from state i. Observe that the sum of the entries in row i is equal to 0, for all $i \in \mathcal{X}$. The entries of column i are transition rates into state i (for $j \neq i$).

11.2.1 Procedure to Establish the Transition Rate Matrix

To establish the transition rate matrix $\mathbb{A}$, we have to:

(1) List and describe all relevant system states. Nonrelevant states should be removed, and identical states should be merged (e.g. see Example 11.3). Each of the remaining states must be given a unique identification. This book uses the integers from 0 up to r, where r denotes the best functioning state of the system, and 0 denotes the worst state. The state space of the system is thus $\mathcal{X} = \{0, 1, \dots, r\}$, but any other sequence of numbers, or letters may be used.
(2) Specify all transition rates a_{ij} for all $i \neq j$ and $i, j \in \mathcal{X}$. Each transition usually involves a failure or a repair. The transition rates are therefore failure rates and repair rates, and combinations of these.
(3) Arrange the transition rates a_{ij} for $i \neq j$ as a matrix, similar to the matrix (11.8) (leave the diagonal entries a_{ii} open).
(4) Fill in the diagonal elements a_{ii} such that the sum of all entries in each *row* is equal to zero, or by using (11.9).

A Markov chain may be represented graphically by a *state transition diagram* that records the a_{ij} of the possible transitions of the Markov chain. The state transition diagram is also known as a Markov diagram. In the state transition diagram, circles are used to represent states, and directed arcs are used to represent transitions between the states. An example of a state transition diagram is given in Figure 11.2.

Example 11.2 (Parallel structure – cont.)
Reconsider the parallel structure of two independent components in Example 11.1. Assume that the following corrective maintenance strategy is adopted: When a component fails, a repair action is initiated to bring this component back to its

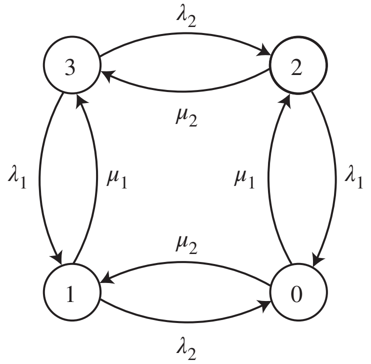


Figure 11.2 State transition diagram of the parallel structure in Example 11.2.

initial functioning state. After the repair is completed, the component is assumed to be as-good-as-new. Each component is assumed to have its own dedicated repair crew.

Assume that the components have constant failure rates λ_i and constant repair rates μ_i, for $i = 1, 2$. The transitions between the four system states in Table 11.1 are illustrated in the state transition diagram in Figure 11.2.

Assume that the system is in a state of 3 at time 0. The first transition may either be to state 2 (failure of component 2), or to state 1 (failure of component 1). The transition rate to state 2 is $a_{32} = \lambda_2$, and the transition rate to state 1 is $a_{31} = \lambda_1$. The sojourn time in state 3 is therefore $\widetilde{T}_3 = \min\{T_{31}, T_{32}\}$, where T_{ij} is the time to the first transition from state i to state j. $\widetilde{T}_3$ is exponentially distributed with rate $a_{31} + a_{32} = \lambda_1 + \lambda_2$, and the mean sojourn time in state 3 is $1/(\lambda_1 + \lambda_2)$.

When the system is in state 2, the next transition may either be to state 3 (with rate $a_{23} = \mu_2$), or to state 0 (with rate $a_{20} = \lambda_1$). The probability that the transition is to state 3 is $\mu_2/(\mu_2 + \lambda_1)$, and the probability that it goes to state 0 is $\lambda_1/(\mu_2 + \lambda_1)$. The memoryless property of the exponential distribution ensures that component 1 is as-good-as-new when the system enters state 2. In this example, we assume that component 1 has the same failure rate λ_1 in state 3, where both components are functioning, as it has in state 2, where only component 1 is functioning. The failure rate a_{20} of component 1 in state 2 may, however, easily be changed to a failure rate λ_1' that is different from (e.g. higher than) λ_1.

When the system is in state 0, both components are in a failed state, and two independent repair crews are working to bring the components back to a functioning state. The repair times T_{01} and T_{02} are independent and exponentially distributed with repair rates μ_1 and μ_2, respectively. The sojourn time $\widetilde{T}_0$ in state 0, $\min\{T_{01}, T_{02}\}$ is exponentially distributed with rate $(\mu_1 + \mu_2)$, and the mean downtime (MDT) of the system is therefore $1/(\mu_1 + \mu_2)$. When the system enters state 0, one of the components will already have failed and be under repair when the other component fails. The memoryless property of the exponential distribution ensures, however, that the time to complete the repair is independent of how long time the component has been under repair.

The transition rate matrix of the system is thus

$$\mathbb{A} = \begin{pmatrix} -(\mu_1 + \mu_2) & \mu_2 & \mu_1 & 0 \\ \lambda_2 & -(\lambda_2 + \mu_1) & 0 & \mu_1 \\ \lambda_1 & 0 & -(\lambda_1 + \mu_2) & \mu_2 \\ 0 & \lambda_1 & \lambda_2 & -(\lambda_1 + \lambda_2) \end{pmatrix}. \tag{11.10}$$

This model disregards the possibility of common-cause failures (CCFs). Thus, a transition between states 3 and 0 is assumed to be impossible during a time interval of length Δt.

Observe that when drawing the state transition diagram, we consider a very short time interval, such that the transition diagram only records events of single transitions. Analogous with the Poisson process, the probability of having two or more events in a short time Δt is $o(\Delta t)$, and hence, events of multiple transitions are not included in the state transition diagram. It is therefore not possible to have a transition from states 1 to 2 in Figure 11.2, because this would involve failure of component 2 and at the same time completed repair of component 1. A CCF can be modeled as a transition from states 3 to 0 in Figure 11.2. Such a transition involves the failure of two components, but may be considered a single event. □

Example 11.3 (Parallel structure – cont.)
Reconsider the parallel structure in Example 11.1, but assume that the two components are independent and identical with the same failure rate λ. In this case, it is not necessary to distinguish between the states 1 and 2 in Table 11.1, and we may reduce the state space to the three states:

2	Both components are functioning
1	One component is functioning and one is failed
0	Both components are in a failed state

Assume that the system is taken care of by a single repair crew, that has adopted a first-fail-first-repair policy. The repair time of a component is assumed to be exponentially distributed with repair rate μ. The mean repair time is then $1/\mu$. The transitions between the three system states are illustrated in Figure 11.3.

A transition from states 2 to 1 takes place as soon as one of the two independent components fails. The transition rate is therefore $a_{21} = 2\lambda$. When the system is

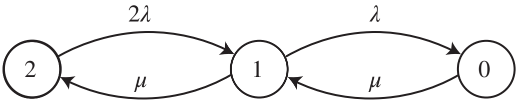


Figure 11.3 State transition diagram for the parallel structure in Example 11.3.

in state 1, it either goes to state 2 (with probability $\mu/(\mu+\lambda)$), or to state 0 (with probability $\lambda/(\mu+\lambda)$).

The transition rate matrix of the system is

$$\mathbb{A} = \begin{pmatrix} -\mu & \mu & 0 \\ \lambda & -(\mu+\lambda) & \mu \\ 0 & 2\lambda & -2\lambda \end{pmatrix}.$$

The mean sojourn times in the three states are seen to be

$$E(\widetilde{T}_0) = \frac{1}{\mu}, \quad E(\widetilde{T}_1) = \frac{1}{\mu+\lambda}, \quad E(\widetilde{T}_2) = \frac{1}{2\lambda},$$

that is, the inverse of the absolute value of the corresponding diagonal entry in $\mathbb{A}$.

An alternative repair strategy for state 0 would be to repair both components at the same time and only start up the system when both components are functioning again. If the repair time for this common repair action has rate μ_C, we have to modify the state transition diagram in Figure 11.2 and introduce $a_{01} = 0$ and $a_{02} = \mu_C$ (a_{12} is still μ). □

Example 11.4 (Homogeneous Poisson process)
Consider a homogeneous Poisson process (HPP) $\{X(t), t \geq 0\}$ with rate λ. The HPP is a Markov process with countable *infinite* state space $\mathcal{X} = \{0, 1, 2, \dots\}$. In this case, we have $\alpha_i = \lambda$ for $i = 0, 1, 2, \dots$, and $a_{ij} = \lambda$ for $j = i + 1$, and 0 for $j \neq i + 1$. The transition rate matrix for the HPP is thus

$$\mathbb{A} = \begin{pmatrix} -\lambda & \lambda & 0 & \cdots \\ 0 & -\lambda & \lambda & \cdots \\ 0 & 0 & -\lambda & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

The state transition diagram for the HPP is illustrated in Figure 11.4. □

11.2.2 Chapman–Kolmogorov Equations

By using the Markov property and the law of total probability, we realize that

$$P_{ij}(t+s) = \sum_{k=0}^{r} P_{ik}(t) P_{kj}(s) \qquad \text{for all } i, j \in \mathcal{X}, t, s > 0. \tag{11.11}$$

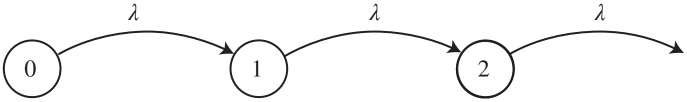


Figure 11.4 State transition diagram for a homogeneous Poisson process (HPP).

Equation (11.11) is known as the *Chapman–Kolmogorov equation*.[2] The equations may, by using (11.2), be written in matrix terms as

$$\mathbb{P}(t+s) = \mathbb{P}(t) \cdot \mathbb{P}(s).$$

11.2.3 Kolmogorov Differential Equations

We will try to establish a set of differential equations that may be used to find $P_{ij}(t)$, and therefore start by considering the Chapman–Kolmogorov equations

$$P_{ij}(t+\Delta t) = \sum_{k=0}^{r} P_{ik}(\Delta t) P_{kj}(t).$$

Observe that we here split the interval $(0, t+\Delta t)$ into two parts. First, we consider a transition from state i to state k in the small interval $(0, \Delta t)$, and thereafter, a transition from state k to state j in the rest of the interval. We now consider

$$P_{ij}(t+\Delta t) - P_{ij}(t) = \sum_{\substack{k=0\\k\neq i}}^{r} P_{ik}(\Delta t) P_{kj}(t) - [1 - P_{ii}(\Delta t)] P_{ij}(t).$$

By dividing by Δt and then taking the limit as $\Delta t \to 0$, we obtain

$$\lim_{\Delta t \to 0} \frac{P_{ij}(t+\Delta t) - P_{ij}(t)}{\Delta t} = \lim_{\Delta t \to 0} \sum_{\substack{k=0\\k\neq i}}^{r} \frac{P_{ik}(\Delta t)}{\Delta t} P_{kj}(t) - \alpha_i P_{ij}(t). \tag{11.12}$$

Because the summing index is finite, we may interchange the limit and summation on the right-hand side of (11.12), and obtain, using (11.7)

$$\dot{P}_{ij}(t) = \sum_{\substack{k=0\\k\neq i}}^{r} a_{ik} P_{kj}(t) - \alpha_i P_{ij}(t) = \sum_{k=0}^{r} a_{ik} P_{kj}(t), \tag{11.13}$$

where $a_{ii} = -\alpha_i$, and the following notation for the time derivative is introduced

$$\dot{P}_{ij}(t) = \frac{d}{dt} P_{ij}(t).$$

The differential equations (11.13) are known as the Kolmogorov equations. They are called backward equations because we start with a transition back by the start of the interval.

The Kolmogorov backward equations may also be written in matrix format as

$$\dot{\mathbb{P}}(t) = \mathbb{A} \cdot \mathbb{P}(t). \tag{11.14}$$

[2] Named after the British mathematician Sydney Chapman (1888–1970) and the Russian mathematician Andrey N. Kolmogorov (1903–1987).

We may also start with the following equation

$$P_{ij}(t+\Delta t)=\sum_{k=0}^{r}P_{ik}(t)P_{kj}(\Delta t).$$

Here, we split the time interval $(0, t+\Delta t)$ into two parts. We consider a transition from i to k in the interval $(0, t)$, and then a transition from k to j in the small interval $(t, t+\Delta t)$. We consider

$$P_{ij}(t+\Delta t)-P_{ij}(t)=\sum_{\substack{k=0\\k\neq i}}^{r}P_{ik}(t)P_{kj}(\Delta t)-[1-P_{jj}(\Delta t)]P_{ij}(t).$$

By dividing by Δt and then taking the limit as $\Delta t \to 0$ we obtain

$$\lim_{\Delta t\to 0}\frac{P_{ij}(t+\Delta t)-P_{ij}(t)}{\Delta t}=\lim_{\Delta t\to 0}\left[\sum_{\substack{k=0\\k\neq i}}^{r}P_{ik}(t)\frac{P_{kj}(\Delta t)}{\Delta t}-\frac{1-P_{jj}(\Delta t)}{\Delta t}P_{ij}(t)\right].$$

Because the summation index is finite, we may interchange limit with summation and obtain

$$\dot{P}_{ij}(t)=\sum_{\substack{k=0\\k\neq i}}^{r}a_{kj}P_{ik}(t)-\alpha_j P_{ij}(t)=\sum_{k=0}^{r}a_{kj}P_{ik}(t), \tag{11.15}$$

where, as before, $a_{jj}=-\alpha_j$. The differential equations (11.15) are known as the Kolmogorov *forward equations*. The interchange of the limit and the sum above does not hold in all cases, but is always valid when the state space is finite.

The Kolmogorov forward equations may be written in matrix terms as

$$\dot{\mathbb{P}}(t)=\mathbb{P}(t)\cdot\mathbb{A}. \tag{11.16}$$

For the Markov processes studied in this book, the backward and the forward equations have the same unique solution $\mathbb{P}(t)$, where $\sum_{j=0}^{r}P_{ij}(t)=1$ for all i in $\mathcal{X}$. In the following, we mainly use the forward equations.

11.2.4 State Equations

Let us assume that we know that the Markov process has state i at time 0, that is, $X(0)=i$. This can be expressed as

$$P_i(0)=\Pr(X(0)=i)=1$$
$$P_k(0)=\Pr(X(0)=k)=0 \qquad \text{for } k\neq i.$$

Because we know the state at time 0, we may simplify the notation by writing $P_{ij}(t)$ as $P_j(t)$. The vector $\boldsymbol{P}(t)=[P_0(t),P_1(t),\dots,P_r(t)]$, then denotes the distribution of the Markov process at time t, when we *know* that the process started in state i at time 0. As in (11.3), we know that $\sum_{j=1}^{r}P_j(t)=1$.

The distribution $\boldsymbol{P}(t)$ may be found from the Kolmogorov forward equations (11.15)

$$\dot{P}_j(t) = \sum_{k=0}^{r} a_{kj} P_k(t), \tag{11.17}$$

where, as before, $a_{jj} = -\alpha_j$. In matrix terms, this may be written

$$[P_0(t), \dots, P_r(t)] \cdot \begin{pmatrix} a_{00} & a_{01} & \cdots & a_{0r} \\ a_{10} & a_{11} & \cdots & a_{1r} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r0} & a_{r1} & \cdots & a_{rr} \end{pmatrix} = [\dot{P}_0(t), \dots, \dot{P}_r(t)] \tag{11.18}$$

or, in a more compact form as

$$\boldsymbol{P}(t) \cdot \mathbb{A} = \dot{\boldsymbol{P}}(t). \tag{11.19}$$

Equation (11.19) is called the *state equation* for the Markov process.

Remark 11.1 (An alternative way of writing the state equations)
Some authors prefer to present the state equations as the transpose of (11.19), that is $\mathbb{A}^T \cdot \boldsymbol{P}(t)^T = \dot{\boldsymbol{P}}(t)^T$. In this case, the vectors are column vectors, and Eq. (11.18) may be written in a slightly more compact form, as

$$\begin{pmatrix} a_{00} & a_{10} & \cdots & a_{r0} \\ a_{01} & a_{11} & \cdots & a_{r1} \\ \vdots & \vdots & \ddots & \vdots \\ a_{0r} & a_{1r} & \cdots & a_{rr} \end{pmatrix} \cdot \begin{bmatrix} P_0(t) \\ P_1(t) \\ \vdots \\ P_r(t) \end{bmatrix} = \begin{bmatrix} \dot{P}_0(t) \\ \dot{P}_1(t) \\ \vdots \\ \dot{P}_r(t) \end{bmatrix}.$$

In this format, the indexes do not follow standard matrix notation. The entries in *column i* represent the departure rates from state i, and the sum of all the entries in a *column* is 0. The reader may choose which format she wants to present the state equations. Both formats give the same result. This book presents the state equations in the format of (11.18) and (11.19). □

Because the sum of the entries in each row in $\mathbb{A}$ is equal to 0, the determinant of $\mathbb{A}$ is 0 and the matrix is singular. Consequently, Eq. (11.19) do not have a unique solution, but by using that

$$\sum_{j=0}^{r} P_j(t) = 1,$$

and the known initial state ($P_i(0) = 1$), we are often able to compute all the probabilities $P_j(t)$ for $j = 0, 1, 2, \dots, r$. (Conditions for existence and uniqueness of the solutions are discussed, for example, by Cox and Miller (1965).)

Example 11.5 (Single component)
Consider a single component. The component has two possible states:

State 1	The component is functioning
State 0	The component is in a failed state.

Transition from state 1 to state 0 means that the component fails, and transition from state 0 to state 1 means that the component is repaired. The transition rate a_{10} is thus the failure rate of the component, and the transition rate a_{01} is the repair rate of the component. In this example, we use the following notation

$a_{10} = \lambda$	The failure rate of the component
$a_{01} = \mu$	The repair rate of the component.

The state transition diagram for the single component is illustrated in Figure 11.5.

The state equations for this simple system is

$$[P_0(t), P_1(t)] \cdot \begin{pmatrix} -\mu & \mu \\ \lambda & -\lambda \end{pmatrix} = [\dot{P}_0(t), \dot{P}_1(t)]. \tag{11.20}$$

The component is assumed to be functioning at time $t = 0$,

$$P_1(0) = 1, \quad P_0(0) = 0.$$

Because the two equations we get from (11.20) are linearly dependent, we use only one of them, for example

$$-\mu P_0(t) + \lambda P_1(t) = \dot{P}(t).$$

And combine this equation with $P_0(t) + P_1(t) = 1$. The solution is:

$$P_1(t) = \frac{\mu}{\mu + \lambda} + \frac{\lambda}{\mu + \lambda}\, e^{-(\lambda+\mu)t} \tag{11.21}$$

$$P_0(t) = \frac{\lambda}{\mu + \lambda} - \frac{\lambda}{\mu + \lambda}\, e^{-(\lambda+\mu)t}. \tag{11.22}$$

For a detailed solution of the differential equation, see Ross (1996). $P_1(t)$ denotes the probability that the component is functioning at time t, that is, the *availability* of the component. The limiting availability $P_1 = \lim_{t\to\infty} P_1(t)$ is from (11.21),

$$P_1 = \lim_{t\to\infty} P_1(t) = \frac{\mu}{\lambda + \mu}. \tag{11.23}$$

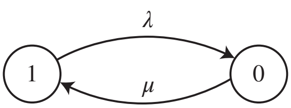


Figure 11.5 State transition diagram for a single component (function-repair cycle).

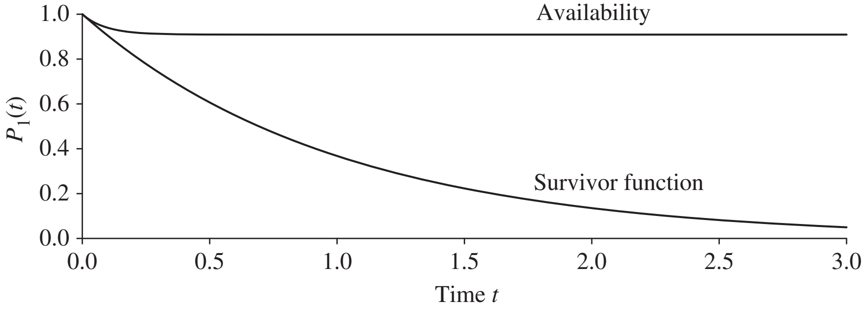


Figure 11.6 Availability and survivor function for a single component ($\lambda = 1$, $\mu = 10$).

The mean time-to-failure, MTTF, is equal to $1/\lambda$, and the mean time to repair, MTTR, is equal to $1/\mu$. The limiting availability may therefore be written as the well-known formula

$$P_1 = \frac{\text{MTTF}}{\text{MTTF+MTTR}}. \tag{11.24}$$

When there is no repair ($\mu = 0$), the availability is $P_1(t) = e^{-\lambda t}$ which coincides with the survivor function of the component. The availability $P_1(t)$ is illustrated in Figure 11.6. □

11.3 Asymptotic Solution

In many applications, only the long-run (steady state) probabilities are of interest, that is, the values of $P_j(t)$ when $t \to \infty$. In Example 11.5, the state probabilities $P_j(t)$ ($j = 0, 1$) approached a steady state P_j when $t \to \infty$. The same steady state value would have been found irrespective of whether the system started in the operating state or in the failed state.

Convergence toward steady state probabilities is assumed for the Markov processes studied in this chapter. The process is said to be *irreducible* if every state is reachable from every other state (see Ross 1996). For an irreducible Markov process, it can be shown that the limits

$$\lim_{t\to\infty} P_j(t) = P_j \qquad \text{for } j = 0, 1, 2, \ldots, r,$$

always exist and are independent of the initial state of the process (at time $t = 0$). For a proof, see Ross (1996). Hence, a process that has been running for a long time, has lost its dependency of its initial state $X(0)$. The process converges to a process where the probability of being in state j is

$$P_j = P_j(\infty) = \lim_{t\to\infty} P_j(t) \qquad \text{for } j = 0, 1, \ldots, r.$$

These asymptotic probabilities are often called the *steady state probabilities* for the Markov process.

If $P_j(t)$ tends to a constant value when $t \to \infty$, then

$$\lim_{t\to\infty} \dot{P}_j(t) = 0 \qquad \text{for } j = 0, 1, \dots, r.$$

The steady state probabilities $\boldsymbol{P} = [P_0, P_1, \dots, P_r]$ must therefore satisfy the matrix equation:

$$[P_0, P_1, \dots, P_r] \cdot \begin{pmatrix} a_{00} & a_{01} & \cdots & a_{0r} \\ a_{10} & a_{11} & \cdots & a_{1r} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r0} & a_{r1} & \cdots & a_{rr} \end{pmatrix} = [0, 0, \dots, 0], \tag{11.25}$$

which may be abbreviated to

$$\boldsymbol{P} \cdot \mathbb{A} = \boldsymbol{0}, \tag{11.26}$$

whereas before

$$\sum_{j=0}^{r} P_j = 1.$$

To calculate the steady state probabilities, $P_0, P_1, \dots, P_r$ of such a process, we use r of the $r + 1$ linear algebraic equation from the matrix equation (11.25), and in addition the fact that the sum of the state probabilities always is equal to 1. The initial state of the process has no influence on the steady state probabilities. Observe that P_j also may be interpreted as the average, long-run proportion of time the system spends in state j.

Remark 11.2 (Numerical solution with R)
Equation (11.25) is seen to be a set of *linear equations* on a matrix format. When you have numerical values for the transition rates, you may use the basic command `solve` or the `matlib` package in R to solve (11.25) and find $[P_0, P_1, \dots, P_r]$.[3,4] □

Example 11.6 (Power station with two generators)
Consider a power station with two generators, 1 and 2. Each generator can have two states: a functioning state (1) and a failed state (0). A generator is considered to

[3] The `matlib` package is available on https://cran.r-project.org/web/packages/matlib/index.html, where brief user guides are found under the heading "Vignettes."

[4] Several other computer programs can be used for the same purpose. Among these are Python, Octave, and MATLAB®. Observe that R can read and write MATLAB (and Octave) m-files by using the package `R.matlab`.

be in the failed state (0) also during repair. Generator 1 is supplying 100 MW when it is functioning, and 0 MW when it is not functioning. Generator 2 is supplying 50 MW when it is functioning, and 0 MW when it is not functioning.

The possible states of the system are

System state	State of generator 1	State of generator 2	System output (MW)
3	1	1	150
2	1	0	100
1	0	1	50
0	0	0	0

Assume that the generators fail independent of each other, and that they are operated on a continuous basis. The failure rates of the generators are

λ_1	Failure rate of generator 1
λ_2	Failure rate of generator 2

When a generator fails, a repair action is started to bring the generator back into operation. The two generators are assumed to be repaired independent of each other, by two independent repair crews. The repair rates of the generators are

μ_1	Repair rate for generator 1
μ_2	Repair rate for generator 2

The corresponding state transition diagram is shown in Figure 11.7.

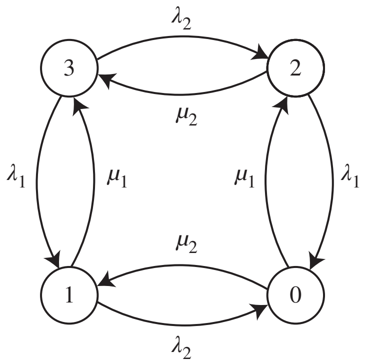


Figure 11.7 State transition diagram of the generators in Example 11.6.

The transition matrix is

$$\mathbb{A} = \begin{pmatrix} -(\mu_1+\mu_2) & \mu_2 & \mu_1 & 0 \\ \lambda_2 & -(\lambda_2+\mu_1) & 0 & \mu_1 \\ \lambda_1 & 0 & -(\lambda_1+\mu_2) & \mu_2 \\ 0 & \lambda_1 & \lambda_2 & -(\lambda_1+\lambda_2) \end{pmatrix}.$$

We can use (11.26) to find the steady state probabilities P_j for $j = 0, 1, 2, 3$, and get the following equations:

$$\begin{aligned} -(\mu_1+\mu_2)P_0 + \lambda_2 P_1 + \lambda_1 P_2 &= 0 \\ \mu_2 P_0 - (\lambda_2+\mu_1)P_1 + \lambda_1 P_3 &= 0 \\ \mu_1 P_0 - (\lambda_1+\mu_2)P_2 + \lambda_2 P_3 &= 0 \\ P_0 + P_1 + P_2 + P_3 &= 1. \end{aligned}$$

Observe that we use three of the steady state equations from (11.26) and in addition the fact that $P_0 + P_1 + P_2 + P_3 = 1$. Observe also that we may choose any three of the four steady state equations, and get the same solution.

The solution is

$$\begin{aligned} P_0 &= \frac{\lambda_1\lambda_2}{(\lambda_1+\mu_1)(\lambda_2+\mu_2)} \\ P_1 &= \frac{\lambda_1\mu_2}{(\lambda_1+\mu_1)(\lambda_2+\mu_2)} \\ P_2 &= \frac{\mu_1\lambda_2}{(\lambda_1+\mu_1)(\lambda_2+\mu_2)} \\ P_3 &= \frac{\mu_1\mu_2}{(\lambda_1+\mu_1)(\lambda_2+\mu_2)}. \end{aligned} \tag{11.27}$$

Now for $i = 1, 2$ let:

$$\begin{aligned} q_i &= \frac{\lambda_i}{\lambda_i+\mu_i} = \frac{\text{MTTR}_i}{\text{MTTF}_i + \text{MTTR}_i} \\ p_i &= \frac{\mu_i}{\lambda_i+\mu_i} = \frac{\text{MTTF}_i}{\text{MTTF}_i + \text{MTTR}_i}, \end{aligned}$$

where $\text{MTTR}_i = 1/\mu_i$ is the mean time to repair of component i, and $\text{MTTF}_i = 1/\lambda_i$ is the mean time-to-failure of component i ($i = 1, 2$). q_i thus denotes the average, or limiting, unavailability of component i, whereas p_i denotes the average (limiting) availability of component i, ($i = 1, 2$). The steady state probabilities may thus be written as

$$\begin{aligned}P_0 &= q_1 q_2\\ P_1 &= q_1 p_2\\ P_2 &= p_1 q_2\\ P_3 &= p_1 p_2.\end{aligned} \tag{11.28}$$

In this example, where the components fail and are repaired independently of each other, we may use direct reasoning to obtain the results in (11.28):

$$\begin{aligned}P_0 =&\ \text{Pr(Component 1 is failed) Pr(Component 2 is failed)} &= q_1 q_2\\ P_1 =&\ \text{Pr(Component 1 is failed) Pr(Component 2 is functioning)} &= q_1 p_2\\ P_2 =&\ \text{Pr(Component 1 is functioning) Pr(Component 2 is failed)} &= p_1 q_2\\ P_3 =&\ \text{Pr(Component 1 is functioning) Pr(Component 2 is functioning)} &= p_1 p_2\end{aligned}.$$

Because all failures and repairs are independent events, we do not need to use Markov methods to find the steady state probabilities. The steady state probabilities may easily be found by using standard probability rules for independent events. Please observe that this only applies for systems with independent failures and repairs.

Assume now that we have the following data:

	Generator 1	**Generator 2**
MTTF$_i$	6 mo ≈ 4380 h	8mo ≈ 5840 h
Failure rate per hour, λ_i	2.3×10^{-4}	1.7×10^{-4}
MTTR$_i$	12 h	24 h
Repair rate per hour, μ_i	8.3×10^{-2}	4.2×10^{-2}

Observe that the steady state probabilities can be interpreted as the mean proportion of time the system stays in the state concerned. The steady state probability of state 1 is, for example, equal to

$$P_1 = \frac{\lambda_1 \mu_2}{(\lambda_1 + \mu_1)(\lambda_2 + \mu_2)} = q_1 p_2 \approx 2.72 \times 10^{-3}.$$

Hence,

$$P_1 = 0.002\ 72 \left[\frac{\text{yr}}{\text{yr}}\right] = 0.002\ 72 \cdot 8760 \left[\frac{\text{h}}{\text{yr}}\right] \approx 23.8 \left[\frac{\text{h}}{\text{yr}}\right].$$

In the long run, the system stays in state 1 approximately 23.8 h/yr. This does *not* mean that state 1 occurs on average once per year and lasts for 23.8 hours each time.

With the given data, we obtain:

System state	System output (MW)	Steady state probability	Average hours in state per year
3	150	0.9932	8700.3
2	100	4.08×10^{-3}	35.8
1	50	2.72×10^{-3}	23.8
0	0	1.12×10^{-5}	0.1

□

11.3.1 System Performance Metrics

Several system performance metrics for the *steady state* situation are introduced in this section. Examples are provided in Sections 11.4–11.6.

Visit Frequency

The Kolmogorov forward equation (11.15) is

$$\dot{P}_{ij}(t) = \sum_{\substack{k=0 \\ k\neq i}}^{r} a_{kj} P_{ik}(t) - \alpha_j P_{ij}(t).$$

When we let $t \to \infty$, then $P_{ij}(t) \to P_j$, and $\dot{P}_{ij}(t) \to 0$. Because the summation index in (11.15) is finite, we may interchange the limit and the sum and get, as $t \to \infty$,

$$0 = \sum_{\substack{k=0 \\ k\neq i}}^{r} a_{kj} P_k - \alpha_j P_j,$$

which can be written as

$$P_j \ \alpha_j = \sum_{\substack{k=0 \\ k\neq i}}^{r} P_k a_{kj}. \tag{11.29}$$

The (unconditional) probability of a departure from state j in the time interval $(t, t + \Delta t]$ is

$$\sum_{\substack{k=0 \\ k\neq i}}^{r} \Pr[(X(t+\Delta t) = k) \ \cap \ (X(t) = j)]$$

$$= \sum_{\substack{k=0 \\ k\neq i}}^{r} \Pr(X(t+\Delta t) = k \mid X(t) = j) \Pr(X(t) = j) = \sum_{\substack{k=0 \\ k\neq i}}^{r} P_{jk}(\Delta t) P_j(t).$$

When $t \to \infty$, this probability tends to $\sum_{\substack{k=0\\k\neq i}}^{r} P_{jk}(\Delta t)P_j$, and the steady state frequency of departures from state j is, with the same argument as we used to derive Eq. (11.5),

$$\nu_j^{\text{dep}} = \lim_{\Delta t\to 0} \frac{\sum_{\substack{k=0\\k\neq i}}^{r} P_{jk}(\Delta t)P_j}{\Delta t} = P_j\ \alpha_j.$$

The left-hand side of (11.29) is hence the steady state frequency of departures from state j. The frequency of departures from state j is seen to be the proportion of time P_j spent in state j, times the transition rate α_j out of state j.

Similarly, the frequency of transitions from state k into state j is $P_k\ a_{kj}$. The total frequency of arrivals into state j is therefore

$$\nu_j^{\text{arr}} = \sum_{\substack{k=0\\k\neq i}}^{r} P_k a_{kj}.$$

Equation (11.29) says that the frequency of departures from state j is equal to the frequency of arrivals into state j, for $j = 0, 1, \dots, r$, and is therefore sometimes referred to as the *balance equations*. In the steady state situation, we define the *visit frequency* to state j as

$$\nu_j = P_j\ \alpha_j = \sum_{\substack{k=0\\k\neq i}}^{r} P_k a_{kj}, \tag{11.30}$$

and the mean time between visits to state j is $1/\nu_j$.

Mean Duration of a Visit

When the process enters a state j, the system stays in this state a time $\widetilde{T}_j$ until the process departures from that state, $j = 0, 1, \dots, r$. We have called $\widetilde{T}_j$ the sojourn time in state j, and shown that $\widetilde{T}_j$ is exponentially distributed with rate α_j. The mean sojourn time, or mean duration of a visit, is hence

$$\theta_j = E(\widetilde{T}_j) = \frac{1}{\alpha_j} \qquad \text{for } j = 0, 1, \dots, r. \tag{11.31}$$

By combining (11.30) and (11.31), we obtain

$$\nu_j = P_j\ \alpha_j = \frac{P_j}{\theta_j}$$

$$P_j = \nu_j \theta_j. \tag{11.32}$$

The mean proportion of time, P_j, the system is spending in state j is thus equal to the visit frequency to state j multiplied by the mean duration of a visit in state j for $j = 0, 1, \dots, r$.

System Availability

Let $\mathcal{X} = \{0, 1, \dots, r\}$ be the set of all possible states of a system. Some of these states represent system functioning according to some specified criteria. Let B denote the subset of states in which the system is functioning, and let $F = \mathcal{X} - B$ denote the states in which the system is failed.

The average, or long-term *availability* of the system is the mean proportion of time when the system is functioning; that is, its state is a member of B. The average system availability A_S is thus defined as follows:

$$A_S = \sum_{j \in B} P_j. \tag{11.33}$$

In the following, we omit the term "average" and call A_S the system availability. The system unavailability $(1 - A_S)$ is then

$$1 - A_S = \sum_{j \in F} P_j. \tag{11.34}$$

The unavailability $(1 - A_S)$ of the system is the mean proportion of time when the system is in a failed state.

Frequency of System Failures

The frequency ω_F of system failures is the steady state frequency of transitions from a functioning state (in B) to a failed state (in F),

$$\omega_F = \sum_{j \in B} \sum_{k \in F} P_j a_{jk}. \tag{11.35}$$

Mean Duration of a System Failure

The mean duration θ_F of a system failure is defined as the mean time from the system enters into a failed state (F) until it is repaired/restored and brought back into a functioning state (B).

Analogous with (11.32) it is obvious that the system unavailability $(1 - A_S)$ is equal to the frequency of system failures multiplied by the mean duration of a system failure. Hence,

$$1 - A_S = \omega_F \theta_F. \tag{11.36}$$

Mean Time Between System Failures

The mean time between system failures, MTBF_S is the mean time between consecutive transitions from a functioning state (B) into a failed state (F). The MTBF_S may be computed from the frequency of system failures by

$$\text{MTBF}_S = \frac{1}{\omega_F}. \tag{11.37}$$

Mean Functioning Time Until System Failure

The mean functioning time ("up-time") until system failure, $E(U)_s$, is the mean time from a transition from a failed state (F) into a functioning state (B) until the first transition back to a failed state (F). It is obvious that

$$\text{MTBF}_S = E(U)_s + \theta_F. \tag{11.38}$$

Observe the difference between the mean functioning time ("up-time") and the mean time to system failure MTTF_S. The MTTF_S is normally calculated as the mean time until system failure when the system initially is in a *specified* functioning state.

11.4 Parallel and Series Structures

This section studies the steady state properties of parallel and series structures of independent components.

11.4.1 Parallel Structures of Independent Components

Reconsider the parallel structure of two independent components in Example 11.6. For this structure, we get

Mean Duration of the Visits

From (11.31), we get

$$\begin{aligned}
\theta_0 &= 1/(\mu_1 + \mu_2) \\
\theta_1 &= 1/(\lambda_2 + \mu_1) \\
\theta_2 &= 1/(\lambda_1 + \mu_2) \\
\theta_3 &= 1/(\lambda_1 + \lambda_2).
\end{aligned} \tag{11.39}$$

Visit Frequency

From (11.31) and (11.39), we get

$$\begin{aligned}
\nu_0 &= P_0(\mu_1 + \mu_2) \\
\nu_1 &= P_1(\lambda_2 + \mu_1) \\
\nu_2 &= P_2(\lambda_1 + \mu_2) \\
\nu_3 &= P_3(\lambda_1 + \lambda_2).
\end{aligned} \tag{11.40}$$

The parallel structure is functioning when at least one of its two components is functioning. When the system is in state 1, 2, or 3 the system is functioning, whereas state 0 corresponds to system failure.

The average system unavailability is

$$1 - A_S = P_0 = q_1 q_2, \tag{11.41}$$

and the average system availability is

$$A_S = P_1 + P_2 + P_3 = 1 - q_1 q_2.$$

The frequency of system failures ω_F is equal to the visit frequency to state 0, which is

$$\omega_F = \nu_0 = P_0(\mu_1 + \mu_2) = (1 - A_S)(\mu_1 + \mu_2). \tag{11.42}$$

The mean duration of a system failure θ_F is in this case equal to the mean duration of a stay in state 0. Thus,

$$\theta_F = \theta_0 = \frac{1}{\mu_1 + \mu_2} = \frac{1 - A_S}{\omega_F}. \tag{11.43}$$

For a parallel structure of n independent components, the above results may be generalized as follows: For system unavailability,

$$1 - A_S = \prod_{i=1}^{n} q_i = \prod_{i=1}^{n} \frac{\lambda_i}{\lambda_i + \mu_i}. \tag{11.44}$$

For frequency of system failures,

$$\omega_F = (1 - A_S) \sum_{i=1}^{n} \mu_i. \tag{11.45}$$

For mean duration of a system failure,

$$\theta_F = \frac{1}{\sum_{i=1}^{n} \mu_i}. \tag{11.46}$$

The mean functioning time (up-time) $E(U)_P$ of the parallel structure can be determined from

$$1 - A_S = \frac{\theta_F}{\theta_F + E(U)_P}.$$

Hence,

$$E(U)_P = \frac{\theta_F A_S}{1 - A_S} = \frac{1 - \prod_{i=1}^{n} \lambda_i/(\lambda_i + \mu_i)}{\prod_{i=1}^{n} \lambda_i/(\lambda_i + \mu_i) \sum_{j=1}^{n} \mu_j}. \tag{11.47}$$

When the component availabilities are very high: (i.e. $\lambda_i \ll \mu_i$ for all $i = 1, 2, \ldots, n$), then

$$\frac{\lambda_i}{\lambda_i + \mu_i} = \frac{\lambda_i \text{ MTTR}_i}{1 + \lambda_i \text{ MTTR}_i} \approx \lambda_i \text{ MTTR}_i.$$

The frequency ω_F of system failures can now be approximated as

$$\omega_F = (1 - A_S) \sum_{i=1}^{n} \mu_i = \prod_{i=1}^{n} \frac{\lambda_i}{\lambda_i + \mu_i} \sum_{j=1}^{n} \mu_j$$
$$\approx \prod_{i=1}^{n} \lambda_i \text{ MTTR}_i \sum_{j=1}^{n} \frac{1}{\text{MTTR}_j}. \tag{11.48}$$

For two components, (11.48) reduces to

$$\omega_F \approx \lambda_1 \lambda_2 (\text{MTTR}_1 + \text{MTTR}_2). \tag{11.49}$$

For three components, (11.48) reduces to

$$\omega_F \approx \lambda_1 \lambda_2 \lambda_3 (\text{MTTR}_1 \text{MTTR}_2 + \text{MTTR}_1 \text{MTTR}_3 + \text{MTTR}_2 \text{MTTR}_3).$$

11.4.2 Series Structures of Independent Components

Consider a series structure of two independent components. The states of the system and the transition rates are as defined in Example 11.6. The state transition diagram of the series structure is shown in Figure 11.8. The corresponding steady state equations are equal to those found for the parallel structure in Example 11.6.

The average availability of the structure, A_S, is equal to P_3 which was found in (11.28) to be

$$A_S = P_3 = \frac{\mu_1 \mu_2}{(\lambda_1 + \mu_1)(\lambda_2 + \mu_2)} = p_1 p_2, \tag{11.50}$$

where

$$p_i = \frac{\mu_i}{\lambda_i + \mu_i} \qquad \text{for } i = 1, 2.$$

The frequency of system failures, ω_F, is the same as the frequency of visits to state 3. Thus,

$$\omega_F = \nu_3 = P_3(\lambda_1 + \lambda_2) = A_S(\lambda_1 + \lambda_2). \tag{11.51}$$

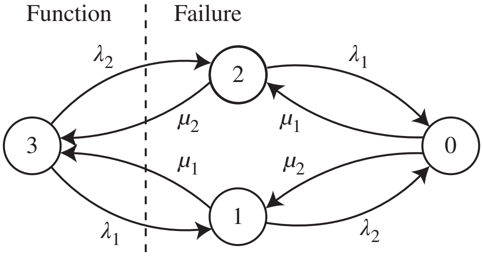


Figure 11.8 Partitioning the state transition diagram of a series structure of two independent components.

The mean duration of a system failure θ_F is equal to

$$\theta_F = \frac{1 - A_S}{\omega_F}. \tag{11.52}$$

For a series structure of n independent components, the above results can be generalized as follows: For system availability

$$A_S = \prod_{i=1}^{n} p_i = \prod_{i=1}^{n} \frac{\mu_i}{\lambda_i + \mu_i}. \tag{11.53}$$

For frequency of system failures

$$\omega_F = A_S \sum_{i=1}^{n} \lambda_i. \tag{11.54}$$

For mean duration of a system failure

$$\begin{aligned}\theta_F &= \frac{1 - A_S}{\omega_F} \\ &= \frac{1}{\sum_{i=1}^{n} \lambda_i} \frac{1 - A_S}{A_S} \\ &= \frac{1 - \prod_{i=1}^{n} \mu_i/(\lambda_i + \mu_i)}{\prod_{i=1}^{n} \mu_i/(\lambda_i + \mu_i) \sum_{j=1}^{n} \lambda_j}.\end{aligned} \tag{11.55}$$

When all the component availabilities are very high such that $\lambda_i \ll \mu_i$ for all i, then $A_S \approx 1$ and the frequency of system failures is approximately

$$\omega_F \approx \sum_{i=1}^{n} \lambda_i, \tag{11.56}$$

which is the same as the failure rate of a nonrepairable series structure of n independent components.

The mean duration of a system failure θ_F may be approximated as

$$\begin{aligned}\theta_F &= \frac{1}{\sum_{i=1}^{n} \lambda_i} \frac{1 - A_S}{A_S} = \frac{1}{\sum_{i=1}^{n} \lambda_i} \left(\frac{1}{A_S} - 1 \right) = \frac{1}{\sum_{i=1}^{n} \lambda_i} \left(\prod_{i=1}^{n} \frac{1}{p_i} - 1 \right) \\ &= \frac{1}{\sum_{i=1}^{n} \lambda_i} \left(\prod_{i=1}^{n} \left(1 + \frac{\lambda_i}{\mu_i} \right) - 1 \right) \approx \frac{1}{\sum_{i=1}^{n} \lambda_i} \left(1 + \sum_{i=1}^{n} \frac{\lambda_i}{\mu_i} - 1 \right) \\ &= \frac{\sum_{i=1}^{n} \lambda_i/\mu_i}{\sum_{i=1}^{n} \lambda_i} = \frac{\sum_{i=1}^{n} \lambda_i \text{MTTR}_i}{\sum_{i=1}^{n} \lambda_i},\end{aligned} \tag{11.57}$$

where $\text{MTTR}_i = 1/\mu_i$ as before is the mean time to repair component i, $i = 1, 2, \ldots, n$. Equation (11.57) is a commonly used approximation for the mean duration of a failure in series structures of high reliability.

11.4.3 Series Structure of Components Where Failure of One Component Prevents Failure of the Other

Consider a series structure of two components. When one of the components fails, the other component is immediately taken out of operation until the failed component is repaired.[5] After a component is taken out of operation, it is not exposed to any stress, and we therefore assume that it will not fail. This dependence between the failures prevents a simple solution by direct reasoning as was possible in Example 11.6. This system has three possible states as described in Table 11.2.

The following transition rates are assumed:

$a_{21} = \lambda_1$	Failure rate of component 1
$a_{20} = \lambda_2$	Failure rate of component 2
$a_{12} = \mu_1$	Repair rate of component 1
$a_{02} = \mu_2$	Repair rate of component 2

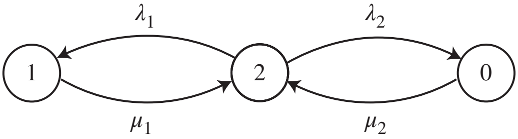
The state transition diagram of the series structure is illustrated in Figure 11.9. The steady state equations for this system are

$$[P_0, P_1, P_2] \cdot \begin{pmatrix} -\mu_2 & 0 & \mu_2 \\ 0 & -\mu_1 & \mu_1 \\ \lambda_2 & \lambda_1 & -(\lambda_1 + \lambda_2) \end{pmatrix} = [0, 0, 0]. \tag{11.58}$$

Table 11.2 Possible states of a series structure of two components where failure of one component prevents failure of the other.

State	Component 1	Component 2
2	Functioning	Functioning
1	Taken out of operation	Functioning
0	Functioning	Taken out of operation

Figure 11.9 State transition diagram of a series structure of two components. Where failure of one component prevents failure of the other component.



[5] The same model is discussed by Barlow and Proschan (1975, pp. 194–201) in a more general context that does not assume constant failure and repair rates.

The steady state probabilities may be found from the equations

$$\begin{aligned} -\mu_2 P_0 + \lambda_2 P_2 &= 0 \\ -\mu_1 P_1 + \lambda_1 P_2 &= 0 \\ P_0 + P_1 + P_2 &= 1. \end{aligned}$$

The solution is

$$P_2 = \frac{\mu_1 \mu_2}{\lambda_1 \mu_2 + \lambda_2 \mu_1 + \mu_1 \mu_2} = \frac{1}{1 + (\lambda_1/\mu_1) + (\lambda_2/\mu_2)}. \tag{11.59}$$

$$P_1 = \frac{\lambda_1}{\mu_1} P_2. \tag{11.60}$$

$$P_0 = \frac{\lambda_2}{\mu_2} P_2. \tag{11.61}$$

Because the series structure is only functioning when both the components are functioning (state 2), the average system availability is,

$$A_S = P_2 = \frac{\mu_1 \mu_2}{\lambda_1 \mu_2 + \lambda_2 \mu_1 + \mu_1 \mu_2} = \frac{1}{1 + (\lambda_1/\mu_1) + (\lambda_2/\mu_2)}.$$

Observe that in this case, the availability of the series structure is *not* equal to the product of the component availabilities.

The mean durations of the stays in each state are

$$\begin{aligned} \theta_2 &= \frac{1}{\lambda_1 + \lambda_2} \\ \theta_1 &= \frac{1}{\mu_1} \\ \theta_0 &= \frac{1}{\mu_2}. \end{aligned}$$

The frequency of system failures ω_F is the same as the frequency of visits to state 2.

$$\omega_F = \nu_2 = P_2(\lambda_1 + \lambda_2) = A_S(\lambda_1 + \lambda_2). \tag{11.62}$$

The mean duration of a system failure θ_F is

$$\begin{aligned} \theta_F &= \frac{1 - A_S}{\omega_F} = \frac{1}{\lambda_1 + \lambda_2} \frac{1 - A_S}{A_S} \\ &= \frac{1}{\mu_1} \frac{\lambda_1}{\lambda_1 + \lambda_2} + \frac{1}{\mu_2} \frac{\lambda_2}{\lambda_1 + \lambda_2}. \end{aligned} \tag{11.63}$$

Equation(11.63) may also be written

$$\begin{aligned} \theta_F = {} & \text{MTTR}_1 \text{ Pr(Component 1 fails|system failure)} \\ & + \text{MTTR}_2 \text{ Pr(Component 2 fails|system failure)}. \end{aligned}$$

This formula is obvious because the duration of a system failure is equal to the repair time of component 1 when component 1 fails and equal to the repair time of component 2 when component 2 fails.

The mean time between system failures, MTBF_S, is

$$\begin{aligned}\text{MTBF}_S = \text{MTTF}_S + \theta_F &= \frac{1}{\lambda_1+\lambda_2} + \frac{1}{\mu_1}\frac{\lambda_1}{\lambda_1+\lambda_2} + \frac{1}{\mu_2}\frac{\lambda_2}{\lambda_1+\lambda_2} \\ &= \frac{1+(\lambda_1/\mu_1)+(\lambda_2/\mu_2)}{\lambda_1+\lambda_2}.\end{aligned} \tag{11.64}$$

The frequency of system failures may also be expressed as

$$\omega_F = \frac{1}{\text{MTBF}_S} = (\lambda_1+\lambda_2)\frac{1}{1+(\lambda_1/\mu_1)+(\lambda_2/\mu_2)} = A_S(\lambda_1+\lambda_2).$$

For a series structure of n components, the above results can be generalized as follows: For system availability

$$A_S = \frac{1}{1+\sum_{i=1}^n(\lambda_i/\mu_i)}. \tag{11.65}$$

For mean time to system failure

$$\text{MTTF} = \frac{1}{\sum_{i=1}^n \lambda_i}. \tag{11.66}$$

For mean duration of a system failure

$$\theta_F = \sum_{i=1}^n \frac{1}{\mu_i}\frac{\lambda_i}{\sum_{j=1}^n \lambda_j} = \frac{1}{\sum_{j=1}^n \lambda_j}\sum_{i=1}^n \frac{\lambda_i}{\mu_i}. \tag{11.67}$$

For frequency of system failures

$$\omega_F = A_S\sum_{i=1}^n \lambda_i = \frac{\sum_{i=1}^n \lambda_i}{1+\sum_{i=1}^n(\lambda_i/\mu_i)}. \tag{11.68}$$

11.5 Mean Time to First System Failure

Before developing formulas for the mean time to system failure, we need to introduce the concept of absorbing states.

11.5.1 Absorbing States

All the processes we have studied so far in this chapter have been irreducible, which means that every state is reachable from every other state.

We now introduce Markov processes with *absorbing states*. An absorbing state is a state that, once entered, cannot be left until the system starts a new mission. The popular saying is that the system is *trapped* in an absorbing state.

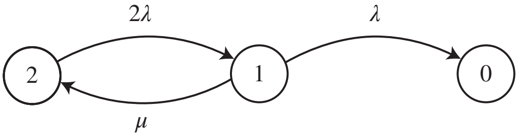


Figure 11.10 State transition diagram for a parallel structure of two identical components.

Example 11.7 (Parallel structure of two independent components) Reconsider the parallel system in Example 11.3 with two independent and identical components with failure rate λ. When one of the components fails, it is repaired. The repair time is assumed to be exponentially distributed with repair rate μ. When both components have failed, the system is considered to have failed and no recovery is possible. Let the number of functioning components denote the state of the system. The state space is thus $\mathcal{X} = \{0, 1, 2\}$, and state 0 is an absorbing state. The state transition diagram of the system is given in Figure 11.10.

Assume that both components are functioning (state 2) at time 0. That is $P_2(0) = 1$. The transition rate matrix of this structure is thus

$$\mathbb{A} = \begin{pmatrix} 0 & 0 & 0 \\ \lambda & -(\lambda + \mu) & \mu \\ 0 & 2\lambda & -2\lambda \end{pmatrix}. \tag{11.69}$$

Because state 0 is an absorbing state, all the transition rates from this state are equal to zero. Thus, the entries of the row corresponding to the absorbing state are all equal to zero.

Because the matrix $\mathbb{A}$ does not have full rank, we may remove one of the three equations without losing any information about $P_0(t)$, $P_1(t)$, and $P_2(t)$. In this case, we remove the first of the three equations. This is accomplished by removing the first column of the matrix. Hence, we get the state equations

$$[P_0(t), P_1(t), P_2(t)] \cdot \begin{pmatrix} 0 & 0 \\ -(\lambda + \mu) & \mu \\ 2\lambda & -2\lambda \end{pmatrix} = [\dot{P}_1(t), \dot{P}_2(t)].$$

Because all the elements of the first row of the matrix are equal to zero, $P_0(t)$ "disappears" in the solution of the equations. The matrix equations may therefore be reduced to

$$[P_1(t), P_2(t)] \cdot \begin{pmatrix} -(\lambda + \mu) & \mu \\ 2\lambda & -2\lambda \end{pmatrix} = [\dot{P}_1(t), \dot{P}_2(t)]. \tag{11.70}$$

The matrix

$$\begin{pmatrix} -(\lambda + \mu) & \mu \\ 2\lambda & -2\lambda \end{pmatrix}$$

has full rank if $\lambda > 0$. Therefore, (11.70) determines $P_1(t)$ and $P_2(t)$. $P_0(t)$ may thereafter be found from $P_0(t) = 1 - P_1(t) - P_2(t)$. This solution of the reduced matrix equations (11.70) is identical to the solution of the initial matrix equations. The reduced matrix is seen to be obtained by deleting the row and the column corresponding to the absorbing state.

Because state 0 is absorbing and reachable from the other states, it is obvious that

$$\lim_{t\to\infty} P_0(t) = 1.$$

The Laplace transforms of the reduced matrix equations (11.70) are

$$(P_1^*(s), P_2^*(s)) \cdot \begin{bmatrix} -(\lambda + \mu) & \mu \\ 2\lambda & -2\lambda \end{bmatrix} = (sP_1^*(s), sP_2^*(s) - 1),$$

when the system is assumed to be in state 2 at time $t = 0$. Thus,

$$-(\lambda + \mu)P_1^*(s) + 2\lambda P_2^*(s) = sP_1^*(s)$$

$$\mu P_1^*(s) - 2\lambda P_2^*(s) = sP_2^*(s) - 1.$$

Solving for $P_1^*(s)$ and $P_2^*(s)$, we get (see Appendix B)

$$P_1^*(s) = \frac{2\lambda}{s^2 + (3\lambda + \mu)s + 2\lambda^2}$$

$$P_2^*(s) = \frac{\lambda + \mu + s}{s^2 + (3\lambda + \mu)s + 2\lambda^2}.$$

Let $R(t)$ denote the survivor function of the system. Because the system is functioning as long as the system is either in state 2 or in state 1, the survivor function is equal to

$$R(t) = P_1(t) + P_2(t) = 1 - P_0(t).$$

The Laplace transform of $R(t)$ is thus

$$R^*(s) = P_1^*(s) + P_2^*(s) = \frac{3\lambda + \mu + s}{s^2 + (3\lambda + \mu)s + 2\lambda^2}. \tag{11.71}$$

The survivor function $R(t)$ may now be determined by inverting the Laplace transform, or we may consider $P_0(t) = 1 - R(t)$ which denotes the distribution function of the time T_s to system failure. The Laplace transform of $P_0(t)$ is

$$P_0^*(s) = \frac{1}{s} - P_1^*(s) - P_2^*(s) = \frac{2\lambda^2}{s[s^2 + (3\lambda + \mu)s + 2\lambda^2]}.$$

Let $f_s(t)$ denote the probability density function of the time T_s to system failure, that is, $f_s(t) = dP_0(t)/dt$. The Laplace transform of $f_s(t)$ is thus

$$f_s^*(s) = sP_0^*(s) - P_0(0) = \frac{2\lambda^2}{s^2 + (3\lambda + \mu)s + 2\lambda^2}. \tag{11.72}$$

The denominator of (11.71) can be written

$$s^2 + (3\lambda + \mu)s + 2\lambda^2 = (s - k_1)(s - k_2),$$

where

$$k_1 = \frac{-(3\lambda + \mu) + \sqrt{\lambda^2 + 6\lambda\mu + \mu^2}}{2}$$

$$k_2 = \frac{-(3\lambda + \mu) - \sqrt{\lambda^2 + 6\lambda\mu + \mu^2}}{2}.$$

The expression for $f_s^*(s)$ can be rearranged so that

$$f_s^*(s) = \frac{2\lambda^2}{k_1 - k_2}\left(\frac{1}{s + k_2} - \frac{1}{s + k_1}\right).$$

By inverting this transform, we get

$$f_s(t) = \frac{2\lambda^2}{k_1 - k_2}(e^{-k_2 t} - e^{-k_1 t}).$$

The mean time to system failure MTTF_S is now given by (the integration is left to the reader as an exercise):

$$\text{MTTF}_S = \int_0^\infty t f_s(t)\, dt = \frac{3}{2\lambda} + \frac{\mu}{2\lambda^2}. \tag{11.73}$$

Observe that the MTTF_S of a two-component parallel system, without any repair (i.e. $\mu = 0$), is equal to $3/2\lambda$. The repair facility thus increases the MTTF_S by $\mu/2\lambda^2$. □

11.5.2 Survivor Function

As discussed in Section 11.4.2, the set of states $\mathcal{X}$ of a system may be grouped in a set B of functioning states and a set $F = \mathcal{X} - B$ of failed states. In the present section, we assume that the failed states are absorbing states.

Consider a system that is in a specified functioning state at time $t = 0$. The survivor function $R(t)$ determines the probability that a system does not leave the set B of functioning states during the time interval $(0, t]$. The survivor function is thus

$$R(t) = \sum_{j \in B} P_j(t). \tag{11.74}$$

The Laplace transform of the survivor function is

$$R^*(s) = \sum_{j \in B} P_j^*(s).$$

11.5.3 Mean Time to the First System Failure

The mean time to system failure, MTTF_S, is determined by

$$\text{MTTF}_S = \int_0^\infty R(t)\, dt. \tag{11.75}$$

The Laplace transform of $R(t)$ is

$$R^*(s) = \int_0^\infty R(t)\, e^{-st}\, dt. \tag{11.76}$$

The MTTF_S of the system may thus be determined from (11.76) by inserting $s = 0$, such that

$$R^*(0) = \int_0^\infty R(t)\, dt = \text{MTTF}_S. \tag{11.77}$$

Example 11.8 (Example 11.7 (Cont.))

The Laplace transform of the survivor function for the two-component parallel system was in (11.71) found to be

$$R^*(s) = \frac{3\lambda + \mu + s}{s^2 + (3\lambda + \mu)s + 2\lambda^2}.$$

By introducing $s = 0$, we get

$$\text{MTTF}_S = R^*(0) = \frac{3\lambda + \mu}{2\lambda^2} = \frac{3}{2\lambda} + \frac{\mu}{2\lambda^2},$$

which is in accordance with (11.73). □

Procedure for Finding MTTF_S

As indicated in Example 11.7, the following procedure may be used to find the mean time to first failure, MTTF, of a system with state space $\mathcal{X} = \{0, 1, \dots, r\}$. See Billington and Allen (1992) and Pagès and Gondran (1980) for details and justification.

(1) Establish the transition rate matrix $\mathbb{A}$. and let $\boldsymbol{P}(t) = [P_0(t), P_1(t), \dots, P_r(t)]$ denote the distribution of the process at time t. Observe that $\mathbb{A}$ is a $(r+1) \times (r+1)$ matrix.

(2) Define the initial distribution $\boldsymbol{P}(0) = [P_0(0), P_1(0), \dots, P_r(0)]$ of the process, and verify that $\boldsymbol{P}(0)$ means that the system has a functioning state.

(3) Identify the failed states of the system, and define these states as absorbing states. Assume that there are k absorbing states.

(4) Delete the rows and columns of $\mathbb{A}$ corresponding to the absorbing states, that is, if j is an absorbing state, remove the entries a_{ji} and a_{ij} for all i from $\mathbb{A}$. Let $\mathbb{A}_R$ denote the reduced transition rate matrix. The dimension of $\mathbb{A}_R$ is $(r+1-k) \times (r+1-k)$.

(5) Let $\boldsymbol{P}^*(s) = [P_0^*(s), P_1^*(s), \dots, P_r^*(s)]$ denote the Laplace transform of $\boldsymbol{P}(t)$ and remove the entries of $\boldsymbol{P}^*(s)$ corresponding to absorbing states. Let $\boldsymbol{P}_R^*(s)$ denote the reduced vector. Observe that $\boldsymbol{P}_R^*(s)$ has dimension $(r + 1 - k)$.

(6) Remove the entries of $s\boldsymbol{P}^*(s) - \boldsymbol{P}(0)$ corresponding to absorbing states. Let $[s\boldsymbol{P}^*(s) - \boldsymbol{P}(0)]_R$ denote the reduced vector.

(7) Establish the equation

$$\boldsymbol{P}_R^*(s) \cdot \mathbb{A}_R = [s\boldsymbol{P}^*(s) - \boldsymbol{P}(0)]_R,$$

set $s = 0$ and determine $\boldsymbol{P}_R^*(0)$.

(8) The mean time-to-failure, MTTF_S, is determined by

$$\text{MTTF}_S = \sum P_j^*(0),$$

where the sum is taken over all j representing the $(r + 1 - k)$ nonabsorbing states.

Example 11.9 (Parallel structure of two independent components)
Reconsider the parallel structure of two independent components in Example 11.2, where the components have failure rates λ_1 and λ_2, and repair rates μ_1 and μ_2, respectively. The states of the system are defined in Table 11.1. The system is assumed to start out at time 0 in state 3 with both components functioning. The system is functioning as long as at least one of the components is functioning. The set B of functioning states is thus $\{1, 2, 3\}$. The system fails when both components are in a failed state, state 0.

In this example, we are primarily interested in determining the mean time to system failure MTTF_S. We therefore define state 0 to be an absorbing state, and set all departure rates from state 0 equal to zero. The transition rate matrix is then

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ \lambda_2 & -(\lambda_2 + \mu_1) & 0 & \mu_1 \\ \lambda_1 & 0 & -(\lambda_1 + \mu_2) & \mu_2 \\ 0 & \lambda_1 & \lambda_2 & -(\lambda_1 + \lambda_2) \end{pmatrix},$$

and the survivor function is

$$R(t) = P_1(t) + P_2(t) + P_3(t).$$

We now reduce the matrix equations by removing the row and the column corresponding to the absorbing state (state 0) and take Laplace transforms:

$$[P_1^*(0), P_2^*(0), P_3^*(0)] \cdot \begin{pmatrix} -(\lambda_2 + \mu_1) & 0 & \mu_1 \\ 0 & -(\lambda_1 + \mu_2) & \mu_2 \\ \lambda_1 & \lambda_2 & -(\lambda_1 + \lambda_2) \end{pmatrix} = [0, 0, -1].$$

This means that

$$P_1^*(0) = \frac{\lambda_1}{\lambda_2 + \mu_1} P_3^*(0) \tag{11.78}$$

$$P_2^*(0) = \frac{\lambda_2}{\lambda_1 + \mu_2} P_3^*(0). \tag{11.79}$$

$$\left(\frac{\lambda_1 \mu_1}{\lambda_2 + \mu_1} + \frac{\lambda_2 \mu_2}{\lambda_1 + \mu_2} - (\lambda_1 + \lambda_2) \right) P_3^*(0) = -1. \tag{11.80}$$

The last equation leads to

$$P_3^*(0) = \frac{1}{\lambda_1 \lambda_2 [1/(\lambda_1 + \mu_2) + 1/(\lambda_2 + \mu_1)]}. \tag{11.81}$$

Finally,

$$\begin{aligned} \text{MTTF}_S &= R^*(0) = P_1^*(0) + P_2^*(0) + P_3^*(0) \\ &= \frac{\lambda_1/(\lambda_2 + \mu_1) + \lambda_2/(\lambda_1 + \mu_2) + 1}{\lambda_1 \lambda_2 [1/(\lambda_1 + \mu_2) + 1/(\lambda_2 + \mu_1)]}, \end{aligned} \tag{11.82}$$

where $P_1^*(0)$ and $P_2^*(0)$ are determined by inserting (11.80) in (11.78) and (11.79), respectively.

Some Special Cases

(1) Nonrepairable system ($\mu_1 = \mu_2 = 0$)

$$\text{MTTF}_S = \frac{(\lambda_2/\lambda_1) + (\lambda_1/\lambda_2) + 1}{\lambda_1 + \lambda_2}.$$

When the two components have identical failure rates, $\lambda_1 = \lambda_2 = \lambda$, this expression is reduced to

$$\text{MTTF}_S = \frac{3}{2} \frac{1}{\lambda}. \tag{11.83}$$

(2) The two components have identical failure rates and identical repair rates ($\lambda_1 = \lambda_2 = \lambda$ and $\mu_1 = \mu_2 = \mu$). Then

$$\text{MTTF}_S = \frac{3}{2\lambda} + \frac{\mu}{2\lambda^2}.$$

□

11.6 Systems with Dependent Components

This section illustrates how a Markov model can be used to model dependent failures. Two simple situations are described: Systems exposed to CCFs and load-sharing systems, that are exposed to *cascading failures*. Dependent failures are discussed in Chapter 8.

11.6.1 Common Cause Failures

Consider a parallel structure of two identical components. The components may fail due to aging or other inherent defects. Such failures occur independent of each other with failure rate λ_I. The components are repaired independent of each other with repair rate μ.

An external event may occur that causes all functioning components to fail at the same time. Failures caused by the external event are CCFs. The external events occur with rate λ_C that is called the CCF rate.

The states of the system are named according to the number of components functioning. Thus, the state space is $\{0, 1, 2\}$. The state transition diagram of the parallel system with CCFs is shown in Figure 11.11.

The corresponding transition rate matrix is

$$\mathbb{A} = \begin{pmatrix} -2\mu & 2\mu & 0 \\ \lambda_C + \lambda_I & -(\lambda_I + \lambda_C + \mu) & \mu \\ \lambda_C & 2\lambda_I & -(2\lambda_I + \lambda_C) \end{pmatrix}.$$

Assume that we are interested in determining the mean time to system failure MTTF_S. Because the system fails as soon as it enters state 0, we define state 0 as an absorbing state, and remove the row and the column from the transition rate matrix corresponding to state 0.

As before, we assume that the system is in state 2 (both components are functioning) at time $t = 0$. By introducing Laplace transforms, we get the following matrix equations

$$[P_1^*(0), P_2^*(0)] \cdot \begin{pmatrix} -(\lambda_I + \lambda_C + \mu) & \mu \\ 2\lambda_I & -(2\lambda_I + \lambda_C) \end{pmatrix} = [0, -1].$$

The solutions are

$$P_1^*(0) = \frac{2\lambda_I}{(2\lambda_I + \lambda_C)(\lambda_I + \lambda_C) + \lambda_C\mu}$$

$$P_2^*(0) = \frac{\lambda_I + \lambda_C + \mu}{(2\lambda_I + \lambda_C)(\lambda_I + \lambda_C) + \lambda_C\mu},$$

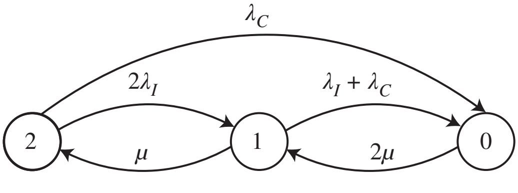


Figure 11.11 State transition diagram for a parallel structure of two components exposed to CCF.

and the mean time to system failure is

$$\text{MTTF}_S = P_2^*(0) + P_1^*(0) = \frac{3\lambda_I + \lambda_C + \mu}{(2\lambda_I + \lambda_C)(\lambda_I + \lambda_C) + \lambda_C\mu}. \tag{11.84}$$

As for the beta-factor model in Chapter 8, the common cause factor β is defined as

$$\beta = \frac{\lambda_C}{\lambda_C + \lambda_I} = \frac{\lambda_C}{\lambda},$$

where $\lambda = \lambda_C + \lambda_I$ is the total failure rate of a component, and the factor β denotes the fraction of CCFs among all failures of a component. To investigate how the beta-factor affects the MTTF_S, we insert β and λ into (11.84) to obtain

$$\begin{aligned}\text{MTTF}_S &= \frac{3(1-\beta)\lambda + \beta\lambda + \mu}{(2(1-\beta)\lambda + \beta\lambda)\lambda + \beta\lambda\mu} \\ &= \frac{3 - 2\beta\lambda + \mu}{(2-\beta)\lambda^2 + \beta\lambda\mu} = \frac{1}{\lambda}\frac{\lambda(3-2\beta) + \mu}{(2-\beta)\lambda + \beta\mu}.\end{aligned} \tag{11.85}$$

Figure 11.12 illustrates how the MTTF_S of a parallel system depends on the common cause factor β.

Consider two simple cases.

(1) $\beta = 0$ (i.e. only *independent failures*, $\lambda = \lambda_I$):

$$\text{MTTF}_S = \frac{3}{2\lambda_I} + \frac{\mu}{2\lambda_I^2},$$

which is what we obtained in Example 11.8.

(2) $\beta = 1$ (i.e. all failures are CCFs, $\lambda = \lambda_C$):

$$\text{MTTF}_S = \frac{1}{\lambda_C}\frac{\lambda_C + \mu}{\lambda_C + \mu} = \frac{1}{\lambda_C}.$$

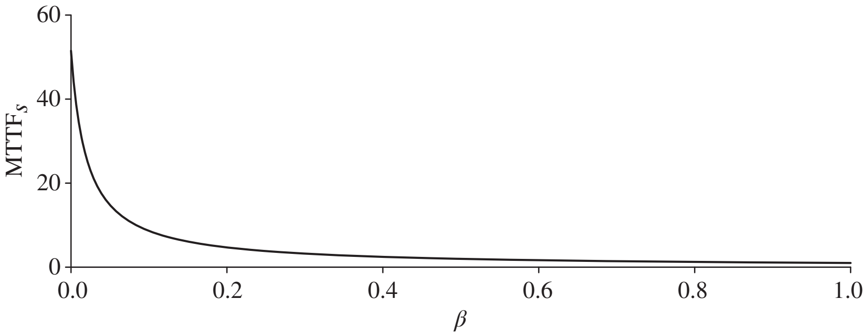


Figure 11.12 The MTTF_S of a parallel structure as a function of the common-cause factor β ($\lambda = 1$, and $\mu = 100$).

The last result is evident. Only CCFs are occurring, and they affect both components simultaneously with failure rate λ_C. For further details about the beta-factor model, see Chapter 8.

11.6.2 Load-Sharing Systems

Consider a parallel structure of two identical components. The components share a common load. If one component fails, the other component has to carry the whole load and the failure rate of this component is assumed to increase immediately when the load is increased. Hence, the failures of the two components are dependent. In Chapter 8, this type of dependency is referred to as *cascading failures*. The components may, for example, be pumps, compressors, or power generators. The following failure rates are assumed:

λ_h	Failure rate at normal load (i.e. when both components are functioning)
λ_f	Failure rate at full load (i.e. when one of the components is failed)

Let μ_h denote the repair rate of a component when only one component has failed, and let μ_f denote the repair rate of a component when both components have failed. Let the number of components that are functioning denote the state of the system. The state space is thus $\{0, 1, 2\}$. When the system has failed (state 0), all available repair resources are used to repair one of the components (usually the component that failed first). The system is stated up again (in state 1) as soon as this component is repaired. The state transition diagram of the system is given in Figure 11.13.

The transition rate matrix is

$$\begin{pmatrix} -\mu_f & \mu_f & 0 \\ \lambda_f & -(\mu_h + \lambda_f) & \mu_h \\ 0 & 2\lambda_h & -2\lambda_h \end{pmatrix}.$$

The system fails when both components fail (i.e. in state 0). To determine the mean time to system failure, MTTF_S, we define state 0 as an absorbing state, and remove the row and the column corresponding to this state from the transition rate matrix. If we assume that the system starts out at time $t = 0$ with both components

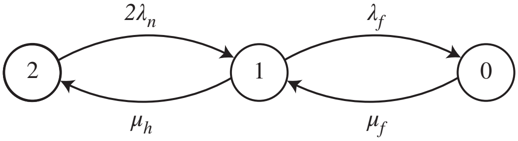


Figure 11.13 Parallel structure of two components sharing a common load.

functioning (state 2), and take Laplace transforms with $s = 0$, we get

$$[P_1^*(0), P_2^*(0)] \cdot \begin{pmatrix} -(\mu_h + \lambda_f) & \mu_h \\ 2\lambda_h & -2\lambda_h \end{pmatrix} = [0, -1].$$

The solution is

$$P_1^*(0) = \frac{1}{\lambda_f}$$

$$P_2^*(0) = \frac{\lambda_f + \mu_h}{2\lambda_h \lambda_f}.$$

The survivor function is $R(t) = P_1(t) + P_2(t)$, and the mean time to system failure is thus

$$\text{MTTF}_S = R^*(0) = P_1^*(0) + P_2^*(0) = \frac{1}{\lambda_f} + \frac{1}{2\lambda_h} + \frac{\mu_h}{2\lambda_h \lambda_f}. \tag{11.86}$$

Observe that when no repair is carried out ($\mu_h = 0$)

$$\text{MTTF}_S = \frac{1}{\lambda_f} + \frac{1}{2\lambda_h}. \tag{11.87}$$

When the load on the remaining component is not increased, such that $\lambda_f = \lambda_h$, we get $\text{MTTF}_S = 3/(2\lambda_h)$ in accordance with (11.83).

Example 11.10 (System of two generators)

Consider a power station with two generators of the same type. During normal operation, the generators are sharing the load and each generator has failure rate $\lambda_h = 1.6 \times 10^{-4}$ h^{-1}. When one of the generators fails, the load on the remaining generator is increased, and the failure rate increases to $\lambda_f = 8.0 \times 10^{-4}$ h^{-1} (five times as high as the normal failure rate). In addition, the system is exposed to CCFs. All generators in operation will fail at the same time when common cause events occur. The CCF rate is $\lambda_C = 2.0 \times 10^{-5}$ h^{-1}. When one generator fails, it is repaired. The mean time to repair, MTTR$_h$ is 12 hours, and the repair rate is therefore $\mu_h \approx 8.3 \times 10^{-2}$ h^{-1}. When the system fails, the MTTR one generator is MTTR$_f$ = 8 hours, and the repair rate is $\mu_f = 1.25 \times 10^{-1}$ h^{-1}. The state transition diagram of the generator system with load-sharing and CCFs is shown in Figure 11.14.

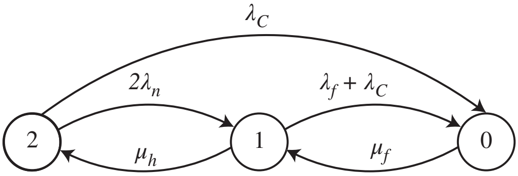


Figure 11.14 State transition diagram for the generator system with load-sharing and CCFs.

The steady state probabilities can be found by the same approach as we have shown several times (see, for example, Example 11.6).

$$P_2 = \frac{\mu_h\mu_f}{(\lambda_f + \lambda_C + \mu_f)(\lambda_C + 2\lambda_h) + \lambda_C\mu_h + \mu_h\mu_f} \approx 0.995\ 75$$

$$P_1 = \frac{(\lambda_C + 2\lambda_h)\mu_f}{(\lambda_f + \lambda_C + \mu_f)(\lambda_C + 2\lambda_h) + \lambda_C\mu_h + \mu_h\mu_f} \approx 0.004\ 06$$

$$P_0 = \frac{(\lambda_f + \lambda_C)(\lambda_C + 2\lambda_h) + \lambda_C\mu_h}{(\lambda_f + \lambda_C + \mu_f)(\lambda_C + 2\lambda_h) + \lambda_C\mu_h + \mu_h\mu_f} \approx 0.000\ 19.$$

The mean time to system failure is found from the Laplace transforms:

$$[P_1^*(0), P_2^*(0)] \cdot \begin{pmatrix} -(\lambda_f + \lambda_C + \mu_h) & \mu_h \\ 2\lambda_h & -(\lambda_C + 2\lambda_h) \end{pmatrix} = [0, -1].$$

We find that

$$\text{MTTF}_S = P_1^* + P_2^* = \frac{2\lambda_h + \lambda_f + \lambda_C + \mu_h}{(\lambda_C + 2\lambda_h)(\lambda_f + \lambda_C + \mu_h) - 2\lambda_n\mu_h}$$

$$\approx 43\ 421\ \text{h} \ \approx\ 4.96\ \text{yr}.$$

□

11.7 Standby Systems

Standby systems are introduced in Section 6.4 where the survivor function $R(t)$ and the mean time-to-failure MTTF_S are determined for some simple nonrepairable standby systems. This section deals with some simple two-item repairable standby systems analyzed by Markov methods. The system considered is illustrated in Figure 11.15. Item A is initially (at time $t = 0$) the operating item and S is the sensing and changeover device.

A standby system may be operated and repaired in a number of different ways:

- The standby item may be cold or partly loaded.
- The changeover device may have several failure modes, such as "fail to switch," "spurious switching," and "disconnect."
- Failure of the standby item may be hidden (nondetectable) or detectable.

In the present section, a few operation and repair modes of a standby system are illustrated. Generalizations to more complicated systems and operational modes are often straightforward, at least in theory, but the computations may require a computer.

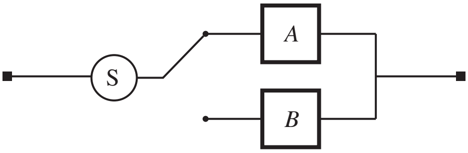


Figure 11.15 Two-item standby system.

11.7.1 Parallel System with Cold Standby and Perfect Switching

Because the standby item is passive, it is assumed not to fail in the standby state. The switching is assumed to be perfect. Failure of the active item is detected immediately, and the standby item is activated with probability 1. The failure rate of item i in operating state is denoted λ_i for $i = A, B$. When the active item has failed, a repair action is initiated immediately. The time to repair is exponentially distributed with repair rate μ_i for $i = A, B$. When a repair action is completed, the item is placed in standby state.

The possible states of the system are listed in Table 11.3 where O denotes operating state, S denotes standby state, and F denotes failed state.

System failure occurs when the operating item fails before repair of the other item is completed. The failed state of the system is thus state 0 in Table 11.3. When both items have failed, they are repaired simultaneously, and the system is thus brought back to state 4. The repair rate in this case is denoted μ. The state transition diagram of the standby system is illustrated in Figure 11.16. The transition rate matrix is

$$\mathbb{A} = \begin{pmatrix} -\mu & 0 & 0 & 0 & \mu \\ \lambda_A & -(\lambda_A + \mu_B) & 0 & 0 & \mu_B \\ 0 & \lambda_B & -\lambda_B & 0 & 0 \\ \lambda_B & 0 & \mu_A & -(\lambda_B + \mu_A) & 0 \\ 0 & 0 & 0 & \lambda_A & -\lambda_A \end{pmatrix}. \tag{11.88}$$

Table 11.3 The possible states of a two-item parallel system with cold standby and perfect switching.

System state	State of item *A*	State of item *B*
4	O	S
3	F	O
2	S	O
1	O	F
0	F	F

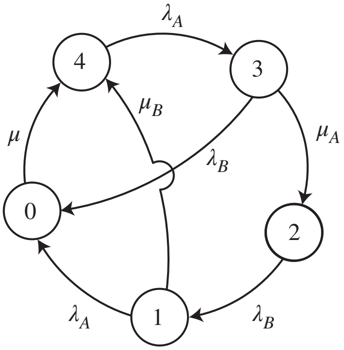


Figure 11.16 State transition diagram of a two-item parallel structure with cold standby and perfect switching.

The steady state probabilities may be determined according to the procedures described in Section 11.3. The survivor function $R(t)$ and the mean time-to-failure MTTF_S of the system can be determined by considering the failed state of the system (state 0) to be an absorbing state. Suppose that the initial state at $t = 0$ is state 4. By deleting the row and the column of the transition rate matrix corresponding to the absorbing state 0, we get the reduced matrix $\mathbb{A}_R$

$$\mathbb{A}_R = \begin{pmatrix} -(\lambda_A + \mu_B) & 0 & 0 & \mu_B \\ \lambda_B & -\lambda_B & 0 & 0 \\ 0 & \mu_A & -(\lambda_B + \mu_A) & 0 \\ 0 & 0 & \lambda_A & -\lambda_A \end{pmatrix}.$$

By taking Laplace transforms (with $s = 0$), we get the equations

$$[P_1^*(0), P_2^*(0), P_3^*(0), P_4^*(0)] \cdot \mathbb{A}_R = [0, 0, 0, -1].$$

The solution is

$$\begin{aligned} P_2^*(0) &= \frac{\lambda_A + \mu_B}{\lambda_B}\, P_1^*(0) \\ P_3^*(0) &= \frac{\lambda_A + \mu_B}{\mu_A}\, P_1^*(0) \\ P_4^*(0) &= \frac{\lambda_B + \mu_A}{\lambda_A}\, P_3^*(0) \\ &= \frac{(\lambda_A + \mu_B)(\lambda_B + \mu_A)}{\lambda_A \mu_A}\, P_1^*(0) \\ &= \frac{1 + \mu_B P_1^*(0)}{\lambda_A}. \end{aligned}$$

Thus,

$$P_1^*(0) = \frac{\mu_A}{\lambda_A \lambda_B + \lambda_A \mu_A + \lambda_B \mu_B}.$$

The mean time-to-failure MTTF_S of the system is now, by using (11.77),

$$\begin{aligned}\text{MTTF}_S &= R^*(0) = P_1^*(0) + P_2^*(0) + P_3^*(0) + P_4^*(0)\\ &= \frac{1}{\lambda_A} + \frac{1}{\lambda_B} + \frac{\mu_A}{\lambda_B}\left(\frac{1}{\lambda_B} - \frac{1}{\lambda_B + \mu_A + \frac{\lambda_B}{\lambda_A}\mu_B}\right).\end{aligned} \tag{11.89}$$

For a nonrepairable system, $\mu_A = \mu_B = 0$. Then

$$\text{MTTF}_S = \frac{1}{\lambda_A} + \frac{1}{\lambda_B},$$

which is an obvious result.

11.7.2 Parallel System with Cold Standby and Perfect Switching (Item *A* is the Main Operating Item)

Reconsider the standby system in Figure 11.15. Assume that item A is the main operating item. This means that item B is only used when A is in a failed state and under repair. Item A is thus put into operation again as soon as the repair action is completed. System failure occurs when the operating item B fails before repair of item A is completed. The failed state of the system is thus state 0 in Table 11.3. When both items have failed, they are repaired simultaneously and brought back to state 4. The repair rate in this case is denoted μ. States 1 and 2 in Table 11.3 are therefore irrelevant states for this system. The state transition diagram of this system is illustrated in Figure 11.17.

The transition rate matrix is

$$\begin{pmatrix} -\mu & 0 & \mu \\ \lambda_B & -(\lambda_B + \mu_A) & \mu_A \\ 0 & \lambda_A & -\lambda_A \end{pmatrix}. \tag{11.90}$$

The steady state probabilities are determined by

$$[P_0, P_3, P_4] \cdot \begin{pmatrix} -\mu & 0 & \mu \\ \lambda_B & -(\lambda_B + \mu_A) & \mu_A \\ 0 & \lambda_A & -\lambda_A \end{pmatrix} = [0, 0, 0]$$

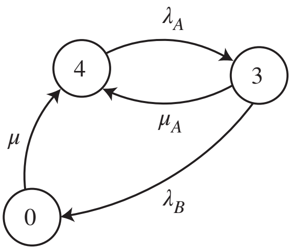


Figure 11.17 State transition diagram of a two-item parallel structure with cold standby and perfect switching (item *A* is the main operating item).

and

$$P_0 + P_3 + P_4 = 1.$$

The solution is

$$P_0 = \frac{\lambda_A\lambda_B}{\lambda_A\lambda_B + \lambda_A\mu + \lambda_B\mu + \mu\mu_A}$$

$$P_3 = \frac{\lambda_A\mu}{\lambda_A\lambda_B + \lambda_A\mu + \lambda_B\mu + \mu\mu_A}$$

$$P_4 = \frac{\lambda_B\mu + \mu\mu_A}{\lambda_A\lambda_B + \lambda_A\mu + \lambda_B\mu + \mu\mu_A},$$

where P_j is the mean proportion of time the system is spending in state j for $j = 0, 3, 4$.

The frequency of system failures, ω_F, is in this case equal to the visit frequency to state 0, i.e.

$$\omega_F = \nu_0 = \frac{P_0}{\mu}.$$

The MTTF$_S$ of the system is determined as in Section 11.5.3. By deleting the row and the column of the transition rate matrix in (11.90) and taking Laplace transforms (with $s = 0$), we obtain

$$[P_3^*(0), P_4^*(0)] \cdot \begin{pmatrix} -(\lambda_B + \mu_A) & \mu_A \\ \lambda_A & -\lambda_A \end{pmatrix} = [0, -1].$$

The solution is

$$P_3^*(0) = \frac{1}{\lambda_B}$$

$$P_4^*(0) = \frac{1}{\lambda_A} + \frac{\mu_A}{\lambda_A\lambda_B}.$$

The mean time-to-failure of the system is thus

$$\text{MTTF}_S = R^*(0) = P_3^*(0) + P_4^*(0) = \frac{1}{\lambda_A} + \frac{1}{\lambda_B} + \frac{\mu_A}{\lambda_A\lambda_B}. \tag{11.91}$$

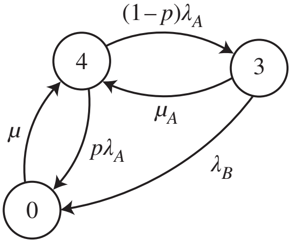
The MTTR of the system is

$$\text{MTTR}_S = \frac{1}{\mu}.$$

The average availability A of the system is thus

$$A = \frac{\text{MTTF}_S}{\text{MTTF}_S + \text{MTTR}_S} = \frac{1/\lambda_A + 1/\lambda_B + \mu_A/(\lambda_A\lambda_B)}{1/\lambda_A + 1/\lambda_B + \mu_A/(\lambda_A\lambda_B) + 1/\mu}.$$

Figure 11.18 State transition diagram of a two-item parallel structure with cold standby and imperfect switching (item *A* is the main operating item).



11.7.3 Parallel System with Cold Standby and Imperfect Switching (Item *A* is the Main Operating Item)

Reconsider the standby system in Figure 11.15 (again let item 1 be item *A* and item 2 be item *B*). Assume that the switching is no longer perfect. When the active item *A* fails, the standby item *B* will be activated properly with probability $(1 - p)$. The probability p may also include a "fail to start" probability of the standby item. The state transition diagram of the system is illustrated in Figure 11.18. From state 4, the system may have a transition to state 3 with rate $(1 - p)\lambda_A$ and to state 0 with rate $p\lambda_A$.

The steady state probabilities are determined by

$$[P_0, P_3, P_4] \cdot \begin{pmatrix} -\mu & 0 & \mu \\ \lambda_B & -(\lambda_B + \mu_A) & \mu_A \\ p\lambda_A & (1-p)\lambda_A & -\lambda_A \end{pmatrix} = [0, 0, 0] \tag{11.92}$$

and

$$P_0 + P_3 + P_4 = 1.$$

The solution is

$$P_0 = \frac{\lambda_A\lambda_B + p\lambda_A\mu_A}{\lambda_A\lambda_B + p\lambda_A\mu_A + (1-p)\lambda_A\mu + \lambda_B\mu + \mu\mu_A}$$

$$P_3 = \frac{\lambda_A\mu(1-p)}{\lambda_A\lambda_B + p\lambda_A\mu_A + (1-p)\lambda_A\mu + \lambda_B\mu + \mu\mu_A}$$

$$P_4 = \frac{\lambda_B\mu + \mu\mu_A}{\lambda_A\lambda_B + p\lambda_A\mu_A + (1-p)\lambda_A\mu + \lambda_B\mu + \mu\mu_A}.$$

The MTTF$_S$ can be determined from

$$[P_3^*(0), P_4^*(0)] \cdot \begin{pmatrix} -(\lambda_B + \mu_A) & \mu_A \\ (1-p)\lambda_A & -\lambda_A \end{pmatrix} = [0, -1],$$

which leads to

$$P_3^*(0) = \frac{1-p}{\lambda_B + p\mu_A}$$

$$P_4^*(0) = \frac{\lambda_B + \mu_A}{\lambda_A(\lambda_B + p\mu_A)}.$$

Thus,

$$\text{MTTF}_S = R^*(0) = P_3^*(0) + P_4^*(0) = \frac{(1-p)\lambda_A + \lambda_B + \mu_A}{\lambda_A(\lambda_B + p\mu_A)}. \tag{11.93}$$

11.7.4 Parallel System with Partly Loaded Standby and Perfect Switching (Item *A* is the Main Operating Item)

Reconsider the standby system in Figure 11.15 but assume that the standby item *B* (i.e. item 2 in the figure) may fail in standby mode and have a hidden failure when activated. The failure rate of item *B* in standby mode is denoted λ_B^s and is normally less than the corresponding failure rate during operation. In addition to the transition in Figure 11.17, this system may also have transitions from states 4 to 1 (in Table 11.3) and from states 1 to 0. The state transition diagram is illustrated in Figure 11.19.

The steady-state probabilities are determined by

$$[P_0, P_1, P_3, 4] \cdot \begin{pmatrix} -\mu & 0 & 0 & \mu \\ \lambda_A & -\lambda_A & 0 & 0 \\ \lambda_B & 0 & -(\lambda_B + \mu_A) & \mu_A \\ 0 & \lambda_B^s & \lambda_A & -(\lambda_A + \lambda_B^s) \end{pmatrix} = [0, 0, 0, 0]$$

and

$$P_0 + P_1 + P_3 + P_4 = 1.$$

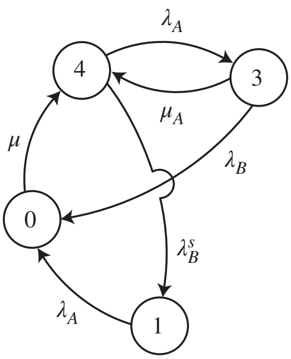


Figure 11.19 State transition diagram of a two-item parallel structure with partly loaded standby and perfect switching (item *A* is the main operating item).

The MTTF_S can be determined from

$$[P_1^*(0), P_3^*(0), P_4^*(0)] \cdot \begin{pmatrix} -\lambda_A & 0 & 0 \\ 0 & -(\lambda_B + \mu_A) & \mu_A \\ \lambda_B^s & \lambda_A & -(\lambda_A + \lambda_B^s) \end{pmatrix} = [0, 0, -1].$$

$$P_1^*(0) = \frac{\frac{\lambda_B^s}{\lambda_A}(\lambda_B + \mu_A)}{\lambda_A \lambda_B + \lambda_B \lambda_B^s + \lambda_B^s \mu_A}$$

$$P_3^*(0) = \frac{\lambda_A}{\lambda_A \lambda_B + \lambda_B \lambda_B^s + \lambda_B^s \mu_A}$$

$$P_4^*(0) = \frac{\lambda_B + \mu_A}{\lambda_A \lambda_B + \lambda_B \lambda_B^s + \lambda_B^s \mu_A}.$$

Thus,

$$\begin{aligned} \text{MTTF}_S &= R^*(0) = P_1^*(0) + P_3^*(0) + P_4^*(0) \\ &= \frac{\left(\frac{\lambda_B^s}{\lambda_A} + 1\right)(\lambda_B + \mu_A) + \lambda_A}{\lambda_A \lambda_B + \lambda_B \lambda_B^s + \lambda_B^s \mu_A}. \end{aligned} \tag{11.94}$$

Let us now assume that we have two items of the same type and no repair is carried out. Let $\lambda_A = \lambda_B = \lambda$, and $\lambda_A^S = \lambda_B^S = \lambda^S$. In this case, the mean time-to-failure is

$$\text{MTTF}_S = \frac{1}{\lambda + \lambda^S}\left(2 + \frac{\lambda^S}{\lambda}\right). \tag{11.95}$$

Observe that when $\lambda = \lambda^S$, Equation (11.95) reduces to the mean time-to-failure of an active parallel system.

11.8 Markov Analysis in Fault Tree Analysis

We now illustrate how results from Markov analysis can be used in fault tree analysis. Assume that a fault tree has been established with respect to a TOP event (a system failure or accident) in a specific system. The fault tree has n basic events (components) and k minimal cut sets $K_1, K_2, \ldots, K_k$.

The probability of the fault tree TOP event may be approximated by the upper bound approximation (6.94

$$Q_0(t) \approx 1 - \prod_{j=1}^{k} [1 - \check{Q}_j(t)]. \tag{11.96}$$

Let us assume that the TOP event is a system failure, such that $Q_0(t)$ is the system unavailability. The average (limiting) system unavailability is thus approximately

$$Q_0 \approx 1 - \prod_{j=1}^{k} (1 - \check{Q}_j), \tag{11.97}$$

where $\check{Q}_j$ denotes the average unavailability of the minimal cut parallel structure corresponding to the minimal cut set $K_j, j = 1, 2, \dots, k$.

In the rest of this section, assume that component i has constant failure rate λ_i, mean time to repair MTTR_i, and constant repair rate $\mu_i = 1/\text{MTTR}_i$ for $i = 1, 2, \dots, n$. Furthermore, assume that $\lambda_i \ll \mu_i$ for all $i = 1, 2, \dots, n$.

The average unavailability q_i of component i is $\lambda_i/(\mu_i + \lambda_i)$, which may be approximated by $\lambda_i \text{MTTR}_i$, such that

$$\check{Q}_j = \prod_{i \in K_j} \frac{\lambda_i}{\mu_i + \lambda_i} \approx \prod_{i \in K_j} \lambda_i \text{MTTR}_i. \tag{11.98}$$

The TOP event probability (system unavailability) is thus approximately

$$Q_0 \approx 1 - \prod_{j=1}^{k} \left(1 - \prod_{i \in K_j} \lambda_i \text{MTTR}_i \right) \tag{11.99}$$

or

$$Q_0 \approx \sum_{j=1}^{k} \prod_{i \in K_j} \lambda_i \text{MTTR}_i. \tag{11.100}$$

11.8.1 Cut Set Information

Consider a specific minimal cut parallel structure K_j, for $j = 1, 2, \dots, k$. As before we assume that the components fail and are repaired independent of each other.

When all the components of the cut set K_j are in a failed state, we have a *cut set failure*. The mean duration of a failure of cut set K_j is from (11.47)

$$\text{MTTR}_j = \frac{1}{\sum_{i \in K_j} \mu_i}. \tag{11.101}$$

The expected frequency of cut set failures, ω_j is from (11.48)

$$\omega_j \approx \left(\prod_{i \in K_j} \frac{\lambda_i}{\mu_i} \right) \left(\sum_{i \in K_j} \mu_i \right) \tag{11.102}$$

and, the mean time between failures (MTBF) of cut set K_j is

$$\text{MTBF}_j = \frac{1}{\omega_K}.$$

Observe that MTBF_j also includes the MDT of the cut parallel structure. The downtime is, however, usually negligible compared to the uptime.

11.8.2 System Information

The system may be considered as a series structure of its k minimal cut parallel structures. If the cut parallel structures were independent and the downtimes were negligible, the frequency, ω_S of system failures would be

$$\omega_S = \sum_{j=1}^{k} \omega_j. \tag{11.103}$$

In general, this formula is not correct, because (i) the minimal cut parallel structures are usually not independent, and (ii) the downtimes of the minimal cut parallel structures are often not negligible.

For a system with very high availability, (11.102) is an adequate approximation for the expected frequency ω_S of system failures.

The mean time between system failures, MTBF_S in the steady state situation is approximately

$$\text{MTBF}_S \approx \frac{1}{\omega_S}.$$

The mean system downtime per system failure is from (11.57) approximately

$$\text{MTTR}_S \approx \frac{\sum_{j=1}^{k} \omega_j \text{MTTR}_j}{\sum_{j=1}^{k} \omega_j}.$$

The average system availability may now be approximated by

$$A_S = \frac{\text{MTBF}_S}{\text{MTBF}_S + \text{MTTR}_S}.$$

The formulas in this section are used in some of the computer programs for fault tree analysis.

11.9 Time-Dependent Solution

Reconsider the Kolmogorov forward equations (11.19)

$$\boldsymbol{P}(t) \cdot \mathbb{A} = \dot{\boldsymbol{P}}(t),$$

where $\boldsymbol{P}(t) = [P_0(t), P_1(t), \dots, P_r(t)]$ is the distribution of the process at time t. Assume that we know the distribution of the system state at time 0, $\boldsymbol{P}(0)$. Usually, we know that the system is in a specific state i at time 0, but sometimes we only know that it has a specific distribution.

It is, in principle, possible to solve the Kolmogorov equations and find $\boldsymbol{P}(t)$ by

$$\boldsymbol{P}(t) = \boldsymbol{P}(0) \cdot e^{t\mathbb{A}} = \boldsymbol{P}(0) \cdot \sum_{k=0}^{\infty} \frac{t^k \mathbb{A}^k}{k!}, \tag{11.104}$$

where $\mathbb{A}^0$ is the identity matrix $\mathbb{I}$. To determine $\boldsymbol{P}(t)$ from (11.104) is sometimes time-consuming and inefficient. The numerical computation of this formula by discretization of the time is proposed on the `book companion site` with a Python script.

When we study a system with absorbing states, such as the parallel system in Example 11.7, we may define a column vector $\boldsymbol{C}$ with entries 1 and 0, where 1 corresponds to a functioning state, and 0 corresponds to a failed state. In Example 11.7, the states 1 and 2 are functioning, and state 0 is failed. The (column) vector is therefore $\boldsymbol{C} = [0, 1, 1]^T$. The survivor function of the system is then given by

$$R(t) = \boldsymbol{P}(0) \cdot \sum_{k=0}^{\infty} \frac{t^k \mathbb{A}^k}{k!} \cdot \boldsymbol{C}. \tag{11.105}$$

It is also possible to use that

$$e^{t\mathbb{A}} = \lim_{k\to\infty} (\mathbb{I} + t \cdot \mathbb{A}/k)^k,$$

and approximate $\boldsymbol{P}(t)$ by

$$\boldsymbol{P}(t) \approx \boldsymbol{P}(0) \cdot (\mathbb{I} + t \cdot \mathbb{A}/n)^n, \tag{11.106}$$

for a "sufficiently" large n. See Bon (1995, pp. 176–182) for further approximations and discussions.

11.9.1 Laplace Transforms

An alternative approach is to use Laplace transforms. A brief introduction to Laplace transforms is given in Appendix B.

Again, assume that we know $\boldsymbol{P}(0)$, the distribution of the Markov process at time 0. The state equations (11.19) for the Markov process at time t are seen to be a set of linear, first-order differential equations. The easiest and most widely used method to solve such equations is by Laplace transforms.

The Laplace transform of the state probability $P_j(t)$ is denoted by $P_j^*(s)$, and the Laplace transform of the time derivative of $P_j(t)$ is, according to Appendix B:

$$\mathcal{L}[\dot{P}_j(t)] = sP_j^*(s) - P_j(0) \qquad \text{for } j = 0, 1, 2, \dots, r.$$

The Laplace transform of the state equations (11.19) is thus in matrix terms

$$\boldsymbol{P}^*(s) \cdot \mathbb{A} = s\boldsymbol{P}^*(s) - \boldsymbol{P}(0). \tag{11.107}$$

By using Laplace transforms, the differential equations are reduced to a set of linear equations. The Laplace transforms $P_j^*(s)$ may now be computed from (11.107). Afterward the state probabilities $P_j(t)$ may be determined from the inverse Laplace transforms.

Example 11.11 Reconsider the single component in Example 11.5, with transition rate matrix

$$\mathbb{A} = \begin{pmatrix} -\mu & \mu \\ \lambda & -\lambda \end{pmatrix}.$$

Assume that the component is functioning at time $t = 0$, such that $\boldsymbol{P}(0) = (P_0(0), P_1(0)) = (0, 1)$. The Laplace transform of the state equation is then from (11.107)

$$(P_0^*(s), P_1^*(s)) \cdot \begin{bmatrix} -\mu & \mu \\ \lambda & -\lambda \end{bmatrix} = (sP_0^*(s) - 0, sP_1^*(s) - 1).$$

Thus,

$$\begin{aligned} -\mu P_0^*(s) + \lambda P_1^*(s) &= sP_0^*(s) \\ \mu P_0^*(s) - \lambda P_1^*(s) &= sP_1^*(s) - 1. \end{aligned} \tag{11.108}$$

By adding these two equations, we get

$$sP_0^*(s) + sP_1^*(s) = 1.$$

Thus,

$$P_0^*(s) = \frac{1}{s} - P_1^*(s).$$

By inserting this $P_0^*(s)$ into (11.108), we obtain

$$\frac{\mu}{s} - \mu P_1^*(s) - \lambda P_1^*(s) = sP_1^*(s) - 1.$$

$$P_1^*(s) = \frac{1}{\lambda + \mu + s} + \frac{\mu}{s} \frac{1}{\lambda + \mu + s}.$$

To find the inverse Laplace transform, we rewrite this expression as

$$P_1^*(s) = \frac{\lambda}{\lambda + \mu} \frac{1}{\lambda + \mu + s} + \frac{\mu}{\lambda + \mu} \frac{1}{s}. \tag{11.109}$$

From Appendix B, the inverse Laplace transform of (11.109) is

$$P_1(t) = \frac{\mu}{\mu + \lambda} + \frac{\lambda}{\mu + \lambda}\, e^{-(\lambda+\mu)t},$$

which is the same result we gave in Example 11.5. □

To find the time-dependent state probabilities for a complicated system is usually a difficult task and is not discussed any further in this book. In most practical applications, we are primarily interested in the steady state probabilities, and do not need to find the time-dependent probabilities.

11.10 Semi-Markov Processes

Section 11.2 defines a continuous-time Markov process as a stochastic process having the properties that each time it enters a state i:

(1) The amount of time the process spends in state i before making a transition into a different state is exponentially distributed with rate α_i.
(2) When the process leaves state i, it will next enter state j with some probability P_{ij}, where $\sum_{\substack{j=0 \\ j\neq i}}^{r} P_{ij} = 1$.

An obvious extension to this definition is to allow the time the process spends in state i (the sojourn time in state i) to have a general "life" distribution, and also to let this distribution be dependent on the state to which the process will go. A semi-Markov process may be defined as (e.g. see Ross 1996):

Definition 11.2 (Semi-Markov process)
A stochastic process $\{X(t), t \geq 0\}$ with state space $\mathcal{X} = \{0, 1, 2, \dots, r\}$ such that, whenever the process enters state i:

(1) The next state it will enter is state j with probability P_{ij}, for i, j in $\mathcal{X}$.
(2) Given that the next state to be entered is state j, the time until the transition from i to j occurs has distribution F_{ij}.

□

The skeleton of the semi-Markov process is defined in the same way as for the continuous-time Markov process (see Section 11.2), and is a discrete-time Markov chain. The semi-Markov process is said to be irreducible if the skeleton is irreducible.

The distribution of the sojourn time $\widetilde{T}_i$ in state i is

$$F_i(t) = \sum_{\substack{j=0 \\ j\neq i}}^{r} P_{ij}\, F_{ij}(t).$$

The mean sojourn time in state i is

$$\mu_i = E(\widetilde{T}_i) = \int_0^{\infty} t\, dF_i(t).$$

Observe that if $F_{ij}(t) = 1 - e^{-\alpha_i t}$, the semi-Markov process is a (continuous-time) Markov process.

Let T_{ii} denote the time between successive transitions into state i, and let $\mu_{ii} = E(T_{ii})$. The visits to state i will now be a renewal process, and we may use the theory of renewal processes described in Chapter 10.

If we let $N_i(t)$ denote the number of times in $[0, t]$ that the process is in state i, the family of vectors

$$[N_0(t), N_1(t), \dots, N_r(t)] \qquad \text{for} \;\; t \geq 0$$

is called a *Markov renewal process*.

If the semi-Markov process is irreducible and if T_{ii} has a nonlattice distribution with finite mean, then

$$\lim_{t\to\infty} \Pr(X(t) = i \mid X(0) = j) = P_i$$

exists, and is independent of the initial state. Furthermore,

$$P_i = \frac{\mu_i}{\mu_{ii}}.$$

For a proof, see Ross (1996). P_i is the proportion of transitions into state i, and is also equal to the long-run proportion of time the process is in state i.

When the skeleton (the embedded process) is irreducible and positive recurrent, we may find the stationary distribution of the skeleton $\pi = [\pi_0, \pi_1, \dots, \pi_r]$ as the unique solution of

$$\pi_j = \sum_{i=0}^{r} \pi_i P_{ij},$$

where $\sum_i \pi_i = 1$ and $\pi_j = \lim_{n\to\infty} \Pr(X_n = j$ (because we assume that the Markov process is aperiodic). Because the π_j is the proportion of transitions that are into state j, and μ_j is the mean time spent in state j per transition, it seems intuitive that the limiting probabilities should be proportional to $\pi_j \mu_j$. In fact,

$$P_j = \frac{\pi_j \mu_j}{\sum_i \pi_i \mu_i}.$$

For a proof, see Ross (1996).

Semi-Markov processes are not discussed any further in this book. More information may be found in Ross (1996), Cocozza-Thivent (1997), Limnios and Oprisan (2001), and Grabski (2015). There are not so many applications for such processes in reliability. For example, a system with two redundant identical items and time-dependent transition rates is not a semi-Markov process: when one of the items is failed, we do not have enough information for the current state and the time spent in the current state, to know the time for the next transition. We also need to know how long time the surviving item has been functioning.

One of the only reasonable applications of semi-Markov processes in reliability is for one item with intermediate degraded states and time-dependent transition rates. When a new state is reached, the rate for the next transition can be calculated with information from the current state only.

11.11 Multiphase Markov Processes

A multiphase Markov process is defined as

Definition 11.3 (Multiphase Markov process)
A Markov process where the parameters and the state of the system can be changed at predefined points in time, such as when PM tasks are carried out. The *phases* indicate the time periods between the changes. □

Two situations are considered:

(1) A PM task alters the transition matrix of the Markov process. This may happen when:
 - the PM task reduces some failure rates, or when
 - stresses during the PM task increases some transition rates.

(2) A PM task changes the state in which the system is restarted.

11.11.1 Changing the Transition Rates

Let $t_1, t_2, \dots, t_n$ be the predefined dates for PM tasks, and let $t_0 = 0$. Between t_{i-1} and t_i, the process evolves according to a homogeneous Markov process with transition matrix $\mathbb{A}_i$. The transition rates in matrix $\mathbb{A}_i$ may change just after time t_i depending on the effects of the PM task. Assume that PM tasks take no time. We want to establish the probability distribution of the chain at any time t and assume that the state probability vector at time 0 is given: $\boldsymbol{P}(t_0) = \boldsymbol{P}(0)$. In practice, this vector tells the probabilities to be in each and every state of the Markov process at time 0. We usually specify the system to be in the "new" state by putting probability 1 to the new state and 0 to the others.

It is possible to calculate the probability distribution of the process at any time t for $0 \le t \le t_1$:

$$\boldsymbol{P}(t) = \boldsymbol{P}(0) \cdot e^{\mathbb{A}_1 t}.$$

Until time t_1, the process evolves with transition matrix $\mathbb{A}_1$, then:

$$\boldsymbol{P}(t_1) = \boldsymbol{P}(0) \cdot e^{\mathbb{A}_1 t_1}.$$

Between t_1 and t_2, the process evolves with transition matrix $\mathbb{A}_2$ and the initial distribution $\boldsymbol{P}(t_1)$. For any time t, such that $t_1 < t \leq t_2$:

$$\boldsymbol{P}(t) = \boldsymbol{P}(t_1) \cdot e^{\mathbb{A}_2(t-t_1)} = \boldsymbol{P}(0) \cdot e^{\mathbb{A}_1 t_1} \cdot e^{\mathbb{A}_2(t-t_1)},$$

and so on. For any t_i, such that $i \geq 1$:

$$\boldsymbol{P}(t_i) = \boldsymbol{P}(0) \cdot \prod_{k=1}^{k=i} e^{\mathbb{A}_k(t_k - t_{k-1})},$$

and for any t such that $t_i < t \leq t_{i+1}$, the distribution is

$$\boldsymbol{P}(t) = \boldsymbol{P}(0) \cdot e^{A_1 t} \qquad \text{for } \ i = 0$$

$$\boldsymbol{P}(t) = \boldsymbol{P}(0) \cdot \left(\prod_{k=1}^{k=i} e^{\mathbb{A}_k(t_k - t_{k-1})} \right) \cdot e^{\mathbb{A}_{i+1}(t-t_i)} \qquad \text{for } \ i \geq 1.$$

The numerical computation of this formula by discretization of the time is proposed on the `book companion site` with a Python script.

11.11.2 Changing the Initial State

A maintenance task at time t_i may change the transition rates of the matrix $\mathbb{A}_i$ but also the state in which the process restarts after a maintenance task or an inspection. This can be modeled by a linear transformation of the probability $\boldsymbol{P}(t_i)$: the probability vector after the maintenance at time t_i is $\boldsymbol{P}(t_i) \cdot \mathbb{B}_i$, where $\mathbb{B}_i$ is an $N \times N$ matrix such that the sum of each row is equal to 1. The term b_{lj} in the matrix $\mathbb{B}_i$ is the probability that the item is in state j after maintenance, given that it was in state l just before the maintenance task is completed. If the maintenance task duration is neglected, for $t_i < t \leq t_{i+1}$, we have

$$\boldsymbol{P}(t) = \boldsymbol{P}(0) \cdot e^{A_1 t} \qquad \text{for } \ i = 0$$

$$\boldsymbol{P}(t) = \boldsymbol{P}(0) \cdot \left(\prod_{k=1}^{i} e^{\mathbb{A}_k(t_k - t_{k-1})} \cdot \mathbb{B}_k \right) \cdot e^{\mathbb{A}_{i+1}(t-t_i)} \qquad \text{for } \ i \geq 1.$$

If the item is taken out of operation during the maintenance/inspections task and the duration of the task is not negligible (considered as constant), the same formalism can be used with a time lag. The maintenance task duration is denoted m_a and the item is taken out of operation during the task. This means that after a maintenance task, beginning at time t_i, the system is restarted at time $t_i + m_a$ with the distribution $\boldsymbol{P}(t_i) \cdot \mathbb{B}_i$. This case yields:

$$\begin{aligned} \boldsymbol{P}(t_1) &= \boldsymbol{P}(0) \cdot e^{\mathbb{A}_1 t_1} \\ \boldsymbol{P}(t_1 + m_a) &= \boldsymbol{P}(t_1) \cdot \mathbb{B}_1 = \boldsymbol{P}(0) \cdot e^{\mathbb{A}_1 t_1} \cdot \mathbb{B}_1, \end{aligned}$$

and for $t_1 + m_a < t \leq t_2$

$$\begin{aligned}\boldsymbol{P}(t) &= \boldsymbol{P}(t_1 + m_a) \cdot e^{\mathbb{A}_2(t-t_1-m_a)}\\ &= \boldsymbol{P}(0) \cdot e^{\mathbb{A}_1 t_1} \cdot \mathbb{B}_1 \cdot e^{\mathbb{A}_2(t-t_1-m_a)}.\end{aligned}$$

In the same way

$$\boldsymbol{P}(t_2 + m_a) = \boldsymbol{P}(0) \cdot e^{\mathbb{A}_1 t_1} \cdot \mathbb{B}_1 \cdot e^{\mathbb{A}_2(t_2-t_1-m_a)} \cdot \mathbb{B}_2.$$

And for $t_2 + m_a < t \leq t_3$

$$\boldsymbol{P}(t) = \boldsymbol{P}(0) \cdot e^{\mathbb{A}_1 t_1} \cdot \mathbb{B}_1 \cdot e^{\mathbb{A}_2(t_2-t_1-m_a)} \cdot \mathbb{B}_2 \cdot e^{\mathbb{A}_3(t-t_2-m_a)}.$$

Generalizing the formula, we get if $t_i + m_a \leq t \leq t_{i+1} \quad$ for $i \geq 2$

$$\boldsymbol{P}(t) = \boldsymbol{P}(0) \cdot e^{\mathbb{A}_1 t_1} \cdot \mathbb{B}_1 \cdot \prod_{k=2}^{i} e^{\mathbb{A}_k(t_k-t_{k-1}-m_a)} \cdot \mathbb{B}_k \cdot e^{\mathbb{A}_{i+1}(t-t_i-m_a)}.$$

A Python script for numerical computation of this formula is provided on the `book companion site`.

11.12 Piecewise Deterministic Markov Processes

In many simple situations, we may be unable to obtain an analytical formula for the system survivor function or its availability by using Markov processes and theirs extensions. As an example, consider the following case: two items are used in a parallel structure, their times-to-failure are not exponentially distributed, and they are separately maintained using preventive or corrective tasks. Even for such a simple case, we are not able to obtain the survivor function of the system by using a Markov process (the time-to-failure is not exponentially distributed), a semi-Markov process (the time spent in the current state is not enough information to calculate the transition rates for the next transitions), or a multiphase Markov process (the system does not behave as a Markov process with changes at deterministic points of time).

Such situations require the use a more generic modeling framework based on *PDMPs*. PDMPs are widely used in dynamic reliability analyses to model phenomena that are assumed to be deterministic most of the time with continuous state space (e.g. evolution of the fluid level in a vessel) and that are influenced from time to time by stochastic events with discrete state space (e.g. failures in the control loop for the fluid level). Usually, a PDMP is made of a set of differential equations (continuous part) whose solutions can experience random “jumps” (effect of discrete stochastic events). Further details are given by Davis (1984).

This book uses a specific type of PDMP, which is also known as a *piecewise linear process*: discrete and stochastic events model degradation increments and failure times, whereas continuous variables are used to model deterministic repair durations or delays, time between inspections, time spent in different states, the age of the items, and so on. Roughly speaking, the continuous part of the PDMP is not related to any physical phenomena but is used to introduce continuous variables to count time and to compensate for the lack of Markov property for the discrete part.

11.12.1 Definition of PDMP

A PDMP may be defined as

Definition 11.4 (Piecewise deterministic Markov process)
A hybrid Markov process $(X(t), \overline{m}\ (t), t \geq 0)$ where $X(t)$ is a discrete random variable or vector with values in a finite state space $\mathcal{X}$ and $\overline{m}\ (t)$ is a vector in a continuous space $\mathcal{M}$. □

In the PDMP, $X(t)$ and $\overline{m}\ (t)$ may interact with each other. $X(t)$ is used to model the discrete system states and $\overline{m}\ (t)$ is used to model time-dependent continuous variables, such as the age of the items, the repair durations, and so on. The PDMP experiences "jumps," meaning that a path of the process is described by jumps of $X(t)$ between some discrete system states in $\mathcal{X}$. We distinguish "jumps due to discrete event" when the jumps are due to a change of the system state itself (e.g. failure of one item) and "jumps due to continuous variable" when the jumps are due to a continuous variable in $\overline{m}\ (t)$ that reaches a boundary in $\mathcal{M}$ (e.g. the delay before maintenance starts is elapsed). Examples of discrete and continuous jumps are given in Section 11.12.3.

11.12.2 State Probabilities

The time-dependent state probabilities are solutions of the Kolmogorov equations, but a closed form of the solution is usually not obtainable. They are, therefore, approximated by discretization of the Kolmogorov equations. Assume that the time is discretized with step Δ. Because $\overline{m}\ (t)$ is a time vector, its components are also discretized with the same time step Δ. According to the law of total probabilities, at time $(n+1)\Delta$, the probability to be in state (x', m') may be derived with the following recursive equation:

$$P_{n+1}(x', m') = \sum_{x} \sum_{m} P_n(x, m) G_n[(x, m)(x', m')], \tag{11.110}$$

where $P_n(x', m')$ is the probability of being in state (x', m') at time $(n+1)\Delta$, $P_n(x, m)$ is the probability of being in state (x, m) at time $n\Delta$, and $G_n[(x, m)(x', m')]$ is the probability that the process moves to state (x', m') at time $(n+1)\Delta$, given that it was in state (x, m) at time $n\Delta$. Observe that m and m' are discretized with the same step Δ. Then, a numerical scheme can be built to calculate $G_n[(x, m)(x', m')]$ step-by-step and by using (11.110). If G_n and the initial state are known, everything is known.

11.12.3 A Specific Case

We study a specific case to illustrate the method. Further details and examples are provided by Arismendi et al. (2019), Cocozza-Thivent et al. (2006a,b), Lair et al. (2011), and Lin et al. (2018). Consider a redundant system with two identical items (1 and 2). A repair action is initiated as soon as an item fails and the associated downtime is d_c. In addition, each item is preventively maintained as soon as its age (time in operation) reaches a predetermined value a. For clarity, assume that the preventive maintenance (PM) duration is negligible. The time-to-failure of an item is assumed to be Weibull distributed with failure rate $\lambda(t)$. Both corrective and the PM tasks are assumed to return the item to an as-good-as-new state.

Discrete States

The process with discrete states has state space $\mathcal{X} = (0, 1, 2, 3)$, where 0 indicates zero functioning items, 1 (respectively 2) indicates that item 1 (respectively item 2) is functioning, and 3 indicates that both items are functioning. We cannot merge states 1 and 2 because we need each item's age to model the PM. At any time t, $X(t) = i$ for $i = 0, 1, 2, 3$. The PDMP experiences a (discrete) jump each time one of the item fails.

Continuous States

The process with continuous states has state space is $\mathcal{M} = ([0, d_c], [0, d_c], [0, a], [0, a])$, where a is the age when an item is preventively repaired. The value of a is a parameter that can be optimized. At any time t, $\overline{m}\ (t) = (m_1(t), m_2(t), m_3(t), m_4(t))$, where $m_1(t)$, $m_2(t)$ denote the time spent under repair for items 1 and 2 at time t, respectively, and $m_3(t)$, $m_4(t)$ denote the age of items 1 and 2 at time t. For clarity, we write: $m_1(t) = r_1(t), m_2(t) = r_2(t)$ and $m_3(t) = a_1(t), m_4(t) = a_2(t)$ in the following. Assume that if $r_i(t) = 0$, the item is not under repair.

Observe that there is no unique way to define a PDMP for a given system and a given maintenance strategy. There may be several solutions that are more or less elegant and numerically efficient, depending on the number of discrete and continuous variables that are used. The one proposed for this example may not be the most concise, but it is the most simple and intuitive.

State Probabilities

We are interested in the survivor function of the redundant system and we need to calculate its state probabilities. The system state probabilities have to be deduced from the state probabilities of the PDMP. A PDMP state is defined by a hybrid vector of discrete and continuous variables at any time t: $(X(t), \overline{m}\ (t)) = (i, r_1(t), r_2(t), a_1(t), a_2(t))$ with $i = 0, 1, 2, 3$. The PDMP state probabilities can be approximated by discretization of $\mathcal{M}$ and by using (11.110). For this purpose, we have to calculate the function G_n by writing a numerical scheme.

Numerical Scheme

The system starts in the new state so $\overline{m}\ (0) = (0, 0, 0, 0)$ and $X(0) = 3$. The function G_n is built by calculating for every discrete state the nonnull transition probabilities for the next possible discrete states, at any time step $n\Delta$. We develop here only some few cases by starting in state 3, in order to illustrate the method. The full numerical scheme is provided on the `book companion site` with a Python script.

For every $n\Delta$

- If $a_1(n\Delta) + \Delta < a$ and $a_2(n\Delta) + \Delta < a$ (if none of the items can reach the replacement age a in $[n\Delta, (n+1)\Delta]$), then only jump due to discrete events can occur and if an item failure occurs, its age is kept at its value at $n\Delta$:
  - $G_n[(3, 0, 0, a_1(n\Delta), a_2(n\Delta))(3, 0, 0, a_1(n\Delta) + \Delta, a_2(n\Delta) + \Delta)]$ $\approx [1 - \lambda(a_1(n\Delta))\Delta][1 - \lambda(a_2(n\Delta))\Delta]$.
  - $G_n[(3, 0, 0, a_1(n\Delta), a_2(n\Delta))(2, 0, 0, a_1(n\Delta), a_2(n\Delta) + \Delta)]$ $\approx \lambda(a_1(n\Delta))\Delta[1 - \lambda(a_2(n\Delta))\Delta]$.
  - $G_n[(3, 0, 0, a_1(n\Delta), a_2(n\Delta))(1, 0, 0, a_1(n\Delta) + \Delta, a_2(n\Delta))]$ $\approx [1 - \lambda(a_1(n\Delta))\Delta]\lambda(a_2(n\Delta))\Delta$.
- If $a_1(n\Delta) + \Delta \geq a$ and $a_2(n\Delta) + \Delta < a$ (if item 1 reaches the replacement age a in $[n\Delta, (n+1)\Delta]$), then a jump due to continuous variable a_1 and due to discrete event can occur.
  - If no failure occurs before the PM date of item 1, then $a_1(n\Delta)$ is put back to 0 and there is no jump of the system state: $G_n[(3, 0, 0, a_1(n\Delta), a_2(n\Delta))(3, 0, 0, 0, a_2(n\Delta) + \Delta)]$ $\approx [1 - \lambda(a_1(n\Delta))\Delta][1 - \lambda(a_2(n\Delta))\Delta]$.
  - If a failure of item 1 occurs before the PM date of item 1, then $a_1(n\Delta)$ is left to its current value, and there is a jump of system state from states 3 to 2: $G_n[(3, 0, 0, a_1(n\Delta), a_2(n\Delta))(2, 0, 0, a_1(n\Delta), a_2(n\Delta) + \Delta)]$ $\approx \lambda(a_1(n\Delta))\Delta[1 - \lambda(a_2(n\Delta))\Delta]$.
  - If a failure of item 2 occurs before the PM date of item 1, then $a_1(n\Delta)$ is put back to 0 and there is also a jump of the system state from states 3 to 1: $G_n[(3, 0, 0, a_1(n\Delta), a_2(n\Delta))(1, 0, 0, 0, a_2(n\Delta))] \approx [1 - \lambda(a_1(n\Delta))\Delta]\lambda(a_2(n\Delta))\Delta$.
- …and so on.

The same kind of calculations may be done for every case for every discrete states. The main idea is to look at the possible transitions from the current state to the other ones. Observe that when calculating the survivor function, it is required to consider that all transition rates from the failed state are zero.

A Monte Carlo simulation algorithm can be easily computed for this case, and it is provided on the `book companion site` together with the complete numerical scheme in a Python script. The reader my use them to compare results and computation times.

11.13 Simulation of a Markov Process

A Markov chain is described by a set of possible system states and a set of transitions between these states. The triggering of a transition depends on the occurrence of stochastic events. Therefore, the output of the simulation algorithm, which can be considered a system "history," consists of the sequence of the states taken by the system and by the corresponding sequence of events that have governed the transitions across these different states.

Simulating a Markov chain implies to consider the current state of the system and to treat its outgoing transitions as competitors. The (first) transition triggered by an event leads the system to a new system state. This new state is then the current state and the procedure continues.

More formally, consider a system with discrete state space $\mathcal{X}$ and let state i be its current state. Each of the outgoing transitions from state i has a constant transition rate a_{ij}, for all $j \in \mathcal{X}$. This means that the duration until the transition to state j is triggering has distribution $T_{ij} \sim \exp(a_{ij})$ (if j were the only possible transition). The "competition" between the outgoing transitions, implies that the duration $\widetilde{T}_i$ in state i is $\widetilde{T}_i = \min_{j \in \mathcal{X}}(T_{ij})$. The "winning" transition, say to state k, implies that k becomes the new current state.

The simulation of one system history may run until a predefined condition is met. This condition, depending on the needs, can, for example be

- The time the process enters into a specified state or a set of states. In that case, the system "history" may be complemented, for instance, by the time spent in the different visited states (i.e. the duration until the outgoing transition is triggered, provided that the considered state is visited a single time; if not, the durations associated with these different visits would have to be summed) or the total duration of the history (i.e. the sum of times spent in all visited states).
- A specified simulated time (i.e. a total duration), which is obtained as the sum of the durations associated to triggered transitions. In that case, only the times spent in visited states are used to complement the system "history."

Finally, several system histories are explored by Monte-Carlo simulation, meaning that the simulation algorithm of one history is run several times. The set of obtained histories is used to calculate empirical means, giving an estimate of the mean time spent in each state, the probability of being in a specific state, the mean time before reaching a given state, and so on.

An appealing property of this approach is that the transition rates can be easily changed (e.g. from exponential to Weibull distribution) without altering the general structure of the underlying code, provided that the distributions remain independent. On the other hand, the simulation approach may require complicated codes when there are, for example, a high numbers of states and/or complicated and dependent distributions associated with the transitions. Simulation of a Markov process is illustrated in Example 11.12.

Remark 11.3 (Accuracy)
The simulation approach requires extensive use of Monte-Carlo simulation to provide accurate outcomes. The number of simulated histories should be high enough to guarantee that the outcomes of the simulation are sufficiently accurate. In practice, it is required to check that the empirical means calculated on the basis of the Monte-Carlo simulation do not vary any more when passing a given number N of histories. The reader may consult Fishman (1996) for more details and theoretical framework. □

Example 11.12 (Simulating a Markov process)
Consider a system with a component A in series with two redundant components B_1 and B_2. B_1 is active and B_2 is in standby mode. The structure is shown in Figure 11.20a. Detection of a failure of component B_1, and activation of the standby component B_2 are assumed to be instantaneous. We want to estimate the MTTF of the system and the probabilities of system failure due to (i) failure of component A or (ii) failures of both components B_1 and B_2.

Because there are three components, each having two states, the system might have a state space of $2^3 = 8$ states. In practice, only a subset of the possible states may be required to model the system, depending on the structure, the assumptions

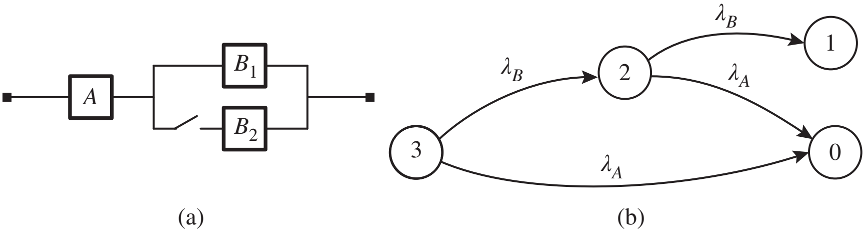


Figure 11.20 Reliability block diagram (a) and state transition diagram (b) of a two-item parallel structure with partly loaded standby and perfect switching.

and the quantities of interest. In the current example, only four states are required:

- *State 3* (A, B_1, B_2). All components are in a functioning state, which is the initial state where the system is as-good-as-new.
- *State 2* $(A,\ \overline{B}_1,\ B_2)$. All components are running except for component B_1. This state may be considered a degraded state of the system.
- *State 1* $(A,\ \overline{B}_1,\ \overline{B}_2)$. Components B_1 and B_2 are failed, such that the system is failed. This is one of the two failed states that has to be studied in this example.
- *State 0*. Comprising the system states $(\overline{A},\ B_1,\ B_2)$ and $(\overline{A},\ \overline{B}_1,\ B_2)$. These two system states may be *merged* because they both imply that the system is in a failed state caused by failure of component A.

Other system states, such as $(A,\ B_1,\ \overline{B}_2)$, $(\overline{A},\ \overline{B}_1,\ \overline{B}_2)$, are, in this example, not considered because they correspond to unreachable system states.

Assume that all the three components have constant failure rates: λ_A for component A and $\lambda_{B_1} = \lambda_{B_2} = \lambda_B$ for components B_1 and B_2. The corresponding state transition diagram is shown in Figure 11.20b. For the exponential distribution, we have that $\Pr(T_{ij} > t + s \mid T_{ij} > s) = \Pr(T_{ij} > t)$ for all $t, s \geq 0$. Therefore, the duration T_{20} associated to the transition from states 2 to 0 does not depend on the duration of the transition T_{32}, that previously triggered the transition to state 2. T_{20} can be drawn directly from the exponential distribution parameter λ_A at the arrival date in state 2. The memoryless property simplifies the management of the duration associated to the concurrent transitions.

Hence, the transitions from states 3 to 0, and from states 2 to 0, have constant rate λ_A; transitions from states 3 to 2, and from states 2 to 1, have constant rate λ_B.

The first pseudocode[6] (GetOneHistory in Figure 11.21) simulates a single history of the system and provides a single observation of the times-to-failure and a single observation of the final system state. This pseudocode may form the basis for a Monte-Carlo simulation that provides an estimate of the system MTTF and states probabilities. A basic pseudocode of the Monte-Carlo simulation for N histories is proposed in Figure 11.22. □

An implementation of this pseudocode in Python is provided on the `book companion site`. When the system is not too large and/or too complicated, it is rather straightforward to extend the pseudocode to use other time-to-failure distributions that the exponential. Reconsider the system in Example 11.12; if component A had a Weibull time-to-failure distribution instead of an exponential

[6] A *pseudocode* is an informal high-level description of the operating principle of a computer program.

```
procedure GetOneHistory($\lambda_A, \lambda_B$)
    $ttf \leftarrow 0$                                                ▷ Time to failure initialization
    $state \leftarrow 3$                                              ▷ Initial state
    $\lambda_{30} \leftarrow \lambda_A$                               ▷ Change of notations
    $\lambda_{32} \leftarrow \lambda_B$
    $\lambda_{20} \leftarrow \lambda_A$
    $\lambda_{21} \leftarrow \lambda_B$
    while $state \neq 1$ and $state \neq 0$ do                        ▷ Loop while any of the final states is reached
        if $state = 3$ then                                           ▷ If current state is 3
            $t_{30} \leftarrow$ draw.exp($\lambda_{30}$)             ▷ Draw duration until component $A$ failure
            $t_{32} \leftarrow$ draw.exp($\lambda_{32}$)             ▷ Draw duration until component $B_1$ failure
            if $t_{30} \leq t_{32}$ then                              ▷ If next event is component $A$ failure
                $state \leftarrow 0$                                  ▷ Update the system state
                $ttf \leftarrow ttf + t_{30}$                         ▷ Update the time to failure
            else                                                      ▷ If next event is component $B_1$ failure
                $state \leftarrow 2$                                  ▷ Update the system state
                $ttf \leftarrow ttf + t_{32}$                         ▷ Update the time to failure
            end if
        else                                                          ▷ If current state is state 2
            $t_{20} \leftarrow$ draw.exp($\lambda_{20}$) ▷ Draw duration until component $A$ failure: *property of exponential law*
            $t_{21} \leftarrow$ draw.exp($\lambda_{21}$)             ▷ Draw duration until component $B_2$ failure
            if $t_{20} \leq t_{21}$ then                              ▷ If next event is component $A$ failure
                $state \leftarrow 0$                                  ▷ Update the system state
                $ttf \leftarrow ttf + t_{20}$                         ▷ Update the time to failure
            else                                                      ▷ If next event is component $B_2$ failure
                $state \leftarrow 1$                                  ▷ Update the system state
                $ttf \leftarrow ttf + t_{21}$                         ▷ Update the time to failure
            end if
        end if
    end while
    return $ttf, state$                                               ▷ Return time to failure $ttf$ and final state $state$
end procedure
```

Figure 11.21 Example of a Markov's chain simulation – single history.

```
procedure SystemMonteCarlo($N, \lambda_A, \lambda_B$)
    $mttf \leftarrow 0$                                               ▷ Initialize variables
    $state0 \leftarrow 0$
    for $i \leftarrow 1, N$ do                                        ▷ Loop on $N$ histories
        $(ttf, state) \leftarrow$ GetOneHistory($\lambda_A, \lambda_B$)   ▷ Get output of a single history
        $mttf \leftarrow mttf + ttf$                                  ▷ Sum time to failure
        if $state = 0$ then
            $state0 \leftarrow state0 + 1$                            ▷ Sum histories ending on state 0
        end if
    end for
    $mttf \leftarrow mttf/N$                                          ▷ Estimate the MTTF of the system
    $state0 \leftarrow state0/N$                                      ▷ Estimate probability system ends on state 0
    $state1 \leftarrow 1 - state0$                                    ▷ Estimate probability system ends on state 1
    return $mttf, state0, state1$                                     ▷ Return MTTF and probabilities estimations
end procedure
```

Figure 11.22 Example of a Markov's chain simulation – estimate of MTTF and states probabilities.

distribution, the pseudocode needs to be adapted in the following way: On the arrival in state 0, the duration t_{03} must be drawn from the correct distribution. The same value must be considered to select the next event at the transition from state 1. This change of assumptions is seen to be managed in a much easier way by simulation than in analytical approaches.

Remark 11.4 (Markov analysis with R)
Markov analysis and Monte Carlo simulation can also be accomplished with R. Several R packages are available. Among these are

- `markovchain` is a general package for discrete time Markov analysis that also includes modules for (continuous time) Markov processes.
- `mcmc` and `mcmcr` implement the Monte Carlo Markov chain approach.
- `mstate` fits multistate models based on Markov chains for survival analysis.
- `simmer` is a package for discrete event simulation and a valuable tool for Markov process simulation. `simmer` is a parallel to the simulation package `simPy` for the Python language.

Further information about these – and several other – packages may be found by visiting `https://cran.r-project.org`. □

11.14 Problems

11.1 Consider an item that is subject to two types of repair. Initially, the item has a constant failure rate λ_1. When the item fails for the first time, a partial repair is performed to restore the item to the functioning state. This partial repair is not perfect, and the failure rate λ_2 after this partial repair is therefore higher than λ_1. After the item fails the second time, a thorough repair is performed that restores the item to an as-good-as-new condition. The third repair is a partial repair, and so on. Let μ_1 denote the constant repair rate for a partial repair, and μ_2 be the constant repair rate of a complete repair ($\mu_1 > \mu_2$). Assume that the item is put into operation at time $t = 0$ in an as-good-as-new condition.

(a) Establish the state transition diagram and the state equations for this process.
(b) Determine the steady state probabilities of the various states.

11.2 Consider a parallel structure of three independent and identical components with failure rate λ and repair rate μ. The components are repaired independently. All the three components are assumed to be functioning at time $t = 0$.

(a) Establish the state transition diagram and the state equations for the parallel structure.

(b) Show that the mean time to the first system failure MTTF is

$$\text{MTTF} = \frac{1}{3\lambda} + \frac{1}{\mu + 2\lambda} + \frac{1}{2\mu + \lambda}.$$

11.3 Consider a parallel structure of four independent and identical components with failure rate λ and repair rate μ. The components are repaired independently. All the four components are assumed to be functioning at time $t = 0$.

(a) Establish the state transition diagram and the state equations for the parallel structure.

(b) Determine the mean time to the first system failure MTTF.

(c) Is it possible to find a general formula for a parallel structure of n components?

11.4 Consider the pitch system represented by the simplified reliability block diagram (RBD) in Figure 11.23. The accumulators are identical and in a parallel structure. The main line of the pitch system is active when the system is started. When the main line of the pitch system fails, the pitch system emergency line takes over. The switch is perfect (no switch failure, when the main line is failed, the emergency line takes over). The failure rates of all the items are constant, denoted λ_1 for the hydraulic cylinder, λ_2 for the two pitch systems (the failure rates are identical for main line and emergency line), λ_3 for the accumulators (the failure rates are identical for the two accumulators), λ_4 for the pump and λ_5 for the filter. Given that this system is embedded in an off-shore wind turbine, we consider that the repair rate is the same for all the items, and mainly due to the time to prepare and go to the spot. The repair rate is assumed to be a constant μ

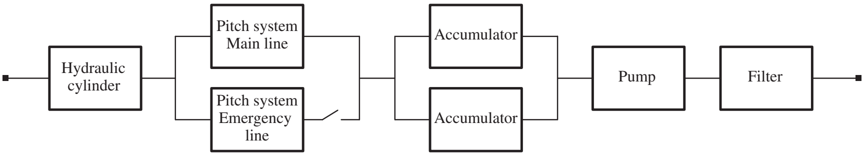


Figure 11.23 RBD for the pitch system in Problem 11.4.

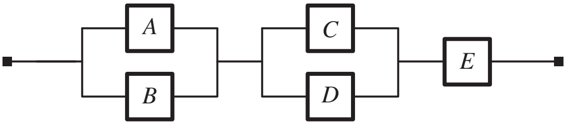


Figure 11.24 RBD for the system in Problem 11.5.

for each item. Consider that one repair team is available for every item at any time.

(a) Define the possible system states and establish a state transition diagram for the system. You may assume that the system is stopped as soon as it is failed (no item will fail after the system has failed).

(b) Establish the transition rate matrix $\mathbb{A}$ for the pitch system.

11.5 Consider the system described by the RBD in Figure 11.24. Items A and B are redundant and have the same constant failure rate λ_1. Items C and D are redundant and have the same constant failure rate λ_2. Item E has a constant failure rate λ_3. When item A (or B) fails, a repair is initiated with a constant repair rate μ_1. While $A(B)$ is under repair, the surviving item $B(A)$ experiences extra load and its failure rate is increased to the constant value $\overline{\lambda}_1$. The same applies to C and D. The failure rate of the surviving item is $\overline{\lambda}_2$ and the repair rate of the failed item is μ_2. When the whole system is failed, a renewal is initiated and the system is put back to the as-good-as-new state with a constant repair rate μ. Let S_p denote the following set of parameters: $S_p = \{\lambda_1, \lambda_3, \ \overline{\lambda}_1, \ \mu_1, \mu\}$.

(a) Consider the system with items A, B, C, D, and E.
   i. Explain why this system can be modeled by a Markov process.
   ii. List the possible states of the system and establish a state transition diagram with as few states as possible.

(b) Now, remove items C and D from the system.
   i. List the possible states of the system and establish the state transition diagram with as few states as possible.
   ii. Establish the transition matrix and the state equations.
   iii. Explain what is meant by steady state and derive the steady state probabilities.
   iv. Calculate the steady state availability of the system expressed by the parameters in S_p.
   v. Calculate the mean number of system failures per hour expressed by the parameters in S_p.
   vi. Calculate the mean number of repairs (any repair) in one year expressed by the parameters in S_p.
   vii. Calculate the mean number of renewals in one year expressed by the parameters in S_p.

viii. Explain the procedure to calculate the system survivor function (without doing the calculations).

11.6 A fail-safe valve has two main failure modes: premature closure (PC)/ spurious closure (SC), and fail to close (FTC), with constant failure rates

$$\lambda_{\text{PC}} = 10^{-3} \qquad \text{PC-failures/h}$$
$$\lambda_{\text{FTC}} = 2 \times 10^{-4} \qquad \text{FTC-failures/h.}$$

The MTTR a PC failure is assumed to be one hour, whereas the MTTR an FTC failure is 24 hours. The repair times are assumed to be exponentially distributed.

(a) Explain why the operation of the valve may be described by a Markov process with three states. Establish the state transition diagram and the state equations for this process.

(b) Calculate the average availability of the valve, and the mean time between failures.

11.7 A production system has two identical channels and is running 24 hours a day all days. Each channel can have three different states, representing 100%, 50%, and 0% capacity, respectively. The failure rate of a channel operating with 100% capacity is assumed to be constant $\lambda_{100} = 2.4 \times 10^{-4}$ h^{-1}. When a failure occurs, the capacity will go to 50% with probability 60% and to 0% capacity with probability 40%. When a channel is operated with 50% capacity, it may fail (and go to 0% capacity) with constant failure rate $\lambda_{100} = 1.8 \times 10^{-3}$ h^{-1}. The system is further exposed to external shocks that will take down the system irrespective of the state it is in. The rate of these shocks is $\lambda_s = 5 \times 10^{-6}$ h^{-1}. The two channels are assumed to operate and fail independent of each other. When both channels have capacity 50% or less, the whole system is closed down, and it is not started up again until both channels have been repaired to an as-good-as-new state. When a channel enters 50% capacity, a repair action is “planned” and then carried out. The planning time includes bringing in spare parts and repair teams. The planning time is 30 hours in which case the channel continues to operate with 50% capacity. The active repair time is so short that it can be neglected. When a channel enters 0% capacity (and the other channel is operating with 100% capacity), the planning time is compressed to 20 hours and the active repair time is still negligible. After a system stop, the mean time to bring the system back to operation is 48 hours, irrespective of state of the system when it entered the idle state. Record any additional assumptions you have to make to answer the questions below.

(a) Define the relevant system states. Use as few states as possible.
(b) Draw the corresponding state transition diagram.
(c) Establish the transition rate matrix $\mathbb{A}$ for the production system.
(d) Establish the Markov steady-state equations on matrix form.
(e) Explain (briefly) what we mean by the concept steady-state probability in this case.
(f) Find the steady-state probability of the production system.
(g) The net income of a channel running at 100% capacity is €500 h^{-1}. The net income of a channel running at 50% capacity is €200 h^{-1}. The cost of a repair (50% → 100% capacity) is €5500. The cost of a repair (0% → 100% capacity) is €10 500. The cost of a system repair from idle to full functioning state (including penalty because of no production) is €280 h^{-1}.
   i. Find the average income per year from operating the system.
   ii. Find the mean time from startup until the first failure.

11.8 A pumping system has three pumps of the same type. Each pump can supply 50% of the required capacity. In normal operation, two pumps are running, whereas the third pump is in standby. When a pump is running, it has a (total) constant failure rate such that MTTF = 550 hours. When one of the active pumps fails, the standby pump is activated, and a repair action of the failed pump is initiated. The switching operation is assumed to take place without any problems. Assume that a pump will not fail in standby mode. The company has only one repair team, and only one pump can therefore be repaired at each time. The mean repair time of a pump is 10 hours. Assume that common cause failures may occur for the active pumps, but that this type of failure will not affect the standby pump. Further, assume that common cause failures may be modeled by a beta-factor model with $\beta = 0.12$. Common cause failures may be regarded as external shocks that will affect all active pumps, regardless of how many pumps that are in operation. Record any additional assumptions you have to make to answer the questions below.
(a) Define the relevant system states. Use as few states as possible.
(b) Draw the corresponding state transition diagram.
(c) Establish the transition rate matrix $\mathbb{A}$ for the pumping system.
(d) Establish the steady-state equations on matrix form.
(e) Explain (briefly) what we mean by the concept steady-state probability in this case.
(f) Find the steady-state probabilities for the pumping system.
(g) The pumping system has a system failure when none of the pumps are in operation. Find the mean time to the first system failure, MTTF,

when the pumping system starts out with two pumps in operation and one pump in standby at time $t = 0$.

11.9 The heater system of a steam producing plant has three identical burners. Only two of the three burners are in use, whereas the third is in standby. Each burner has constant failure rate $\lambda = 2.5 \times 10^{-3}\ \text{h}^{-1}$. When one of the active burners fails, the standby burner is activated, and a repair action of the failed burner is initiated. The probability of a successful activation of a standby burner is assumed to be 98%. The company has only one repair team, and only one burner can therefore be repaired at each time. The mean repair time of a failed burner is two hours. The same repair time also applies for a burner that has "failed to start." The likelihood of common cause failures is considered to be negligible. Record any additional assumptions you have to make to answer the questions below.

(a) Define the relevant system states. Use as few states as possible.
(b) Draw the corresponding state transition diagram.
(c) Establish the transition rate matrix $\mathbb{A}$ for the burner system.
(d) Establish the steady-state equations on matrix form.
(e) Explain (briefly) what we mean by the concept steady-state probability in this case.
(f) Find the steady-state probabilities for the burner system.
(g) The steam production plant fails when no burner is active, and only one burner is active and none of the other burners can be activated within 30 minutes. Find the mean time to the first system failure, MTTF, when the burner system starts out with two burners in operation and one burner in standby at time $t = 0$.

11.10 The degradation of an item can be discretized according to four levels (level 1 is new, level 4 has failed) and the degradation level is known only at periodic inspection dates (period τ). Maintenance tasks can be performed only at inspection dates and their duration is negligible. The transition rates between the four degradation levels are all constant and equal to $10^{-4}\ \text{h}^{-1}$. The corrective and PM tasks bring the item to an as-good-as-new state and a PM task is performed when the item is found in the degradation level 2 or 3 at the inspection date.

(a) Define the relevant system states between two inspections. Use as few states as possible.
(b) Draw the corresponding state transition diagram between two inspections.
(c) Establish the transition rate matrix $\mathbb{A}$ for the item between two inspections.

(d) By using a multiphase Markov process, calculate the unavailability of the item (probability to be in state 4) at any time with and without PM task. Make a plot of it.

(e) Consider now that the monitoring is not perfect at inspection times: there is a probability 0.9 that the unit is diagnosed as being in state 2 or 3 when it is actually in state 2 or 3, and a probability 0.1 that the unit is diagnosed as being in state 1 (new state) when it is actually in state 2 or 3. Modify your previous model to calculate the item availability at any time between two inspections with PM task.

11.11 Consider a redundant system with two identical items (1 and 2). A repair action is initiated as soon as an item fails and the associated downtime is $d_c = 1000$ hours. In addition, each item is preventively maintained as soon as its age (time in operation) reaches a predetermined value $a = 7500$ hours. The downtime due to PM task is $d_p = 500$ hours. The time-to-failure of an item is assumed to be Weibull distributed with parameters $\alpha = 2.25$ and $\theta = 1 \times 10^4$ hours. Both corrective and the PM tasks are assumed to return the item to an as-good-as-new state.

- Define the relevant system states between two inspections. Use as few states as possible.
- Draw the corresponding state transition diagram.
- Establish the transition rate matrix $\mathbb{A}$.
- Use the PDMP proposed in Section 11.12.3 and modify it to take into account PM tasks duration.
- Calculate the surviving function and the availability of the system at any time.

11.12 Consider the system described in Example 11.12 and assume that the items can be repaired upon failure with a constant repair rate μ_A for item A and μ_B for items B_1 and B_2. There are two repair teams available at any time. Numerical values are: $\lambda_A = 10^{-4}\ \text{h}^{-1}$, $\lambda_B = 5 \times 10^{-3}\ \text{h}^{-1}$, $\mu_A = \mu_B = 10^{-1}\ \text{h}^{-1}$. The surviving items are stopped when the whole system is failed and cannot fail anymore.

- Define the relevant system states. Use as few states as possible.
- Draw the corresponding state transition diagram.
- Establish the transition rate matrix $\mathbb{A}$.
- Use the pseudocode proposed in Figures 11.21 and 11.22 and modify it to take into account corrective maintenance.
- Write a script to implement it (Python or R).
- Calculate the MTTF and the MTBF of the system.

- Modify the script to consider that the repair durations are constant and equal to 100 hours. Calculate the MTTF and the MTBF of the system and compare to the previous results.

References

Arismendi, R., Barros, A., Vatn, J., and Grall, A. (2019). Prognostics and maintenance optimization in bridge management. *Proceedings of the 29th European Safety and Reliability Conference(ESREL)*, Hannover, Germany, pp. 653–661.

Barlow, R.E. and Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing, Probability Models*. New York: Holt, Rinehart, and Winston.

Billington, R. and Allen, R.N. (1992). *Reliability Evaluation of Engineering systems: Concepts and Techniques*, 2e. New York: Springer.

Bon, J.L. (1995). *Fiabilité des Systèmes, Méthodes Mathématiques*. Paris: Masson.

Cocozza-Thivent, C. (1997). *Processus Stochastiques et Fiabilité des Systèmes (in French)*. Paris: Springer.

Cocozza-Thivent, C., Eymard, R., and Mercier, S. (2006a). A finite-volume scheme for dynamic reliability models. *IMA Journal of Numerical Analysis* 26 (3): 446–471.

Cocozza-Thivent, C., Eymard, R., Mercier, S., and Roussignol, M. (2006b). Characterization of the marginal distributions of Markov processes used in dynamic reliability. *International Journal of Stochastic Analysis* 26 (3): 1–18.

Cox, D.R. and Miller, H.D. (1965). *The Theory of Stochastic Processes*. London: Methuen & Co..

Davis, M.H.A. (1984). Piecewise-deterministic Markov processes: a general class of non-diffusion stochastic models. *Journal of the Royal Statistical Society: Series B (Methodological)* 46 (3): 353–388.

Fishman, G. (1996). *Monte Carlo; Concepts, Algorithms, and Applications*. New York: Springer.

Grabski, F. (2015). *Semi-Markov Processes: Applications in System Reliability and Maintenance*. Amsterdam: Elsevier.

Lair, W., Mercier, S., Roussignol, M., and Ziani, R. (2011). Piecewise deterministic Markov processes and maintenance modeling: application to maintenance of a train air-conditioning system. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability* 225 (2): 199–209.

Limnios, N. and Oprisan, G. (2001). *Semi-Markov Processes and Reliability*. Basel: Birkhäuser.

Lin, Y.H., Li, Y.F., and Zio, E. (2018). A comparison between Monte Carlo simulation and finite-volume scheme for reliability assessment of multi-state physics systems. *Reliability Engineering & System Safety* 174: 1–11.

Pagès, A. and Gondran, M. (1980). *Fiabilité des Systèmes*. Paris: Eyrolles.

Pukite, J. and Pukite, P. (1998). *Modeling for Reliability Analysis; Markov Modeling for Reliability, Maintainability, Safety, and Supportability Analyses of Complex Computer Systems*. New York: IEEE Press.

Ross, S.M. (1996). *Stochastic Processes*. New York: Wiley.

Trivedi, K.S. and Bobbio, A. (2017). *Reliability and Availability Engineering: Modeling, Analysis, and Applications*. Cambridge: Cambridge University Press.

12

Preventive Maintenance

12.1 Introduction

Preventive maintenance (PM) is introduced briefly in Chapter 9 and is treated in more detail in the current chapter. A PM task is triggered by the age of the item, the calendar time, or the condition of a functioning item. The main challenge of a PM strategy is to decide *what* should be done, *how* thoroughly it should be done, and *when* it should be done to *prevent* item or system failure at the lowest possible long-term *cost*. The term cost is used here with a wide interpretation and may cover production losses, risk to personnel, and pollution of the environment. The cost related to a PM task is sometimes referred to as the *objective function* for the task. Section 9.3 defines PM as maintenance "carried out to mitigate degradation and reduce the probability of failure" (IEV 192-06-05). A slightly different definition is partly based on ISO 14224 (2016):

Definition 12.1 (Preventive maintenance)
Maintenance carried out at predetermined intervals or according to prescribed criteria and intended to reduce the probability of failure or functional degradation of an item. □

A PM task of an item is relevant when (i) the failure rate of the item is increasing and (ii) the cost associated to the PM task is lower than the cost of a CM task that is carried out after an item failure has occurred. When the type and thoroughness of a PM task has been decided, the time to perform the task has to be selected, based on time or item condition. There is typically an optimal time to perform the task that gives a minimal long-term cost. If the task is performed too early or too late, the long-term cost will be higher than the minimal cost. The same applies to the thoroughness of the task. A too brief and a too careful PM task inevitably gives an increased long-term cost. The carefulness of a maintenance task is discussed and classified in Chapter 9.

System Reliability Theory: Models, Statistical Methods, and Applications, Third Edition.
Marvin Rausand, Anne Barros, and Arnljot Høyland.

Companion Website: www.wiley.com/go/SystemReliabilityTheory3e

To select the best PM task and the optimal time to perform this task, we need to use various types of maintenance models. Some of the most relevant models and approaches are presented and discussed in this chapter. The first category of commonly used PM strategies is called time-based PM strategies because the time is the sole decision variable. It can be the calendar time (block replacement strategy) or the time in operation (age-based strategy). The asymptotic cost per time unit is derived and the added value of such strategies is discussed based on several numerical and practical examples.

Next, degradation models are introduced together with the concept of *remaining useful lifetime* (RUL). This is a preliminary step to introduce *condition-based maintenance* (CBM) strategies which also provide direct inputs for *prognostics*. A brief review of the most commonly used degradation models is presented with some approaches to calculate the probability distribution of RUL. Numerical examples, simulations of the degradation paths, and of the probability density function of RUL are provided on the `book companion site` and we strongly recommend the reader to study these to better understand the models.

The most common CBM strategies are reviewed and classified according to the nature of the condition monitoring information (continuous monitoring or inspections) and the nature of the degradation models used for the item (discrete state space or continuous one). For each class, a suitable modeling framework is proposed to evaluate the cost function in some simple but realistic and representative cases. The review is far from exhaustive, but provides relevant and significant inputs to start a modeling work in a wide range of situations.

In Section 12.6, the system is not considered any more as a single item or as a black box but as a collection of several items that are put together to fulfill a main function. A brief review of the modeling challenges and approaches for multi-item systems is presented. Then a generic and rather complete example is studied in detail.

Throughout this chapter, examples are provided to illustrate the concepts, the analyses, and the modeling work. We strongly rely on (i) renewal and counting processes (Chapter 10) and (ii) Markov processes and their extensions (Chapter 11). For each example, a corresponding Python script is available on the `book companion site`. These can be used to simulate the maintained item or system and assess the maintenance strategies by Monte Carlo simulation.

12.2 Terminology and Cost Function

The models and the analyses are based on a set of terms that we define in a decision-theoretic framework.

Maintenance task. A specific task a to maintain an item determined by specifying "what, where, how, and when." Some authors prefer to call this task a *maintenance action*, but we prefer to use the term task. The set of all possible and realistic tasks (or actions) is called the maintenance *task space* $\mathcal{A}$. When the task space is discrete, we write $\mathcal{A} = \{a_1, a_2, a_3, \dots\}$.

Maintenance decision. A process δ, based on the actual operating context, costs, knowledge, and available data $\mathcal{D}$, which is used to select a specific maintenance task $a_i \in \mathcal{A}$. This process can be represented as a function $\delta : \mathcal{D} \to a_i$.

Maintenance strategy. An overall framework that describes how the maintenance decision problem shall be approached and how specific decisions shall be made when actual input information/data in $\mathcal{D}$ is available. Observe that a decision is made based on a given dataset $\mathcal{D}$, whereas a strategy tells how we should approach the decision problem for any dataset $\mathcal{D}$. The strategy must embrace an objective function, a utility function, or a loss function. This function $C(\cdot, t)$ may have one or more dimensions and is defined as a cumulative function on a time horizon $[0, t]$. When $C(\cdot, t)$ has more than one dimension, we say that we face a decision problem with multiple objectives. This book is delimited to single objectives and this objective is called *cost*. The function $C(\cdot, t)$ is called the cost function and is defined below.

Cost Function. The cost function of a maintenance strategy is a function of at least the following items:

(a) A specified set of maintenance tasks $\mathcal{A}$

(b) A specified process δ

(c) The actual operating context $\mathcal{D}_{\text{oc}}$ describing the state of the system, which components are failed, the actual production/operational requirements, and so on

(d) The calendar time t_{cal}, because the cost may depend on whether the task is to be done within or outside normal working hours, the time of the year, and so on

(e) …maybe several more items

If we assume that $\mathcal{D}_{\text{oc}}$ and t_{cal} are specified as part of the decision problem, we may simplify the notation and write the cost function as $C(a, \delta, t)$. The final objective is then to choose the set of relevant maintenance tasks $\mathcal{A}$ and the decision process δ that minimizes the cost $C(a, \delta, t)$. This book focuses on the modeling phase and not on the optimization phase. We demonstrate how reliability models can be used to obtain $C(a, \delta, t)$ for given a and δ. You may use the same routine for different a and δ to choose the best maintenance task. For brevity, the cost function is often denoted $C(t)$, meaning implicitly that a and δ are fixed and known.

In practice, the real cost function $C(t)$ is a random and time-dependent variable. For optimization purposes, it is replaced by its mean value or by the *asymptotic* *cost per time unit*, C_∞, defined as follows:

$$C_\infty = \lim_{t\to\infty} \frac{C(t)}{t}.$$

The use of C_∞ makes sense when the item behavior and the maintenance tasks are studied on the very long term.

If the maintained item experiences renewals (it is put back into its initial state) at random or deterministic times, then the model of the maintained item is a renewal process and the cost function can be derived by using analytical tools related to renewal processes (see Chapter 10) and especially the *renewal theorem*:

$$C_\infty = \lim_{t\to\infty} \frac{C(t)}{t} = \frac{E[C(T_R)]}{E(T_R)}, \tag{12.1}$$

where $C(t)$ is the maintenance cost accumulated from 0 to t for a maintenance strategy, T_R is the renewal cycle, that is, the time elapsing between two renewals, and $C(T_R)$ is the maintenance cost accumulated in a renewal cycle. In the following, T_R is also called the *replacement interval.*

This means that C_∞ can be calculated by considering a single renewal cycle. In most cases, a closed form of C_∞ is not obtainable, and approximation or numerical calculation tools must be used. Another option is to use Monte Carlo discrete event simulation (see Chapter 6) and approximate the mean cost per time unit by:

$$C_\infty = \frac{E[C(T_R)]}{E(T_R)} = \lim_{n\to\infty} \frac{\frac{1}{n}\sum_{k=1}^{n} C(t_R^k)}{\frac{1}{n}\sum_{k=1}^{n} t_R^k} \simeq \frac{\sum_{k=1}^{n_s} C(t_R^k)}{\sum_{k=1}^{n_s} t_R^k}, \tag{12.2}$$

where t_R^k is the length of the k^{th} simulated renewal cycle, $C(t_R^k)$ is the maintenance cost in the kth renewal cycle, and n_s is the number of simulated renewal cycles (which must be high enough to guarantee the approximation quality).

In some cases, a and δ can be reduced to a set of parameters, in which case the cost function C_∞ can be written as a function of the parameters to be optimized $C_\infty(\cdot)$.

12.3 Time-Based Preventive Maintenance

In this section, all PM tasks are assumed to be replacements, such that the item is as-good-as-new after a PM task is completed. It is further assumed that the PM tasks are specified such that the only decision variable is *when* the task is to be carried out. The time may be measured as time in operation or calendar time. Consider two distinct situations:[1]

[1] Professor **Bruno Castanier**, Université d'Angers, made significant contributions to this section.

(1) The item is planned to be preventively replaced after a specified time in operation (i.e. at a specified operational age). If a failure occurs first, the item is replaced at the failure date, the preventive replacement is cancelled, and rescheduled starting from the replacement date. This strategy is called *age replacement*, and its advantage is to avoid preventive replacements of items that have recently been (correctively) replaced. The disadvantage is that the time of the next PM cannot be known in advance because it depends on failure occurrences.

(2) The items are preventively replaced at fixed dates, even if failures occur in-between. This strategy is called *block replacement*, and its advantage is that the times of PM tasks are known in advance. They are usually periodic. This strategy is, for example, relevant when the cost of having available repair teams is high. The disadvantage is that we may preventively replace items that have recently been replaced.

12.3.1 Age Replacement

Under an *age replacement* strategy, an item is replaced upon failure or at a specified operational age t_0, whichever comes first. This strategy makes sense when the replacement cost upon failure is higher than the cost of a planned replacement, and when the failure rate of the item is increasing.

Consider a process where the item is subject to age replacement at age t_0, which is nonrandom. Let T be the (potential) time-to-failure of the item. T is assumed to be continuous with distribution function $F(t)$, density $f(t)$, and mean time-to-failure (MTTF). The time required to replace the failed item is considered to be negligible, and after replacement, the item is assumed to be as-good-as-new. The time between two consecutive replacements is called a *replacement period* or a renewal cycle T_R. The replacement period may be expressed as $T_R = \min(t_0, T)$. The mean length of a replacement period is

$$E(T_R) = \int_0^{t_0} t\,f(t)\,dt + t_0 \Pr(T \geq t_0) = \int_0^{t_0} [1 - F(t)]\,dt. \tag{12.3}$$

Some authors use the term *mean time between replacements* (MTBRs) instead of $E(T_R)$. Observe that $E(T_R)$ is always less than t_0 and that $\lim_{t_0 \to \infty} E(T_R) = \text{MTTF}$. The mean number of replacements, $E[N(t)]$ in a long time interval of length t is therefore approximately

$$E[N(t)] \approx \frac{t}{E(T_R)} = \frac{t}{\int_0^{t_0} [1 - F(t)]\,dt}. \tag{12.4}$$

Including a repair duration is accomplished by adding its mean value to $E(T_R)$.

Let c be the cost of a preventive replacement when the item has reached the age t_0, and let $c + k$ be the cost of replacing a failed item (before age t_0). The cost c

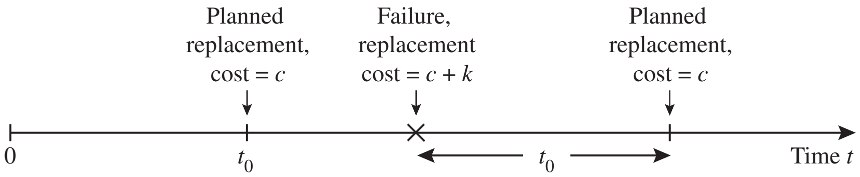

Planned replacement, cost = c
Failure, replacement cost = c + k
Planned replacement, cost = c
0
t_0
t_0
Time t

Figure 12.1 Age replacement strategy and costs.

covers the hardware and man-hour costs, whereas k is the extra cost incurred by the unplanned replacement, such as production loss and extra mobilization cost for the repair team. The costs are illustrated in Figure 12.1.

By the age replacement strategy, the replacement times cannot be fully scheduled, and the strategy may therefore be complicated to manage when having a high number of items. The age of each item has to be monitored, and the replacement tasks will be spread out in time.

The total cost per replacement period with replacement age t_0 is equal to the replacement cost c plus the extra cost k whenever a failure occurs. The mean total cost per replacement period is

$$E[C(T_R)] = c + k\,\Pr(\text{“failure”}) = c + k\,\Pr(T < t_0) = c + kF(t_0). \tag{12.5}$$

The asymptotic cost per time unit, C_∞, with replacement age t_0 is denoted $C_\infty(t_0)$ because it depends on the replacement age t_0 and is determined by

$$C_\infty(t_0) = \frac{E[C(T_R)]}{E(T_R)} = \frac{c + kF(t_0)}{\int_0^{t_0}[1 - F(t)]\,dt}. \tag{12.6}$$

The objective is now to determine the replacement age t_0 that minimizes $C_\infty(t_0)$. An approach to finding the optimal t_0 is shown in Example 12.1.

When $t_0 \to \infty$, (12.6) becomes

$$C_\infty(\infty) = \lim_{t_0\to\infty} C_\infty(t_0) = \frac{c + k}{\int_0^{\infty}[1 - F(t)]\,dt} = \frac{c + k}{\text{MTTF}}. \tag{12.7}$$

Observe that $t_0 \to \infty$ means that no age replacement takes place. All replacements are corrective replacements and the cost of each replacement is $c + k$. The time between replacements is MTTF, and (12.7) is therefore an obvious result. The ratio

$$\frac{C_\infty(t_0)}{C_\infty(\infty)} = \frac{c + kF(t_0)}{\int_0^{t_0}[1 - F(t)]\,dt}\,\frac{\text{MTTF}}{c + k} = \frac{1 + rF(t_0)}{\int_0^{t_0}[1 - F(t)]\,dt}\,\frac{\text{MTTF}}{1 + r}, \tag{12.8}$$

where $r = k/c$, may be used as a measure of the cost efficiency of the age replacement strategy with replacement interval t_0. A low value of $C_\infty(t_0)/C_\infty(\infty)$ indicates a high cost efficiency.

Example 12.1 (Age replacement–Weibull distribution)
Consider an item with Weibull time-to-failure distribution $F(t)$ with scale parameter θ and shape parameter α. To find the optimal replacement age t_0, we have to find the replacement age t_0 that minimizes (12.7), or alternatively (12.8). Using (12.8) yields

$$\frac{C_\infty(t_0)}{C_\infty(\infty)} = \frac{1 + r(1 - e^{-(t_0/\theta)^\alpha})}{\int_0^{t_0} e^{-(t/\theta)^\alpha}\,dt}\,\frac{\theta\Gamma(1/\alpha + 1)}{1 + r}. \tag{12.9}$$

By introducing $x_0 = t_0/\theta$, (12.9) may be written as

$$\frac{C^*_\infty(x_0)}{C_\infty(\infty)} = \frac{1 + r(1 - e^{-x_0^\alpha})}{\int_0^{x_0} e^{-x^\alpha}\,dx}\,\frac{\Gamma(1/\alpha + 1)}{1 + r}, \tag{12.10}$$

where $C^*(\cdot)$ is the cost function obtained by the transform $x_0 = t_0/\theta$. To find the x_0 for which (12.10) attains its minimum by analytical methods is not straightforward. The optimal x_0 may be found graphically by plotting $C^*_\infty(x_0)/C_\infty(\infty)$ as a function of x_0. An example is shown in Figure 12.2 where $C^*_\infty(x_0)/C_\infty(\infty)$ is plotted for $\alpha = 3$, and some selected values of $r = k/c$.

The optimal x_0, and thereby the optimal replacement age $t_0 = \theta x_0$, can be found from Figure 12.2 as the value minimizing the ratio $C^*_\infty(x_0)/C_\infty(\infty)$. Observe that when $C^*_\infty(x_0)/C_\infty(\infty) > 1$, no age replacement should take place. The cost efficiency of the age replacement strategy is seen to decrease when t_0 increases. □

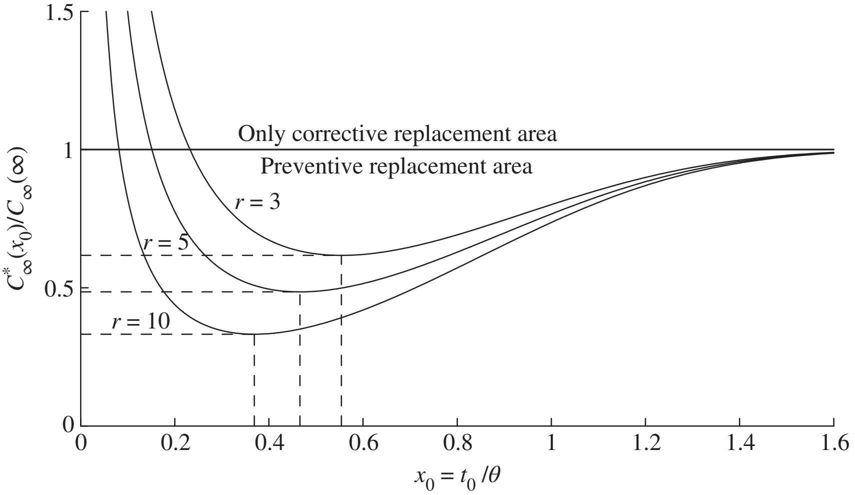


Figure 12.2 The ratio $C^*_\infty(x_0)/C_\infty(\infty)$ as a function of x_0 for the Weibull distribution with shape parameter $\alpha = 3$ and $r = 3, 5$, and 10.

Time Between Failures

Let $T_{F,1}, T_{F,2}, \dots$ be the times between consecutive actual failures. This may be represented as a renewal process where the renewals are the actual failures. The dataset for component i contains a random number, N_i of time periods of length t_0 (corresponding to replacements without failure), plus a last time period in which the item fails at an age Z_i, less than t_0, such that

$$T_{F,i} = N_i t_0 + Z_i \qquad \text{for } i = 1, 2, \dots .$$

The distribution of the random variable N_i can be found by using a *geometric distribution* (see Section 5.8.3)

$$\Pr(N_i = n) = [1 - F(t_0)]^n F(t_0) \qquad \text{for } n = 0, 1, \dots .$$

The mean number of replacements without failure for replacement age t_0 is

$$E(N_i) = \sum_{n=0}^{\infty} n \Pr(N_i = n) = \frac{1 - F(t_0)}{F(t_0)}. \tag{12.11}$$

The distribution of Z_i is

$$\Pr(Z_i \le t) = \Pr(T \le t \mid T \le t_0) = \frac{F(t)}{F(t_0)} \qquad \text{for } 0 < t \le t_0.$$

Hence,

$$E(Z_i) = \int_0^{t_0} \left(1 - \frac{F(t)}{F(t_0)}\right) dt = \frac{1}{F(t_0)} \int_0^{t_0} [F(t_0) - F(t)]\, dt. \tag{12.12}$$

The mean time between actual failures when the replacement age is t_0 becomes

$$\begin{aligned} E(T_{F,i}) &= t_0 E(N_i) + E(Z_i) \\ &= \frac{1}{F(t_0)} \left(t_0[1 - F(t_0)] + \int_0^{t_0} [F(t_0) - F(t)]\, dt \right) \\ &= \frac{1}{F(t_0)} \int_0^{t_0} [1 - F(t)]\, dt. \end{aligned} \tag{12.13}$$

Age Replacement – Availability Criterion

In some applications, the unavailability of the item is more important than the cost of replacement/repair, and it may be of interest to determine the replacement age t_0, that minimizes the average unavailability of the item. Let MDT_P be the mean downtime for a planned replacement, and MDT_F be the mean downtime needed to restore the function after a failure. The total mean downtime for the replacement age t_0 is

$$\begin{aligned} \text{MDT}(t_0) &= \text{MDT}_F F(t_0) + \text{MDT}_P[1 - F(t_0)] \\ &= [\text{MDT}_F - \text{MDT}_P] F(t_0) + \text{MDT}_P \end{aligned}$$

The mean time between replacements is $\text{MTBR}(t_0) = E(T_R)$ with

$$\begin{aligned} E(T_R) &= \int_0^{t_0} [1 - F(t)]\, dt + \text{MDT}_F F(t_0) + \text{MFD}_P[1 - F(t_0)] \\ &= \int_0^{t_0} [1 - F(t)]\, dt + \text{MDT}_P + [\text{MDT}_F - \text{MDT}_P]F(t_0) \end{aligned}$$

The average unavailability for replacement age t_0 is therefore

$$\begin{aligned} \overline{A}_{\text{av}}(t_0) &= \frac{\text{MDT}(t_0)}{\text{MTBR}(t_0)} \\ &= \frac{[\text{MDT}_F - \text{MDT}_P]F(t_0) + \text{MDT}_P}{\int_0^{t_0}[1 - F(t)]\, dt + \text{MDT}_P + [\text{MDT}_F - \text{MDT}_P]F(t_0)}. \end{aligned} \tag{12.14}$$

The optimal replacement age t_0 is the value of t_0 that minimizes $\overline{A}_{\text{av}}(t_0)$ in (12.14). This value may be found by the same approach as for the cost criterion.

12.3.2 Block Replacement

An item that is maintained under a block replacement strategy is preventively replaced at regular time intervals $(t_0, 2t_0, \ldots)$ regardless of age, and correctively replaced at failure dates. The block replacement strategy is easier to manage than an age replacement strategy because only the elapsed (calendar) time since last replacement must be monitored, rather than the operational time since last replacement. The block replacement strategy is therefore commonly used when there are a large number of similar items in service. The main drawback of the block replacement strategy is that it is rather wasteful, because almost new items may be replaced at planned replacement times.

Consider an item that is put into operation at time $t = 0$. The time-to-failure T of the item has distribution function $F(t) = \Pr(T \leq t)$. The item is operated under a block replacement strategy where it is preventively replaced at times $t_0, 2t_0, \ldots$. The preventive replacement cost is c. If the item fails in an interval, it is immediately repaired or replaced. The cost of the unplanned repair is k. Let $N(t_0)$ be the number of failures/replacements in an interval of length t_0, and let $W(t_0) = E[N(t_0)]$ be the mean number of failures/repairs in the interval.

The renewal cycle is $T_R = t_0$, and because t_0 is deterministic, $E(T_R) = t_0$. The average cost in a renewal cycle is $E[C(T_R)] = c + kW(t_0)$. The average cost per time unit $E[C(T_R)]/E(T_R)$ when using a block replacement interval of length t_0 is denoted $C_\infty(t_0)$ as it depends on one parameter t_0 and is equal to

$$C_\infty(t_0) = \frac{c + kW(t_0)}{t_0}. \tag{12.15}$$

Consider a block replacement model where the replacement interval t_0 is considered to be so short that the probability of having more than one failure in a block replacement interval is negligible. In this case, we may use the approximation

$$\begin{aligned} W(t_0) = E[N(t_0)] &= \sum_{n=0}^{\infty} n \Pr(N(t_0) = n) \\ &\approx \Pr(N(t_0) = 1) = \Pr(T \le t_0) = F(t_0) \end{aligned}$$

The average cost $C(t_0)$ per time unit is then

$$C_\infty(t_0) \approx \frac{c + kF(t_0)}{t_0}. \tag{12.16}$$

The minimum of $C_\infty(t_0)$ is found by solving $dC_\infty(t_0)/dt_0 = 0$ and gives

$$\frac{c}{k} + F(t_0) = t_0\, F'(t_0). \tag{12.17}$$

Example 12.2 (Block replacement)

Assume that $F(t)$ is a Weibull distribution with shape parameter $\alpha > 1$ and scale parameter θ. The optimal replacement interval can be found by solving

$$\frac{c}{k} + 1 - e^{-(t_0/\theta)^\alpha} = t_0\, \frac{\alpha}{\theta^\alpha} t_0^{\alpha-1} e^{-(t_0/\theta)^\alpha} = \frac{\alpha}{\theta^\alpha} t_0^\alpha e^{-(t_0/\theta)^\alpha},$$

which can be written as

$$\frac{c}{k} + 1 = (1 + \alpha(t_0/\theta)^\alpha) e^{-(t_0/\theta)^\alpha}. \tag{12.18}$$

For this model to be realistic, the preventive replacement cost c must be small compared to the corrective replacement cost k. By introducing $x = (t_0/\theta)^\alpha$, and using the approximation $e^x \approx 1 + x + x^2/2$, we can solve (12.18) and get the approximative solution (when remembering that t_0 is small)

$$x \approx \frac{\alpha}{1 + c/k} - 1 - \sqrt{\left(\frac{\alpha}{1 + c/k} - 1\right)^2 - 2\left(1 - \frac{1}{1 + c/k}\right)}. \tag{12.19}$$

If we assume that $c/k = 0.1$ and $\alpha = 2$, we get the optimal value $t_0 = 1/\lambda x^{1/\alpha} \approx 0.35\theta \approx 0.39$ MTTF. With the same value of c/k and $\alpha = 3$, we get the optimal value $t_0 \approx 0.391/\lambda \approx 0.44$ MTTF. In Figure 12.3, the optimal replacement interval t_0 is plotted as a function of α. The optimal value t_0 is equal to h MTTF, where MTTF is the mean of the Weibull distribution with parameters α and θ. □

Block Replacement with Minimal Repair

The block replacement strategy may be modified by only carrying out *minimal repair* when items fail in the block interval. The assumption is that a minimal repair is often adequate until the next planned replacement. In this case, we have a nonhomogeneous Poisson process (NHPP) within the block interval of length t_0, and we may use the formulas developed in Section 10.4.1 to determine $W(t_0)$.

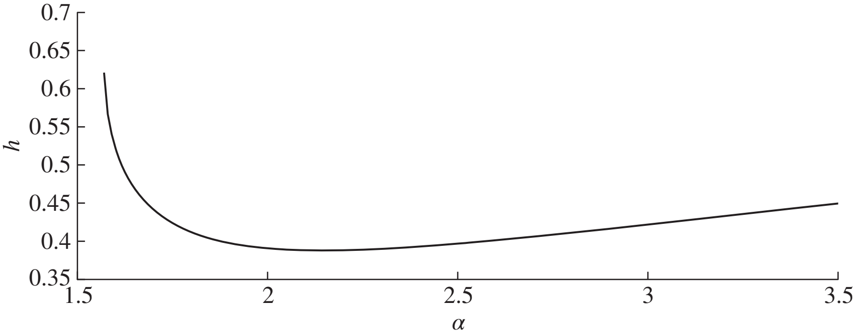


Figure 12.3 The optimal replacement interval t_0 in Example 12.2 as a function of the shape parameter α of the Weibull distribution. The optimal value t_0 is equal to h MTTF.

This modified block replacement model was proposed and studied by Barlow and Hunter (1960).

Another approach would be to assume that we carry out normal (imperfect) repairs in the block interval. In that case, we may use the theory described in Section 10.4.1 to determine $W(t_0)$.

Block Replacement with Limited Number of Spares

Consider an item that is operated under a block replacement strategy. We now assume that the number m of spares that may be used in a replacement interval is limited. In this case, we may run out of spares and the item's function may therefore be unavailable during a part of the replacement interval. The times-to-failure $T_1, T_2, \dots$ of the items are assumed to be independent and identically distributed with distribution function $F(t)$.

Let k_u denote the cost per time unit when the item function is not available, and let $\widetilde{T}_u(t_0)$ be the time the item remains unavailable in a replacement interval of length t_0. Hence, we have $\widetilde{T}_u(t_0; m) = t_0 - \sum_{i=1}^{m+1} T_i$ if the initial item and the m spares fail in the replacement interval, and $\widetilde{T}_u(t_0; m) = 0$ if less than $m + 1$ failures occur.

The same number m of spares are assumed to be made available for each replacement interval. All intervals therefore have the same stochastic properties, and we may therefore confine ourselves to studying the first interval $(0, t_0)$.

The mean cost in a replacement interval is $c + kE[N(t_0)] + k_u E[\widetilde{T}_u(t_0; m)]$, and the cost $C_\infty(t_0; m)$ when using a block replacement interval of length t_0 is

$$C_\infty(t_0; m) = \frac{c + kE[N(t_0)] + k_u E[\widetilde{T}_u(t_0; m)]}{t_0}, \tag{12.20}$$

where $N(t_0)$ is the number of replacements in $(0, t_0)$.

Example 12.3 (Block replacement without spare item)
Consider an item that is operated under a block replacement strategy without any spare item ($m = 0$) in each block interval. In this case, (12.20) can be written

$$C_\infty(t_0; 0) = \frac{c + kF(t_0) + k_u \int_0^{t_0} F(t)\, dt}{t_0}. \tag{12.21}$$

Let $F(t)$ be a Weibull distribution with shape parameter $\alpha = 3$ and scale parameter $\lambda = 0.1$. In Figure 12.4, $C_\infty(t_0; 0)$ is plotted as a function of t_0 for some selected cost values $c,\ k$, and k_u that give three different shapes.

When the replacement period t_0 tends toward infinity, the block replacement strategy is equivalent to leave the item as it is, and not replace it. Then the average cost per time unit will tend to k_u. When $c = 3$, the shape of the curve is quite similar to the corresponding curve for the classical age replacement strategy with optimal replacement period. When $c = 10$, the optimal block replacement cost $C_\infty(t_0; 0)$ is close to k_u. When t_0 increases, $C_\infty(t_0; 0)$ remains close to the replacement cost until the influence of $\widetilde{T}_u(t_0)$ becomes sufficiently large. When $c = 20$, the curve does not have a very distinctive minimum, and we may as well choose a very long replacement interval. □

Example 12.4 (Block replacement with limited number of spare items)
Consider an item that is operated under a block replacement strategy with m spare items in each block replacement interval. The time-to-failure T is assumed to be gamma distributed with parameters λ and α. The density of T is

$$f_T(t) = \frac{\lambda}{\Gamma(\alpha)} (\lambda t)^{\alpha-1}\, e^{-\lambda t}. \tag{12.22}$$

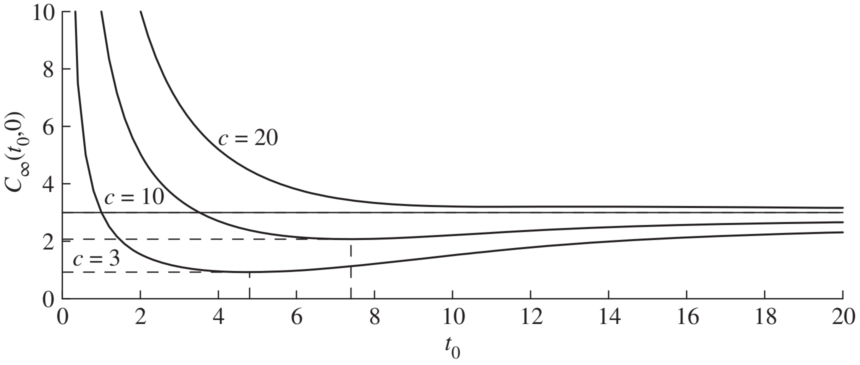


Figure 12.4 The average cost per time unit for a block replacement strategy with no spares when the time-to-failure distribution is a Weibull distribution with $\alpha = 3$ and $\lambda = 0.1$, $k = 10$, $k_u = 3$, and $c = 3,\ 10,\ 20$.

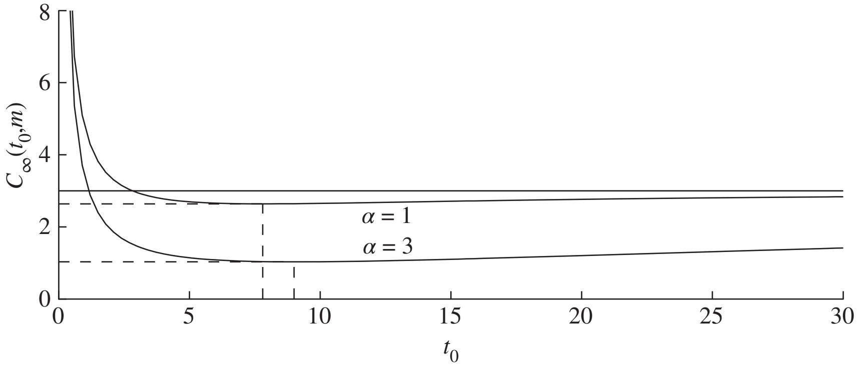


Figure 12.5 The average cost per time unit for a block replacement strategy with $m = 5$ spares when the time-to-failure distribution is a gamma distribution with parameters α and $\lambda = 1$, for $\alpha = 1$ and 3, and $k = 10$, $k_u = 3$, and $c = 3$.

The item function will be unavailable when the initial item and the m spares have failed. The time to system failure is therefore $T_s = \sum_{i=1}^{m+1} T_i$ where the times to individual failure $T_1, T_2, \dots, T_{m+1}$ are assumed to be independent and identically distributed with density $f_T(t)$. Let $F^{(m+1)}(t)$ denote the distribution function of T_s. The distribution $F^{(m+1)}(t)$ can be found by taking the $(m + 1)$-fold convolution of $F(t)$ (see Section 10.3.2). Because the gamma distribution is "closed under addition," T_s is gamma distributed with parameters λ and $(m + 1)\alpha$. In this case, (12.22) may be written

$$C_\infty(t_0; m) = \frac{c + kF^{(m+1)}(t_0) + k_u \int_0^{t_0} F^{(m+1)}(t)\, dt}{t_0}. \tag{12.23}$$

In Figure 12.5, $C_\infty(t_0; m)$ is plotted as a function of t_0 for some selected values of the parameter α, and cost values c, k, and k_u. □

The cost k of a repair/replacement in the block replacement interval may be extended to be time-dependent and to include other types of costs, for example, if the item deteriorates during the interval and will require increasing operating costs.

12.3.3 *P–F* Intervals

We now study an inspection and replacement strategy known as the *P–F* interval approach. The *P–F* interval approach is discussed in most of the main references on reliability centered maintenance (RCM).

Consider an item that is exposed to random shocks (events) and assume that the shocks occur as a homogeneous Poisson process (HPP) with rate λ. The time

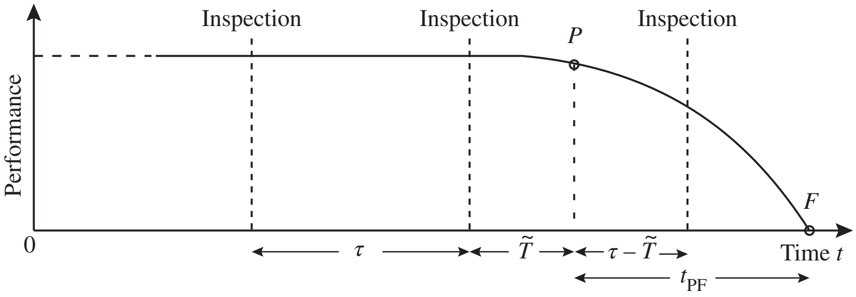


Figure 12.6 Average behavior and concepts used in *P–F* interval models.

between two consecutive shocks is then, according to Section 10.2, exponentially distributed with rate λ, and mean $1/\lambda$. When a shock occurs, it produces a weakness (potential failure) in the item that, in time, will develop/deteriorate into a critical failure. We are not able to observe the shocks, but may be able to reveal indications of potential failures some time after the shock has occurred. Let *P* be the point of time (after a shock) when an indication of a potential failure can be first detected, and let *F* be the point of time where the item has functionally failed. The time interval from *P* to *F* is called the *P–F* interval, and is generally a random variable. If a potential failure is detected between *P* and *F* in Figure 12.6, this is the time interval in which it is possible to carry out a task to prevent the failure and to avoid its consequences. The cost of a preventive replacement (or repair) is c_P, and the cost of a corrective replacement after a critical failure has occurred is c_C.

The item is inspected at regular intervals of length τ, and the cost of each inspection is c_I. The inspections may be observations using human senses (view, smell, sound), or we may use some monitoring equipment. In the most simple setup, we assume that the inspection procedure is perfect, such that all potential failures are detected by the inspection. In many cases, this is not a realistic assumption, and the probability of successful detection may be a function of the time since *P*, the time of the year, and so on. Our main objective in this section is to find the optimal inspection interval τ, that is, the value of τ that gives the lowest mean average cost.

The length of the *P–F* interval generally depends on the materials and characteristics of the item, the failure mode, the failure mechanisms, and the environmental and operational conditions. Estimates of *P–F* intervals are not available in reliability data sources and must be estimated by expert judgment by operators, specialists on deteriorating mechanisms, and equipment designers. The length of the *P–F* interval may be regarded as a random variable T_{PF} with a subjective distribution function (see Chapter 15).

Example 12.5 (Cracks in railway rails)
Vatn and Svee (2002) study crack occurrences and crack detection in (railroad) rails. In their model, cracks are initiated at random. The frequency λ of initiated cracks may be measured as the number of initiated cracks per unit length of rails and per time unit. The frequency generally depends on the traffic load, the material and geometry of the rail, and various environmental factors, but may also be caused by particles on the rails or "shocks" from trains with noncircular wheels. In the first phase, the cracks are very small, and very difficult to detect. A special rail-car equipped with ultrasonic inspection equipment is used to inspect the rails. When a crack has grown to a specific size, it should be detectable by ultrasonic inspection. This crack-size corresponds to the potential failure P described above. The P–F interval is the time interval from an observable crack P is present until a critical failure F occurs. The critical failure F is, in this case, breakage of a rail and possible derailment of a train. Ultrasonic inspection is carried out at regular intervals, at a rather high cost. It is therefore of interest to find an optimal inspection interval, that balances the inspection cost and the costs related to replacements and potential accidents. □

Our objective is to find the inspection interval τ that minimizes the mean total cost. In the general setup, this is a rather difficult task. We therefore start by solving the problem in the most simple situation, with known (deterministic) P–F interval and known repair time. Thereafter, we present some ideas on how to solve the problem in a more realistic setup.

Deterministic *P*–*F* Interval and Repair Time and Perfect Inspection

To simplify the problem, assume that the length of the P–F interval t_{PF} is known (deterministic). The time from a potential failure P is detected (during the first inspection after P), until the failure has been corrected, t_{Rep}, is assumed to be known (deterministic). We further assume that the inspections are perfect such that all potential failures are detected during the inspections. Figure 12.6 shows that we have a preventive replacement when $\tau - t + t_{\text{Rep}} < t_{\text{PF}}$, and a corrective replacement if $\tau - t + t_{\text{Rep}} > t_{\text{PF}}$. If $\tau + t_{\text{Rep}} < t_{\text{PF}}$, all the replacements are preventive, and there is no problem to optimize. We therefore assume that $\tau + t_{\text{Rep}} > t_{\text{PF}}$ (see Remark 12.1).

Assume that we start observing the item at time $t = 0$ and that the potential failure P is observable a short time after the shock occurs. The time T from startup to P is exponentially distributed with failure rate λ. Let $N(\tau)$ be the number of inspection intervals before a shock occurs. The event $N(\tau) = n$ hence means that we observe n inspection intervals without any shock, and the shock occurs in inspection interval $n + 1$. The random variable $N(\tau)$ has a geometric distribution with point probability

$$\Pr(N(\tau) = n) = (e^{-\lambda\tau})^n(1 - e^{-\lambda\tau}) \qquad \text{for } n = 0, 1, \dots ,$$

and mean value

$$E[N(\tau)] = \frac{e^{-\lambda\tau}}{1 - e^{-\lambda\tau}}.$$

Assume that a shock and an observable potential failure P has occurred in inspection interval $n + 1$. Let $\widetilde{T}$ be the time from inspection n till P. The probability distribution of $\widetilde{T}$ is

$$\Pr(\widetilde{T} \le t) = \Pr(T \le t \mid T \le \tau) = \frac{1 - e^{-\lambda t}}{1 - e^{-\lambda\tau}} \qquad \text{for } 0 < t \le \tau.$$

A preventive replacement will therefore take place with probability

$$P_P(\tau) = \Pr(\widetilde{T} > \tau + t_{\text{Rep}} - t_{\text{PF}}) = 1 - \frac{1 - e^{-\lambda(\tau + t_{\text{Rep}} - t_{\text{PF}})}}{1 - e^{-\lambda\tau}}.$$

A corrective replacement will take place with probability

$$P_C(\tau) = \Pr(\widetilde{T} < \tau + t_{\text{Rep}} - t_{\text{PF}}) = \frac{1 - e^{-\lambda(\tau + t_{\text{Rep}} - t_{\text{PF}})}}{1 - e^{-\lambda\tau}}.$$

If we know that the potential failure results in a critical failure (corrective maintenance, CM), the mean time to this failure is $1/\lambda + t_{\text{PF}}$. On the other hand, if we know that the potential failure results in a preventive replacement, the mean time to this replacement is $E(N(\tau) + 1)\tau + t_{\text{Rep}}$. The mean time between replacements is therefore

$$\begin{aligned}\text{MTBR}(\tau) &= \left(\frac{1}{\lambda} + t_{\text{PF}}\right) P_C(\tau) + (E(N(\tau) + 1)\tau + t_{\text{Rep}})P_P(\tau) \\ &= \left(\frac{1}{\lambda} + t_{\text{PF}}\right) P_C(\tau) + \left(\frac{\tau}{1 - e^{-\lambda\tau}} + t_{\text{Rep}}\right) P_P(\tau). \end{aligned} \tag{12.24}$$

The mean total cost in a replacement interval (renewal cycle) $E[C(T_R)]$ is

$$E[C(T_R)] = c_P P_P(\tau) + c_C P_C(\tau) + c_I[E[N(\tau)] + \Pr(\widetilde{T} > \tau - t_{\text{PF}})],$$

where $\Pr(\widetilde{T} > \tau - t_{\text{PF}})$ is the probability that the item will not fail within the inspection interval where the potential failure occurred, and consequently that the next inspection will be carried out. When $\tau - t_{\text{PF}} > 0$, this probability is

$$\Pr(\widetilde{T} > \tau - t_{\text{PF}}) = \Pr(T > \tau - t_{\text{PF}} \mid T \le \tau) = \frac{e^{-\lambda(\tau - t_{\text{PF}})} - e^{-\lambda\tau}}{1 - e^{-\lambda\tau}}.$$

We therefore have that

$$\Pr(\widetilde{T} > \tau - t_{\text{PF}}) = \begin{cases} \dfrac{e^{-\lambda(\tau - t_{\text{PF}})} - e^{-\lambda\tau}}{1 - e^{-\lambda\tau}} & \text{for } \tau - t_{\text{PF}} > 0 \\ 1 & \text{for } \tau - t_{\text{PF}} < 0 \end{cases}.$$

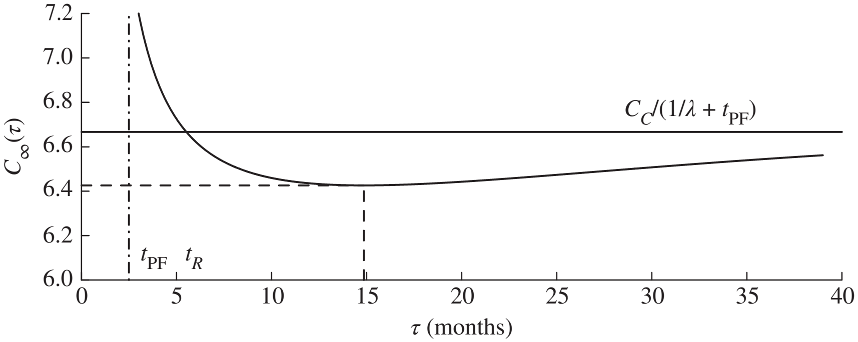


Figure 12.7 The asymptotic cost $C_\infty(\tau)$ per time unit as a function of τ for $\lambda = 1/12$ mo^{-1}, $t_{\text{PF}} = 3$ months, $t_R = 0.5$ month, $C_C = 100$, $C_P = 20$, and $C_I = 15$.

The mean total cost in a replacement interval is therefore

$$E[C(T_R)] = \begin{cases} c_P P_P(\tau) + c_C P_C(\tau) + c_I \frac{e^{-\lambda(\tau - t_{\text{PF}})}}{1-e^{-\lambda\tau}} & \text{for } \tau - t_{\text{PF}} > 0 \\ c_P P_P(\tau) + c_C P_C(\tau) + c_I \left(\dfrac{e^{-\lambda\tau}}{1 - e^{-\lambda\tau}} + 1 \right) & \text{for } \tau - t_{\text{PF}} < 0 \end{cases}.$$

The mean total cost per time unit with inspection interval τ is denoted $C_\infty(\tau)$, depends on one parameter τ, and is equal to

$$C_\infty(\tau) = \frac{E[C(T_R)]}{E(T_R)} = \frac{E[C(T_R)]}{\text{MTBR}(\tau)}. \tag{12.25}$$

To find the value of τ for which (12.25) attains its minimum is not a straightforward task. The optimal τ may be found graphically by plotting $C_\infty(\tau)$ as a function of τ. An example is shown in Figure 12.7.

Remark 12.1

In the case when $\tau + t_{\text{Rep}} < t_{\text{PF}}$, all the replacements are preventive, and the mean time between replacements is $\text{MTBR}(\tau) = (E[N(\tau)] + 1)\tau + t_{\text{Rep}}$. The total cost in a replacement period is $C_T(\tau) = c_P + c_I(E[N(\tau)] + 1)$. The optimal replacement interval (with the restriction that $\tau + t_{\text{Rep}} < t_{\text{PF}}$) can therefore be found by minimizing

$$C_\infty(\tau) = \frac{E[C(T_R)]}{\text{MTBR}(\tau)} = \frac{c_I/(1 - e^{-\lambda\tau}) + c_P}{\tau/(1 - e^{-\lambda\tau}) + t_{\text{Rep}}}.$$

□

Stochastic *P–F* Interval, Deterministic Repair Time, and Nonperfect Inspection

Reconsider the situation described above, but assume that the inspection is not perfect. In general, the probability of detecting a potential failure depends on the time since the potential failure became observable. When a crack in the rail in Example 12.5 has been initiated, it will grow with time. The probability of detecting the crack is assumed to increase with the size of the crack. A model where the probability of successful detection is a function of the crack size will be rather complicated. We therefore simplify the situation and introduce $\theta_i(\tau)$ to be the probability that the potential failure is *not* detected in inspection i after that an observable potential failure P has occurred, for $i = 1, 2, \ldots$. The probability is assumed to be a function of the inspection interval τ. We assume that $1 > \theta_1(\tau) \geq \theta_2(\tau) \geq \cdots$.

The P–F interval, T_{PF} is assumed to be a random variable with distribution function $F_{\text{PF}}(t)$. The repair time t_R is assumed to be known (deterministic). Let $T_F = \widetilde{T} + T_{\text{PF}}$. The variable T_F is hence the time from the last inspection before P until a (possible) critical failure. The distribution of T_F can be found by the convolution of the distribution of $\widetilde{T}$ and $T_{\text{PF}}(t)$.

$$\begin{aligned} F_F(t) = \Pr(T_F \leq t) &= \int_0^\tau F_{\text{PF}}(t-u)\, dF_{\widetilde{T}}(u) \\ &= \frac{\lambda}{1-e^{-\lambda\tau}} \int_0^\tau F_{\text{PF}}(t-u)\, e^{-\lambda u}\, du. \end{aligned} \tag{12.26}$$

Let $R_F(t) = 1 - F_F(t)$, and let $Z(\tau)$ be the number of inspections carried out after a potential failure P has occurred. We want to find the probabilities $\Pr(Z(\tau) \geq k)$ for $k = 0, 1, \ldots$. It is obvious that $\Pr(Z(\tau) \geq 0) = 1$. At least one inspection will be carried out if $T_F = \widetilde{T} + T_{\text{PF}} > \tau$, that is,

$$\Pr(Z(\tau) \geq 1) = \Pr(T_F > \tau) = R_F(\tau).$$

At least two inspections will be carried out if $T_F > \tau$, the failure is not detected in the first inspection, and if $T_F > 2\tau$. Because $\Pr(T_F > \tau \cap T_F > 2\tau) = \Pr(T_F > 2\tau)$ we get

$$\Pr(Z(\tau) \geq 2) = \theta_1(\tau) R_F(2\tau).$$

By continuing this argument, we get in the general case that (we define $\theta_0(\tau) = 1$):

$$\Pr(Z(\tau) \geq k) = \left(\prod_{j=0}^{k-1} \theta_j(\tau) \right) R_F(k\tau) \qquad \text{for } k = 1, 2, \ldots . \tag{12.27}$$

The mean number of inspections is therefore

$$E[Z(\tau)] = \sum_{k=1}^{\infty} \Pr(Z(\tau) \geq k) = \sum_{k=1}^{\infty} \left(\prod_{j=0}^{k-1} \theta_j(\tau) \right) R_F(k\tau).$$

A preventive replacement will take place with probability

$$\begin{aligned} P_P(\tau) = {} & [1 - \theta_1(\tau)] \Pr(T_F > \tau + t_{\text{Rep}}) \\ & + \theta_1(\tau)[1 - \theta_2(\tau)] \Pr(T_F > 2\tau + t_{\text{Rep}}) + \cdots \end{aligned}$$

which can be written as

$$P_P(\tau) = \sum_{k=1}^{\infty}[1 - \theta_k(\tau)] \prod_{j=0}^{k-1} \theta_j(\tau) \Pr(T_F > k\tau + t_{\text{Rep}})$$

$$= \sum_{k=1}^{\infty}[1 - \theta_k(\tau)] \prod_{j=0}^{k-1} \theta_j(\tau)\; R_F(k\tau + t_{\text{Rep}}). \tag{12.28}$$

A corrective replacement will take place with probability $P_C(\tau) = 1 - P_P(\tau)$. Let $Z_P(\tau)$ be the number of inspections that are carried out after a potential failure P has occurred, when we know that the item will be preventively replaced. By using the same argument as we used to find (12.27), we get

$$\Pr(Z_P(\tau) \geq k) = \prod_{j=1}^{k-1} \theta_j(\tau),$$

and the mean value is

$$E[Z_P(\tau)] = \sum_{k=1}^{\infty} \prod_{j=1}^{k-1} \theta_j(\tau).$$

The mean time between replacements is therefore

$$E(T_R) = \text{MTBR}(\tau) = \left(\frac{1}{\lambda} + E(T_{\text{PF}})\right) P_C(\tau) + [(E[N(\tau)] + E[Z_P(\tau)])\tau + t_{\text{Rep}}]P_P(\tau). \tag{12.29}$$

The mean total cost $E[C(T_R)]$ in a replacement interval (renewal cycle) is

$$E[C(T_R)] = c_P P_P(\tau) + c_C P_C(\tau) + c_I(E[N(\tau)] + E[Z(\tau)]). \tag{12.30}$$

The optimal inspection interval τ may in this case be determined as the value of τ that minimizes $C_{\infty}(\tau) = E[C(T_R)]/\text{MTBR}(\tau)$.

The models described in this section may be extended in many different ways. An obvious extension is to let the repair time be a random variable T_{Rep}. Another extension is to let the time to potential failure P have an increasing failure rate function.

Delay Time Models

Few references are available discussing quantitative assessment related to the *P–F* interval approach. Some further developments have been made based on the *delay-time* concept that was introduced in maintenance applications by Christer and Waller (1984). The delay-time model assumes that a failure is dependent on the occurrence of a defect (an incipient or potential failure). The time-to-failure T of an item can therefore be divided in two parts: (i) the time T_P from startup until a defect occurs and (ii) the delay-time T_{PF} from the defect occurred until the item fails.

Several inspection models have been developed based on the delay-time principle covering, for example, imperfect inspections, nonconstant defect rate, nonstationary inspection rules (Wang 2008). Some of these models have been applied in industry (e.g. see Dekker 1996).

12.4 Degradation Models

For the time-based strategies presented in Section 12.3, it is relevant to use a time-to-failure model for the item. When, on the other hand, the maintenance decisions are condition-based and the item is experiencing a degradation that can be observed, the time-to-failure models are not suitable and we have to use *degradation models*. With degradation models, the state space of the items is not reduced to functioning or failed. The functioning and failed states may be split into different substates that are more or less degraded. These substates can be defined in a finite discrete state space, in an infinite discrete state space, or in a continuous state space. Consider an item that gradually *degrades* when it is being used. For mechanical items, the degradation may lead to an increasing number of initiated cracks (possibly infinite discrete state space), crack lengths, corrosion depths, level of vibration (continuous state space), and many other physical quantities.

Some of these quantities may be monitored by using a *degradation indicator*, which may be measured continuously or periodically. In this chapter, the values obtained by using the degradation indicator are assumed to be (univariate) scalars. As an illustration, assume that the degradation indicator measures a crack length at a specific location. At time t, the true value of this crack length is $x(t)$, whereas the measured value obtained by the degradation indicator is $y(t)$, which may be slightly different from $x(t)$ because of measurement errors and "noise." Both the degradation and the measurements are subject to random variations and may be modeled as stochastic processes $\{X(t), t \geq 0\}$ and $\{Y(t), t \geq 0\}$, respectively. At time t, $x(t)$ and $y(t)$ are specific (numerical) outcomes of these two stochastic processes. In the following, $X(t)$ denotes the (random) state of the item and $Y(t)$ its observable (random) condition. In some situations, we may assume that $X(t)$ is available and the noise or measurement errors can be neglected.

A *degradation model* is a stochastic process $\{X(t), t \geq 0\}$ together with a set of assumptions about the probability distribution of $X(t)$ and the development of $X(t)$, as a function of time. This section briefly examines three categories of degradation models: trend models, models with increments, and shock models. All of these have a continuous state space. A brief reminder about discrete state space models (Markov processes) is provided. More details may be found in Chapters 9 and 10.

When the presence of degradation is acknowledged, the maintenance decision can be based on the degradation indicator (Christer and Wang 1992), but also on the RUL distribution (Huynh et al. 2014). RUL is increasingly used to optimize systems beyond maintenance decisions (Langeron et al. 2015; Lee et al. 2011). The quality of the collected data seen in relation to establishing the degradation indicator or the RUL distribution is of primary importance and is extensively discussed in the literature with prognostics and diagnostics perspectives (Nguyen et al. 2019), but this is outside the scope of this book. The RUL (survival) distribution is a natural extension of the survivor function in the sense that it is a kind of survivor function conditioned by the knowledge about the current degradation indicator. RUL is introduced in Section 12.4.1 before describing degradation models.

12.4.1 Remaining Useful Lifetime

RUL of an item at time t_j is briefly introduced in Chapter 5. RUL(t_j) is a random variable that measures the time from t_j until the item is not "useful" any more. What is meant by not being "useful" must be carefully specified. The distribution of RUL is a reliability metric that can be used when a degradation model is available. The RUL distribution is written as

$$\Pr(\mathrm{RUL}(t_j) \leq t) = F_{\mathrm{RUL}(t_j)}(t). \tag{12.31}$$

Remark 12.2 (Another interpretation of RUL)
In some practical applications, RUL(t_j) is given as a fixed number, which is an *estimate* of the mean time period the item can survive after a given time t_j in a specified operating context, and based on the knowledge available about the previous condition development. In this book, RUL(t_j) is always considered to be a random time variable. □

In reliability theory, *prognostics* is a field dedicated to the estimation of the probability distribution of RUL(t_j) or its mean. Prognostics may be addressed from different points of view and with different methods, usually classified as *model-based* and *data-driven* approaches. Model-based approaches are used for physical models dedicated to specific applications and are as such outside the scope of this book. For data-driven prognostics, we distinguish between two categories:

Data-driven prognostics with no probabilistic modeling. These approaches (e.g. kernels, machine learning, and artificial intelligence) rely on the observed data without any prior choice of a unique degradation model between the running state and the failed state. These approaches are not presented here, but a brief introduction is given hereafter to indicate what the approaches consist of and when they may be relevant. With such approaches, the distribution of RUL(t_j) is usually not derived but replaced by an estimate of the mean value of RUL(t_j) with confidence interval.

Data-driven prognostics with probabilistic modeling. These approaches rely on historical data to fit the parameters of an a priori chosen degradation model $\{X(t), t \geq 0\}$. Statistical methods presented in Chapter 14 may be used in combination with physical considerations to choose the degradation model.

Sections 12.4.2 and 12.4.3 introduce the most common degradation models. With these approaches, the distribution of $\mathrm{RUL}(t_j)$ may be derived analytically or estimated empirically by simulating the degradation model. An example is provided by Le Son et al. (2013) and a review of data-driven prognostics with probabilistic modeling is proposed by Si et al. (2011).

Data-Driven Prognostics with no Probabilistic Modeling

In this section, $X(t)$ is not established a priori, and there is no intention to explain the degradation phenomenon by stating that it should be a trend model, a model with increments, or a shock model. Instead, a link is built between the data from condition monitoring and the value of RUL. RUL is defined at any time t_i as a generic function, f, such that

$$\widehat{\mathrm{RUL}}(t_i) = f(t_i, y_i, u_i), \tag{12.32}$$

where t_i is the current time, y_i is the measured value (or a vector of values) related to the current condition of the item, and u_i is a vector of measured values describing the operating context. RUL is estimated by learning the structure of the function f (whether it is a linear function, a polynomial one, an exponential one, etc.) and its parameters. This can be done by linear regression, neural networks, Bayesian networks (especially if expert judgment or qualitative data are available), and so on. Observe that the maintenance decision is made based on the RUL estimate, and not the degradation model $X(t)$. All these methods rely on the existence of a dataset S, with more or less the following structure:

$$S = \{((t_k^j, y_k^j, u_k^j), \mathrm{RUL}_k^j\}_{j=1,2,\ldots,N;\ k=1,2,\ldots,\kappa}, \tag{12.33}$$

where t_k^j, for $k = 1, 2, \ldots, \kappa$, are the sampling times for item j, and RUL_k^j is the recorded RUL of item j at time t_k^j for the measures (y_k^j, u_k^j). The dataset must be divided into two parts, one for estimating the function f (learning dataset) and one for testing the quality of the estimate of f (testing dataset). The way the dataset is divided into two parts may influence the results of the estimation and must be carefully checked by resorting to cross validation methods.

Data-Driven Prognostics with Probabilistic Modeling

Let T be the time-to-failure of the item, $X(t_j + h)$ the future state of the item (i.e. at time t_j), $\mathcal{X}_\ell$ the set of failed (or unacceptable) states of the item, $\mathcal{T}_{t_j}$ the set of times when the item condition has been observed in $[0, t_j]$, and $Y(t)$ the item condition at these times. Then, given that $T > t_j$, $\mathrm{RUL}(t_j)$ is formally defined as

$$\mathrm{RUL}(t_j) = \min\{h; X(t_j + h) \in \mathcal{X}_\ell\}, \tag{12.34}$$

and its distribution is defined as

$$\Pr(\mathrm{RUL}(t_j) \leq t) = \Pr(\min\{h;\ X(t_j + h) \in \mathcal{X}_\ell\} \leq t \mid T \geq t_j,\ Y(t)_{t\in\mathcal{T}_{t_j}}).$$

To define RUL and to find its probability distribution, the following must be available:

(1) A variable $X(t)$ that describes the state of the item at time t.
(2) A set $\mathcal{X}_\ell$ of unacceptable states.
(3) A set $\mathcal{T}_{t_j}$ of observation times and condition observations $Y(t)$ at these times.
(4) We must be able to estimate $X(t)$ at time t, by filtering observed values of $Y(t)$, for $t \in \mathcal{T}_{t_j}$, if necessary.
(5) We must be able to predict the value of $X(t_j + h)$ at any time after t_j.

If $X(t)$ is a time-dependent scalar function, a degradation level ℓ can be defined as the lowest level of degradation that is considered to be a failure.

12.4.2 Trend Models; Regression-Based Models

Let $Y(t)$ be a time-dependent function with a continuous state space. Typical applications for such a model are degradation phenomena that can be monitored through a continuous natural trend, that is to say through variations of quantities, such as temperature, flow, velocity, and pressure as a function of time. The generic form of the model is

$$Y(t_k) = X(t_k) + \varepsilon(t_k), \tag{12.35}$$

where $Y(t_k)$ is the observed condition at time t_k, $X(t_k)$ is the actual degradation (where $X(t)$ is a monotonically increasing function), and $\varepsilon(t_k)$ is a random error (often referred to as *noise* from the monitoring device). In most cases, it is assumed that $\varepsilon(t_k) \sim \mathcal{N}(0, \sigma^2)$.

The following cases may be considered for $k = 1, 2, \dots$

$$\begin{aligned} Y(t_k) &= c + at_k + \varepsilon(t_k) && \text{(linear)}\\ Y(t_k) &= c + at_k + bt_k^2 + \varepsilon(t_k) && \text{(polynomial)}\\ \log[aY(t_k) + b] &= c + at_k + \cdots + \varepsilon(t_k) && \text{(logarithmic)}\\ ae^{bY(t_k)} &= c + at_k + \cdots + \varepsilon(t_k) && \text{(exponential),} \end{aligned}$$

where the model parameters $\{a, b, c, \dots\}$ can be deterministic or random.

Because $X(t)$ and $Y(t)$ are scalar functions of time, a degradation level ℓ can be defined as the lowest degradation level that is considered to be a failure and $\mathrm{RUL}(t_j)$ is defined as

$$\mathrm{RUL}(t_j) = \min\{h;\ X(t_j + h) \geq \ell\}, \tag{12.36}$$

for $T > t_j$. Because the current values of $X(t_k)$ and $Y(t_k)$ are not influenced by the observations of the item condition in the past, all information contained in $\mathcal{T}_{t_j}$ is useless except $Y(t_j)$, the observed condition at time t_j. Then

$$\Pr(\text{RUL}(t_j) \leq t) = \Pr(\min\{h;\ X(t_j + h) > \ell\} \leq t \mid T \geq t_j, Y(t_j) = y(t_j)).$$

If $X(t)$ is monotonically increasing and the noise is not too significant, the following approximation may be used

$$\begin{aligned} \Pr(\text{RUL}(t_j) \leq t) &\simeq \Pr[X(t_j + t) > \ell \mid T \geq t_j, Y(t_j) = y(t_j)] \\ &\simeq \Pr[Y(t_j + t) > \ell \mid T \geq t_j, Y(t_j) = y(t_j)] \\ &\simeq \Pr[Y(t_j + t) - Y(t_j) > \ell - y(t_j)] \quad \text{for } y(t_j) \leq \ell. \end{aligned} \tag{12.37}$$

Wiener Process with Linear Drift

The Wiener process[2] (or Brownian motion with linear drift) is a special case of a trend model and can be defined as:

$$Y(t_k) = at_k + \varepsilon(t_k), \qquad Y(0) = 0,$$

where the constant a is called the *drift parameter* and the *noise* is a random variable $\varepsilon(t_k)$ with probability distribution $\mathcal{N}(0, \sigma^2 t_k)$. Because the normal distribution can take both positive and negative values, the Wiener process is not monotonic. A possible interpretation is that the observed degradation is noisy whereas the true degradation is monotonically increasing in average. Another interpretation is that $Y(t) = X(t)$ and that the true degradation is directly observed with such fluctuations. This may be the case when a crack is randomly clogging. For the sake of clarity, we use $Y(t)$ instead of $X(t)$. A Python script to simulate the paths of the Wiener process is provided on the `book companion site`.

Because a is deterministic:

$$\begin{aligned} E[Y(t_k)] &= at_k \\ E[Y(t_{k+1}) - Y(t_k)] &= a(t_{k+1} - t_k) \end{aligned}$$

and:

$$\begin{aligned} \text{var}[Y(t_k)] &= \sigma^2 t_k \\ \text{var}[Y(t_{k+1}) - Y(t_k)] &= \sigma^2(t_{k+1} + t_k) \end{aligned}$$

This means that (i) in average, the Wiener process is linearly increasing as a function of time with speed a and (ii) its variance is increasing with the time interval $t_{k+1} - t_k$ and the variance of the noise.

[2] Named after the US mathematician and philosopher Norbert Wiener (1894–1964).

The Distribution of RUL(t_j)

Because of the nonmonotonicity, there is no direct link between the distribution of RUL and the probability that the condition level $Y(t)$ is below the failure level ℓ. This calculation is not straightforward, but you may find more details in Kahle et al. (2016). For $y(t_j) \leq l$ the RUL distribution is given by

$$F_{\text{RUL }(t_j)}(t) = \Pr(\text{RUL}(t_j) \leq t)$$
$$= \int_0^{t+t_j} \frac{\ell - y(t_j)}{\sqrt{2\pi\sigma^2(u-t_j)^3}} e^{-\frac{(\ell - y(t_j) - a(u-t_j))^2}{2\sigma^2(u-t_j)}} \, du. \qquad (12.38)$$

The book companion site provides details about Wiener processes, its simulation, and its parameter estimation from degradation data. Examples of more advanced trend models are given by Le Son et al. (2013) and Deng et al. (2016). A polynomial trend model is studied in Problem 12.7.

12.4.3 Models with Increments

Consider a degradation process where $Y(t)$ is not explicitly established. Instead, we use a model of degradation increments, where of degradation $Y(t)$ increases in a time interval (t_j, t_k) in a continuous state space. We usually assume that the *degradation increment* $I_{(t_j,t_k)} = Y(t_k) - Y(t_j)$ is a random variable with a given probability distribution. Typical applications for such a model are degradation phenomena that can be monitored through increments of degradation, such as corrosion and erosion. For a review and some examples of applications of models with increments, see Ghamlouch et al. (2018) and Van Noortwijk (2009).

Example 12.6 (Exponentially distributed increments)

Consider a deteriorating item where the degradation increments are exponentially distributed with rate $\lambda/(t_j - t_k)$ between times t_j and t_k. Assume that the item is studied in two time intervals (t_1, t_2) and (t_2, t_3). The probability density functions of the degradation increments in the intervals are as follows:

$$f_{(t_1,t_2)}(x) = \frac{\lambda}{t_2 - t_1} e^{-\frac{\lambda}{t_2-t_1}x}$$
$$f_{(t_2,t_3)}(x) = \frac{\lambda}{t_3 - t_2} e^{-\frac{\lambda}{t_3-t_2}x}$$

For this example, observe that the mean degradation increment $E[I_{(t_1,t_2)}] = (t_2 - t_1)/\lambda$ and the variance $\text{var}[I_{(t_1,t_2)}] = (t_2 - t_1)^2/\lambda^2$ increase when the length of the interval increases. A Python script to simulate the paths of such a process is provided on the `book companion site`. □

The probability distribution of the increments is chosen such that the degradation model fits to the available dataset, and such that (i) the distribution parameters can be estimated with classical statistical methods and (ii) the distribution of RUL is obtainable. For this purpose, the class of Levy processes is often used. We describe briefly the main features of such processes with a particular case; the homogeneous gamma process.

Levy Process

A Levy process[3] is a continuous-time stochastic process $\{X(t), t \geq 0\}$ where the increments in disjoint time intervals are *independent* random variables. The Levy process satisfies the Markov property and is hence a Markov process, because the next degradation increment does not depend on the past increments. In addition, if the distribution of the increments depends only on $t_j - t_k$, but not on t_j neither t_k, the process is stationary or homogeneous in time. In this case, the increments are identically distributed for intervals of the same length, $t_j - t_k$, and the process is a homogeneous Markov process.

Homogeneous Gamma Process

A homogeneous gamma process is a special case of a Levy process. It is a continuous-time stochastic process $\{Y(t), t \geq 0\}$ where the increments in disjoint time intervals are *independent* random variables such that $Y(0) = 0$ and for any $t_2 > t_1 \geq 0$, the increment $Y(t_2) - Y(t_1)$ has a gamma density:

$$f_{\alpha(t_2-t_1),\beta}(y) = \frac{\beta}{\Gamma[\alpha(t_2 - t_1)]}(\beta y)^{\alpha(t_2-t_1)-1}e^{-\beta y} \qquad \text{for } y \geq 0.$$

Because the gamma density is defined only for positive values, the increments are always positive and the degradation model is always increasing. This means that the gamma process can be a suitable model also for $X(t)$. In that case, we have direct access to the degradation measure, without any additional noise. Then, the degradation increment $X(t_2) - X(t_1)$ has a gamma density $f_{\alpha(t_2-t_1),\beta}(x)$. The mean degradation in the interval (t_1, t_2) is

$$E[X(t_2) - X(t_1)] = \frac{\alpha(t_2 - t_1)}{\beta},$$

and the variance is:

$$\text{var}[X(t_2) - X(t_1)] = \frac{\alpha(t_2 - t_1)}{\beta^2}.$$

This means that the mean degradation in an interval of length t_0 is $\frac{\alpha}{\beta}t_0$, independent of when the interval begins. The parameter β is called the *rate* of the process. The variance of the process increases with the time horizon between t_1 and t_2 and

[3] Named after the French mathematician Paul Pierre Lévy (1886–1971).

can be tuned independently on the mean. A Python script to simulate the paths of the homogeneous gamma process is provided on the `book companion site`. For more details about the gamma process, simulation, and parameter estimation from degradation data, see the `book companion site`.

The Distribution of RUL(t_j)

Because $X(t)$ is a scalar function of time, a degradation level ℓ can be defined as the lowest degradation level that is considered to be a failure and RUL(t_j) is defined as a hitting time. The distribution function of RUL can be derived and computed numerically.

$$\begin{aligned} F_{\text{RUL }(t_j)}(t) &= \Pr(\text{RUL}(t_j) \leq t) \\ &= \Pr\left(X(t_j + t) \leq \ell \mid X(t_j) > \ell,\ X(s)_{s\in\mathcal{T}_{t_j}}\right) \quad \begin{array}{c}\text{Thanks to the}\\ \text{monotonicity}\end{array} \\ &= \Pr(X(t_j + t) - X(t_j) > \ell - x(t_j)) \quad \begin{array}{c}\text{Thanks to the}\\ \text{lack of memory}\end{array} \\ &= \int_{\ell - x(t_j)}^{+\infty} f_{\alpha t,\beta}(u)\, du \quad \text{for } \ x(t_j) \leq \ell. \end{aligned} \tag{12.39}$$

Observe that the gamma process is a jump process. The jumps, whose size lies in the interval $[x, x + dx)$, occur as a Poisson process with an intensity depending on x. In practice, this means that the modeled degradation should occur by "jumps." It has also some implications when simulating the process and looking for the hitting time of the failure level ℓ. You may consult the `book companion site` for more details. The homogeneous gamma process is studied further in Problem 12.8.

12.4.4 Shock Models

Assume that $Y(t)$ is explicitly established as a function of shocks. A shock is an event that can cause degradation or instantaneous item failure. Examples of items experiencing shocks are passive items such as switches and valves, that must act on demand. The impact of a demand on the item condition can be modeled as a shock. The time between two consecutive shocks, the damage caused by each shock, and the criteria for item failure (e.g. damage threshold, number of shocks with a given magnitude, time between shocks) are the three main characteristics of a shock model. Depending on the damage caused by the shock (either it is a continuous variable or a discrete one), the shock model may be defined in a continuous or discrete state space. Shock models are classified as extreme or cumulative shock models. A detailed review is provided by Nakagawa (2007).

In the first category, a single shock can cause item failure, whereas in the second category, each shock causes an additive damage to the item and failure occurs when the cumulative damage exceeds a given threshold. Beyond this classification, extreme and cumulative shock models may be mixed, and dependencies between arrival times and magnitude of the shocks can be introduced. We focus here on cumulative shock models because they are of interest for CBM strategies where the noise is considered to be negligible. Then we have access to $X(t)$.

A generic way to introduce cumulative shock models is to use a *marked point process*. The occurrence times of the shocks are denoted T_k, for $k = 1, 2, 3, \dots$, and are generally random variables. The instantaneous damage caused by the kth shock is defined by a variable D_k, which may be random and dependent on the (random) time T_k and is called a *mark*. Then, the process $\{T_k, D_k; k \geq 1\}$ is a marked point process.

Let $N(t)$ be a counting process representing the number of shocks in the time interval $(0, t]$, see Chapter 10. The cumulative damage $X(t)$ at time t is given by

$$X(t) = \sum_{k=1}^{N(t)} D_k. \tag{12.40}$$

The distribution function of $X(t)$ for $x > 0$ is defined by

$$\Pr(X(t) \leq x) =$$
$$\sum_{k=1}^{\infty} \Pr(D_1 + D_2 + \cdots + D_k \leq x \mid N(t) = k) \Pr(N(t) = k)$$

and for $x = 0$ by

$$\Pr(X(t) \leq 0) = \Pr(N(t) = 0).$$

If the increments are identically distributed with a given probability density function f and are independent from each other and from the process (T_k), then

$$\Pr(X(t) \leq x) = \Pr(N(t) = 0)\ I_{(x \leq 0)}$$
$$+ \sum_{k=1}^{\infty} \int_0^x (f)^{*(k)}(u)\, du\ \ \Pr(N(t) = k)\ I_{(x>0)}$$

where $(f)^{*(k)}$ is the k^{th} convolution of the probability density function f. This follows from the summation rule for k independent and identically distributed random variables with the same density f (see Chapter 10).

The Distribution of RUL(t_j)

Assume that the observed cumulative damage due to shocks at time t_j is m. If the failure level is ℓ, then the distribution of the RUL at time t_j is

$$\Pr(\text{RUL}(t_j) \leq t) = \Pr\left(\sum_{k=1}^{N(t_j+t)} D_k > \ell \mid \sum_{k=1}^{N(t_j)} D_k = m \right). \tag{12.41}$$

Then if $\ell - m > 0$

$$\Pr(\text{RUL}(t_j) \leq t) = \Pr\left(\sum_{k=N(t_j)+1}^{N(t_j+t)} D_k > \ell - m\right)$$

$$= \sum_{k=1}^{\infty} \int_{\ell-m}^{\infty} (f)^{*(k)}(x)\, dx \Pr(N(t_j + t) - N(t_j) = k). \qquad (12.42)$$

Examples of how to use shock models for PM optimization are given by Zhu et al. (2015) and Rafiee et al. (2015).

12.4.5 Stochastic Processes with Discrete States

When the state space is discrete or discretized, discrete state space degradation models can be used. The most common are continuous-time Markov chains (Markov processes) as described in Chapter 11. Several physical degradation phenomena have by nature discrete state space. An example is the high-voltage electrical motors for compressor systems in the oil and gas industry, which are monitored by the amount of partial discharges. The number of partial discharges decides the value of $X(t)$, and the guidelines recommend to define only four degradation states by putting thresholds on $X(t)$. It is also quite common that the degradation phenomenon is continuous, but the state space for $X(t)$ is discretized for convenience by guidelines. This is the case in civil engineering to define the condition of structures such as bridges: by using inspection reports and measures, the decision-maker is ranking a bridge between four degradation levels only.

Consider an item with n states, n is the new state and 0 is the failed state. Intermediate states from $n - 1$ to 1 are degraded states. For a time homogeneous degradation, the calculation of the distribution of RUL(t_j) requires to calculate the probability density function $\tilde{f}_n(x), \tilde{f}_{n-1}(x), \dots, \tilde{f}_1(x)$ of the sojourn times $\widetilde{T}_n, \widetilde{T}_{n-1}, \dots, \widetilde{T}_0$ in nonfailed states. If the monitoring is continuous and we know that the item enters the degraded state m at time t_j, then

$$\Pr(\text{RUL}(t_j) \leq t) = \Pr\left(\sum_{k=1}^{m} \widetilde{T}_k \leq t \mid X(t_j) = m\right)$$

$$= \int_0^t \tilde{f}_m * \tilde{f}_{m-1} * \cdots * \tilde{f}_1(x)\, dx. \qquad (12.43)$$

If the monitoring is not continuous, t_j may not coincide with the exact date upon which the item enters a degraded state. If the model is a Markov process, this does not matter because the only useful information is in which state the system is at time t_j, and (12.43) is still valid. If the model is not a Markov process, the calculation of the RUL distribution is more complicated because the time already spent in the current state at time t_j may influence the results.

12.4.6 Failure Rate Models

A last option that can be mentioned is the one based on time-dependent failure rate. Such models are used for imperfect maintenance policies, mainly to optimize the reduction of the failure rate, the conditional failure rate, or the virtual age after a failure has been repaired. Such models are discussed in Chapter 10, with focus on CM optimization (i.e. what do to after failure).

12.5 Condition-Based Maintenance

As an introduction and a motivation for CBM, consider a single item that may be maintained according to four different maintenance strategies: (i) only corrective replacements, (ii) age-based replacements, (iii) block replacements, or (iv) *ideal* replacements. Ideal replacement means that the item is preventively replaced just before failure. Ideal replacement is obviously not conceivable except for some very special cases.

By using the terminology and the notation in Section 12.3, the asymptotic cost per time unit of a strategy with only corrective replacements is

$$C_\infty = \frac{c+k}{\text{MTTF}}, \tag{12.44}$$

and the asymptotic cost per time unit for ideal maintenance is

$$C_\infty = \frac{c}{\text{MTTF}}. \tag{12.45}$$

To obtain numerical results, assume that $c = k = 50$ cost units and that the time-to-failure of the item is gamma distributed with MTTF = 375 time units and standard deviation 50 time units. The asymptotic cost per time unit for age and block replacements may be derived from (12.6) to (12.16), respectively, and are shown in Figure 12.8 as a function of t_0. The item is replaced preventively after t_0 time units in operation for the age replacement strategy and after t_0 calendar time units for the block replacement strategy. Optimal values for t_0 (i.e. the value of t_0 that minimizes the mean cost per time unit) can be determined for both strategies. The mean costs of the corrective and ideal strategies are constant values. The asymptotic costs for age and block replacements are always lower than the CM cost. In Figure 12.8, the asymptotic cost for block replacement is *seemingly* higher than the CM cost for large values of t_0, but this is due to an approximation error for high values of t_0 when calculating the cost. This error is explained in Section 12.3.2 and is related to the assumption made that only a single failure can occur before a preventive replacement.

The gap in Figure 12.8 between the minimal cost of the age-based strategy and the ideal strategy illustrates the maximum benefit we can hope for with a CBM

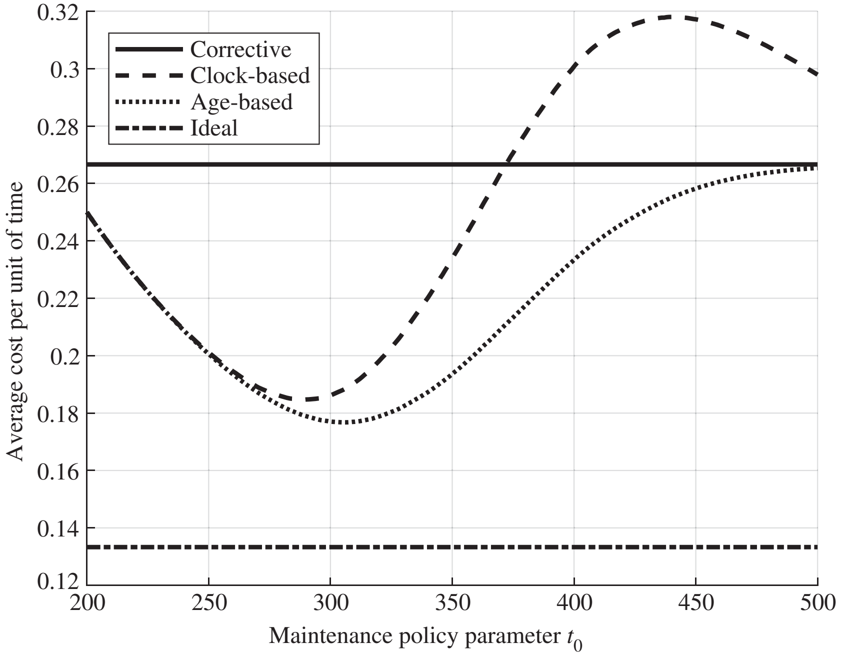


Figure 12.8 Comparison between four different (non-CBM) maintenance strategies.

strategy. The aim of the CBM strategy is, by monitoring, modeling, and predicting degradation phenomena, to plan preventive replacements that come as close as possible to the ideal strategy. The problem is that our monitoring, modeling, and prediction of degradations are not perfect, they have a cost, and they are introducing uncertainties. It is then important to quantify precisely the added value of CBM. A more detailed example is given by Zio and Compare (2013) and a more general discussion on the application of mathematical models in maintenance is given by Scarf (1997).

12.5.1 CBM Strategy

The main elements of a CBM strategy are as follows:

(1) A degraded state or a set of states for which a PM task is to be planned.
(2) A state or a set of states to which the item is put back after the PM task.
(3) A monitoring approach (continuous, inspection-based, or opportunistic) to determine the state of the item.

The models presented in this section are delimited to a single item (or to a uni-dimensional degradation model), and it is assumed that the actual degraded state $X(t)$ is directly measurable.

We provide an overview of different CBM models and distinguish between continuous monitoring and inspection-based monitoring. For continuous monitoring, the current degradation state $X(t)$ is assumed to be known at any time, and the parameters to optimize may include the following:

- The state in which a maintenance task is planned to be started.
- The state to which the item is put back after the maintenance task.
- The maintenance duration, if the maintenance cost and its efficiency are dependent on it.

For inspection-based monitoring, the current degradation state $X(t)$ is assumed to be known at inspection times only, and decisions related to maintenance are taken at these times. Parameters to optimize may include the previous list, plus the inspection dates/intervals.

In both cases, a generic degradation model is assumed with either discrete or continuous state space. The common assumptions to all the models are as follows:

- The monitoring is perfect meaning that the true state of the item is perfectly known, continuously or at inspection date.
- Each PM task brings the item from a degraded state to the as-good-as-new state or to a less degraded state (i.e. imperfect maintenance).
- The CM tasks always bring the item to the as-good-as-new state and have a higher cost per time unit than the preventive ones.
- The cost of PM tasks may increase with the degree of repair.
- The PM tasks may have a higher cost if the item is in a more degraded state when the PM task starts.
- It may be a penalty cost due to failure and possibly due to the sojourn time in degraded states. This penalty cost can be caused by loss of production.

12.5.2 Continuous Monitoring and Finite Discrete State Space

Consider a discrete degradation model $X(t)$ that is known at any time and that takes values in a discrete and finite state space. Several CBM strategies are available based on the following assumptions:

- The degradation $X(t)$ takes values in a discrete finite state space with n states. As an illustration, let $n = 4$.
- One of the states is considered to be as-good-as-new and one is regarded as failed. The other states are regarded as degraded.
- The degradation is gradual, meaning that the item moves from one state to the next more degraded state, until the failed state is reached.

Maintenance Strategies

Consider the state transition diagram in Figure 12.9, where state 3 is the as-good-as-new state, state 0 is the failed state, and states 2 and 1 are intermediate, degraded states. Transitions from state k to state $k - 1$, for $k = 1, 2, 3$, are related to the degradation phenomenon. The transition rate from state k to state $k - 1$ is denoted λ_k. Transitions to a state with a higher number are related to maintenance tasks. Transitions from state 0 are CM. The transition from state 0 to state 3 corresponds to a perfect repair (renewal), a transition from state 0 to state 2 is an imperfect repair and a transition from state 0 to state 1 may be seen as a minimal repair. The degree of repair for CM may be a parameter to optimize. When needed, a repair rate from state k to state $k + 1$ is denoted by $\mu_{k\ k+1}$, as shown in Figure 12.10.

The maintenance cost determines which of the PM strategies of Figure 12.10 that should be preferred. We may, for example, decide whether the PM task should start when the item enters state 2 or 1, and whether or not it should be repaired to the as-good-as-new state.

We start by highlighting some implicit assumptions related to the use of state transition diagrams in PM planning. Case 1 in Figure 12.10 may be regarded as a reference case with only CM, where the item is always repaired to the as-good-as-new state (i.e. state 3). A PM task may be modeled by a transition from state k to state $k + 1$, for $k = 1, 2$, with corresponding transition rates $\mu_{k\ k+1}$. We consider the PM strategies for the cases 2–5.

(1) When in state 1 or 2, the item can either degrade/fail or be maintained to a better/as-good-as-new state. In practice, this means that:
- The item is not taken out of operation while it is maintained or
- The time spent in state 1 or 2 corresponds to a delay (i.e. maintenance is planned but not started) and at the end of the delay, the item is put into the better (or as-good-as-new) state immediately (i.e. the maintenance duration is negligible compared to the delay).

(2) When assuming constant transition rates μ_{13} or μ_{12} in cases 4 and 5, this implies, due to the memoryless property of the exponential distribution, that if a PM task is planned or started in state 2 and the item degrades to state 1 before the maintenance is completed or the delay is over, the remaining time spent in state 1 does not depend on the time already spent in state 2 to preventively maintain the item. This may be acceptable in practice for some

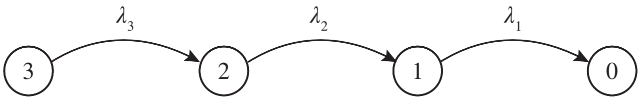


Figure 12.9 State transition diagram for a single item with degraded states.

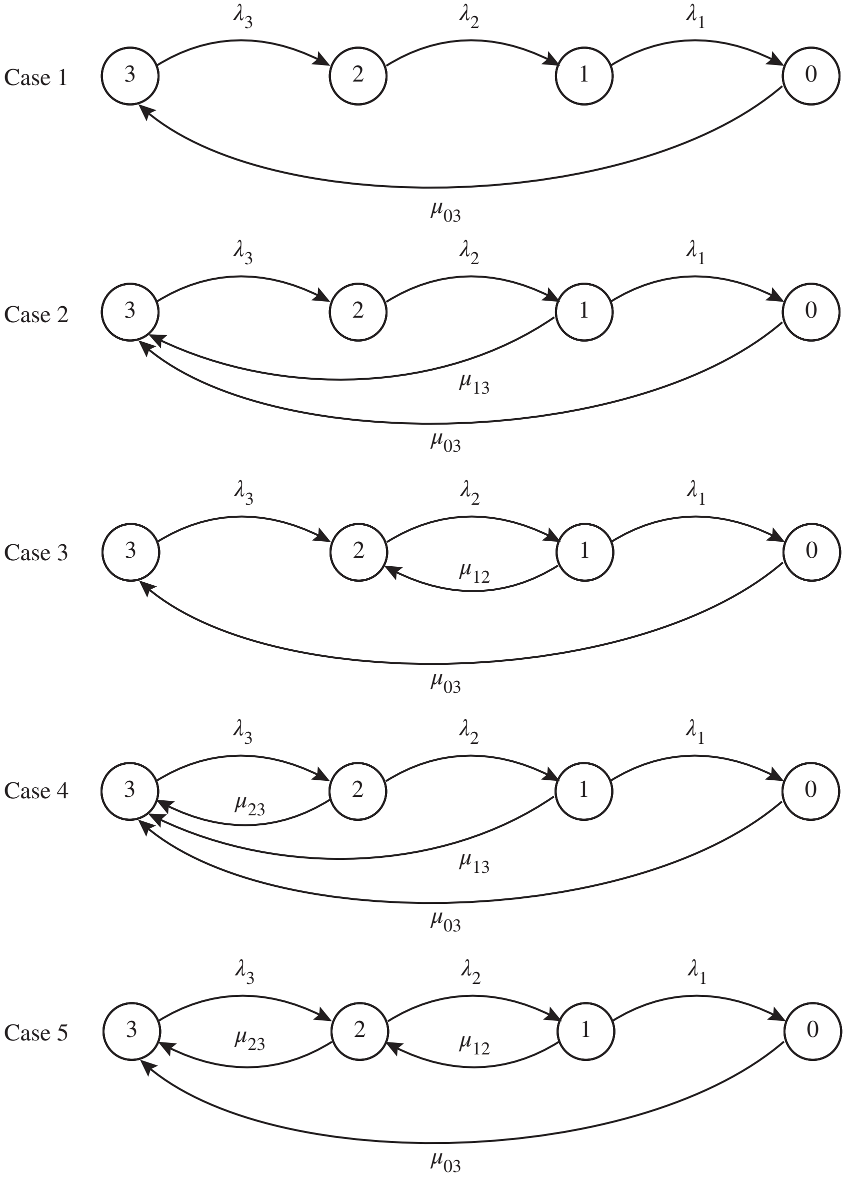


Figure 12.10 State transition diagram for a single item with degraded states and CBM.

cases but not for all, and the modeling of delays with constant transition rates can be questionable.

(3) When assuming constant transition rates λ_1, λ_2, λ_3, this implies, due to the memoryless property of the exponential distribution, that the RUL at time t_k does not depend on the time already spent in the state in which the item is observed at time t_k.

Example 12.7 (Degradation and maintenance of multicomponent systems)
Assume that the item is made of three identical components and one of the three is sufficient for the item to function as required. This means that the item can be modeled as a 1oo3:G parallel structure. In state 3, none of the components is failed, in state 2, one of them is failed and in state 1, two of them are failed. As soon as one component is failed (state 2), a repair task is started. Meanwhile, a second component can fail and the item is put to state 1 before it can be put back to state 3. In such a case, the graphs of Figure 12.10 make sense but it may not be realistic to consider constant repair rates for the two last strategies: the time spent in state 1 may depend on the time already spent in state 2 given the maintenance work already done for the first failed component. □

Example 12.8 (Degradation and maintenance of bridges)
An increasing number of modern bridges are continuously monitored, where sensor data provide an overall assessment of their structural health over time. These data are used together with other information sources to trigger decisions related to maintenance tasks. The state of the bridge is usually characterized by a finite number of degraded states going from as-good-as-new to unacceptable. The Norwegian Road Administration is currently using a scale with four states. If the bridge is diagnosed to be in state 2 or 1, PM is scheduled and the bridge is kept in operation. The maintenance strategy corresponds to case 4 in Figure 12.10. During a maintenance task, the bridge continues to degrade, but with a very low rate (the transition rate for the maintenance is much higher than the degradation rate), but if it degrades to a lower state, the maintenance tasks to renew the bridge are very different, such that the work done in the previous degraded state may be disregarded. In this case, the assumption of constant repair rates may be reasonable. □

Maintenance Cost

Let c_{ij} be the maintenance cost per time unit for bringing the item from state i to state j. According to the assumptions for the maintenance costs, $c_{03} \geq c_{13} \geq c_{23}$, $c_{13} \geq c_{12}$, $c_{12} \geq c_{23}$. Let γ_j be the penalty cost per time unit due to sojourn in a degraded or failed state j (e.g. due to loss of production). We have $\gamma_0 \geq \gamma_1 \geq \gamma_2$.

If all the transition rates are constant, the model describing the maintenance strategies are time-homogeneous Markov processes. The asymptotic cost per time unit C_∞ depends on the steady-state probabilities for each state. These probabilities can be obtained numerically by using the results given for homogeneous Markov processes in Chapter 11. They represent the mean time spent in each state per time unit (mean proportion of time). The cost C_∞^j for each case j in Figure 12.10 is determined as:

$$\begin{aligned}
C_\infty^1 &= c_{03}\mu_{03}P_0 + \sum_{i=0}^{2} \gamma_i P_i \\
C_\infty^2 &= c_{13}\mu_{13}P_1 + C_\infty^1 \\
C_\infty^3 &= c_{12}\mu_{12}P_1 + C_\infty^1 \\
C_\infty^4 &= c_{23}\mu_{23}P_2 + c_{13}\mu_{13}P_1 + C_\infty^1 \\
C_\infty^5 &= c_{23}\mu_{23}P_2 + c_{12}\mu_{12}P_1 + C_\infty^1
\end{aligned}$$

where P_i is the steady-state probability for state i, $\gamma_i P_i$ is the mean loss of production per time unit, and $\mu_{ij}P_j$ is the mean number of maintenance tasks per time unit from state i to state j. The numerical computation of the steady-state probabilities P_i and an example of a Python simulation algorithm are provided on the `book companion site` for each case. The numerical computation is valid only when all the transition rates are constant.

If at least one of the transition rates is not constant, Monte Carlo simulation should be used. The steady state may not exist anymore and other cost functions have to be used, such as the cumulative mean cost per time unit for a given time horizon t. Such a cost function depends on the mean sojourn time in each state within $[0, t]$. An example of a simulation algorithm is provided on the `book companion site` for Case 4 when the transition rate μ_{13} depends on the time spent in state 2. The outputs of the simulation algorithms present the mean time spent in each state, the mean number of failures, and the mean number of maintenance interventions (preventive and corrective) in a specified time horizon. Numerical examples are further studied in Problem 12.10.

It is possible to modify the state transition diagrams in Figure 12.10, such that the item is taken out of operation during maintenance. A transition diagram, which corresponds to case 4 in Figure 12.10, is shown in Figure 12.11, where 2^R, 1^R, and 0^R are the states where the item is under repair, d_0, d_1, d_2 are the waiting rates for a possible delays, and r_0, r_1, r_2 are the repair rates. If there is no delay, one can consider that transitions from states 2 and 1 to states 2^R and 1^R are immediate. If there is a delay, the transition can be decided after a deterministic or a random duration. In case of constant transition rates, a Markov process can be used. If not, a piecewise-deterministic Markov process (PDMP) is recommended. An example of a Python simulation algorithm is provided on the `book companion site`.

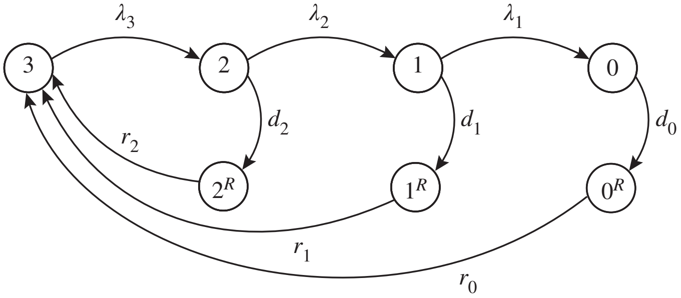


Figure 12.11 State transition diagram for a single component with degraded states taken out of operation during maintenance.

12.5.3 Continuous Monitoring and Continuous State Space

Consider a continuous degradation model $X(t)$ that is known at any time and that takes values in a continuous state space. We further assume that:

- The degradation level ℓ for which the item is considered to be failed is known.
- The repair duration is negligible.
- There is a delay before a maintenance task can be started, it is denoted τ and is deterministic.
- The item is degrading continuously and may fail within the maintenance delay.
- The maintenance tasks, whether corrective or preventive, are perfect, both return the item to the as-good-as-new state (i.e. a renewal).

These assumptions are realistic when the repair duration is short compared to the delay, and when the downtime may be significant because of the delay. This is true for systems that have a high reliability but are difficult to access, such as subsea production systems in the offshore oil and gas industry, offshore platforms, offshore wind farms, hydroelectric dams. For such systems, it is also reasonable to assume perfect CM and PM tasks because for many items, the material cost can be very low compared to the delay cost (including preparation and moving).

Maintenance Strategy

The following PM strategy may be realistic for items with continuous monitoring and continuous degradation: a preventive renewal is planned as soon as the degradation level reaches a given level m and the actual renewal is started after a delay τ. Meanwhile, the item can reach the failure level ℓ and stay in the failed state until the maintenance task is started. If so, the preventive renewal is replaced by a corrective one. The objective is to optimize the level m for which the preventive renewal is planned. Such a strategy has been extensively studied by Bérenguer et al. (2003) and Grall et al. (2006).

Maintenance Cost

With the given assumptions, the number of scenarios in a renewal cycle is very limited: either the item does not fail within the delay or it does. In the first case, the renewal cost is c_m, and in the second case, the renewal cost is c_ℓ plus the downtime cost per time unit γ multiplied with the downtime. The asymptotic mean cost per time unit is:

$$\begin{aligned} C_\infty =& \frac{c_m \Pr(T_\ell^{(\mathrm{h})} > T_m^{(\mathrm{h})} + \tau) + c_\ell \Pr(T_\ell^{(\mathrm{h})} \leq T_m^{(\mathrm{h})} + \tau)}{E(T_m^{(\mathrm{h})}) + \tau} \\ &+ \frac{\gamma E[(T_m^{(\mathrm{h})} + \tau - T_\ell^{(\mathrm{h})}) I_{(T_\ell^{(\mathrm{h})} \leq T_m^{(\mathrm{h})} + \tau)}]}{E(T_m^{(\mathrm{h})}) + \tau} \\ =& \frac{c_m \Pr(T_\ell^{(\mathrm{h})} > T_m^{(\mathrm{h})} + \tau) + c_\ell \Pr(T_\ell^{(\mathrm{h})} \leq T_m^{(\mathrm{h})} + \tau)}{E(T_m^{(\mathrm{h})}) + \tau} \\ &+ \frac{\gamma(\tau - E[\min(\tau, T_\ell^{(\mathrm{h})} - T_m^{(\mathrm{h})})])}{E(T_m^{(\mathrm{h})}) + \tau}, \end{aligned} \tag{12.46}$$

where:

- $T_m^{(\mathrm{h})}$ and $T_\ell^{(\mathrm{h})}$ are the hitting times of degradation levels m and ℓ, respectively.
- $T_R = T_m^{(\mathrm{h})} + \tau$ is the time interval between two renewals and $E(T_R) = E(T_m^{(\mathrm{h})} + \tau) = E(T_m^{(\mathrm{h})}) + \tau$ is the mean value.
- $E[(T_m^{(\mathrm{h})} + \tau - T_\ell^{(\mathrm{h})}) I_{(T_\ell^{(\mathrm{h})} \leq T_m^{(\mathrm{h})} + \tau)}]$ is the mean downtime in case of a failure within the delay, that is, when $T_\ell^{(\mathrm{h})} \leq T_m^{(\mathrm{h})} + \tau$. The term $I_{(T_\ell^{(\mathrm{h})} \leq T_m^{(\mathrm{h})} + \tau)}$ equals 0 if $T_\ell^{(\mathrm{h})} \geq T_m^{(\mathrm{h})} + \tau$ and 1 otherwise (indicator function). This means that the mean of the downtime $T_m^{(\mathrm{h})} + \tau - T_\ell^{(\mathrm{h})}$ is non-zero only when the failure occurs before the PM task is carried out.

The quantities $\Pr(T_\ell^{(\mathrm{h})} > T_m^{(\mathrm{h})} + \tau)$, $E(T_m^{(\mathrm{h})})$, and $E[\min(\tau, T_\ell^{(\mathrm{h})} - T_m^{(\mathrm{h})})]$ need to be determined. If the degradation process is monotonically increasing and homogeneous:

$$\begin{aligned} \Pr(T_\ell^{(\mathrm{h})} > T_m^{(\mathrm{h})} + \tau) &= \Pr(T_\ell^{(\mathrm{h})} - T_m^{(\mathrm{h})} > \tau) \\ &= \Pr(X(\tau) \leq \ell - m) \end{aligned}$$

and:

$$\begin{aligned} E(T_m^{(\mathrm{h})}) &= \int_0^{+\infty} \Pr(T_m^{(\mathrm{h})} > u)\, du \\ &= \int_0^{+\infty} \Pr(X(u) \leq m)\, du \end{aligned}$$

For nonmonotonic processes such as the Wiener processes and trend models, this is not true. It is a valid approximation if the nonmonotonicity may be neglected.

In case of a gamma process, which is monotonically increasing, we get:

$$\Pr(T_\ell^{(\mathrm{h})} > T_m^{(\mathrm{h})} + \tau) = \int_0^{\ell - m} f_{\alpha\tau,\beta}(x)\,dx,$$

$$E(T_m^{(\mathrm{h})}) = \int_0^{+\infty} \int_0^{m} f_{\alpha u,\beta}(x)\,dx\,du.$$

For $E[\min\,(\tau, T_\ell^{(\mathrm{h})} - T_m^{(\mathrm{h})})]$, we need the joint density function of $(T_m^{(\mathrm{h})}, T_\ell^{(\mathrm{h})})$. In case of monotonic degradation:

$$\begin{aligned} f_{T_m^{(\mathrm{h})}, T_\ell^{(\mathrm{h})}} &= \frac{\partial^2}{\partial u \partial v} \Pr(T_m^{(\mathrm{h})} > u, T_\ell^{(\mathrm{h})} > v) \\ &= \frac{\partial^2}{\partial u \partial v} \Pr(X(u) \le m, X(v) \le \ell) \end{aligned}$$

Then the survivor function of $(T_\ell^{(\mathrm{h})} - T_m^{(\mathrm{h})})$ is

$$\overline{G}(s) = \int_0^{+\infty} \int_{v+s}^{+\infty} f_{T_m^{(\mathrm{h})}, T_\ell^{(\mathrm{h})}}(u, v)\,du\,dv.$$

Finally,

$$E[\min\,(\tau, T_\ell^{(\mathrm{h})} - T_m^{(\mathrm{h})})] = \int_0^{\tau} \overline{G}(s)\,ds.$$

In case of a gamma process, we get:

$$\begin{aligned} f_{T_m^{(\mathrm{h})},\ T_\ell^{(\mathrm{h})}(u,v)} &= \frac{\partial^2}{\partial u \partial v} \Pr(T_m^{(\mathrm{h})} > u, T_\ell^{(\mathrm{h})} > v) \\ &= \frac{\partial^2}{\partial u \partial v} \Pr(X(u) \le m,\ X(v) \le \ell) \\ &= \frac{\partial^2}{\partial u \partial v} \int_0^m \int_0^{\ell - x} f_{\alpha u,\beta}(x)\, f_{\alpha(v-u),\beta}(y)\,dy\,dx \\ &= \int_0^m \int_0^{\ell - x} \frac{\partial^2}{\partial u \partial v} f_{\alpha u,\beta}(x)\, f_{\alpha(v-u),\beta}(y)\,dy\,dx \end{aligned}$$

because the increments are independent. Numerical examples are further studied in Problem 12.9.

12.5.4 Inspection-Based Monitoring and Finite Discrete State Space

With inspection-based monitoring, the item state is known at inspection dates, and the maintenance tasks can be triggered only at these dates. This is the case for many passive items such as valves, pipelines, vessels, any standby safety systems (e.g. fire or gas detectors), and many parts of a structure in civil engineering. All these items may not provide by themselves any signal that can be monitored continuously as a degradation indicator. They need to be activated or inspected to be diagnosed in a given state.

Consider a degradation model $X(t)$, which takes values in a finite discrete state space. Further, assume that:

- The degradation model is the same as the one used in Section 12.5.1.
- The item is inspected at deterministic dates $\tau_1, \tau_2, \tau_3, \dots$. At inspection, the item is taken out of operation and can be preventively maintained without any delay.
- The maintenance task durations are negligible (compared to the item lifetime), and there is no delay for intervention (the item is easy to access).
- The CM and PM tasks return the item to an as-good-as-new state.

These assumptions are realistic when the repair duration is very short compared to the item lifetime and when the delay before intervention at inspection date can be neglected. This is true for systems that are very reliable and easy to access once an inspection is launched. In addition, for many production systems, planned inspections, and associated maintenance tasks are often triggered by stopping or reducing the production process when the impact on the production is as low as possible or when the loss can be compensated by redundant systems. The production losses that have to be taken into account are mainly the ones due to unexpected failures between two inspections and not the ones due to stop of production at inspection dates.

Time-Based Inspections Versus Condition-Based Inspections

Inspections may be condition-based, meaning that the date of the next inspection is determined by the degraded state at the current inspection date. The formalism required in such a case is outside the scope of the book.

Assume that the inspections are fixed according to a calendar (they are usually periodic). This is the easiest case to model and it makes sense in practice, when some periods of time are more suitable to reduce the production rate, put some items out of operation and perform inspections or PM.

Maintenance Strategy

If the transition rates from a state k to a more degraded state $k-1$ are constant, the item degradation model between two inspections is a Markov process. The maintenance tasks can be modeled by a transition matrix $\mathbb{B}$ as explained in Section 11.11. The complete model including inspection and maintenance is a multiphase Markov process. An example is given below. Consider the notation in Figure 12.9. The transition rate matrix of the Markov process between two inspections is:

$$\mathbb{A} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \lambda_1 & -\lambda_1 & 0 & 0 \\ 0 & \lambda_2 & -\lambda_2 & 0 \\ 0 & 0 & \lambda_3 & -\lambda_3 \end{pmatrix}.$$

Let $\boldsymbol{P}(t) = [P_0(t), P_1(t), P_2(t), P_3(t)]$ be the time-dependent state probability vector, where $P_j(t)$ is the probability that the Markov process (degradation model) is in state j at time t $[P_j(t) = \Pr(X(t) = j)]$. The vector of state probabilities after the maintenance task at time τ_i is $\boldsymbol{P}(\tau_i)\mathbb{B}$, where $\mathbb{B}$ is a 4×4 matrix such that the sum of the entries in each line is 1. The entry $\mathbb{B}_{ij}$ in the matrix $\mathbb{B}$ is the probability that the item is in state j after the maintenance task, given that it was in state i just before the maintenance task is completed.

The matrix $\mathbb{B}$ depends on the maintenance strategy applied to the item after inspection. If, for example, the PM is triggered in state 1 and if all maintenance tasks are as-good-as-new, we have

$$\mathbb{B} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The analytical expression of the time-dependent state probability vector is given for $\tau_i \le t < \tau_{i+1}$ by

$$\boldsymbol{P}(t) = \boldsymbol{P}(0) \left(\prod_{k=1}^{k=i} e^{(\tau_k - \tau_{k-1})\mathbb{A}} \, \mathbb{B} \right) e^{(t-\tau_i)\mathbb{A}}. \tag{12.47}$$

Python scripts for the numerical computation and for Monte Carlo simulation of the multiphase Markov process are provided on the `book companion site`. If one of the transition rates is time-dependent, it is necessary to use a PDMP as described in Section 11.12.

Maintenance Cost

The process $X(t)$ has no steady state and the maintenance cost per time unit has to be calculated on a given finite time horizon by using the time-dependent state probabilities $\boldsymbol{P}(t)$. Consider, for example, the cumulative maintenance cost between two inspections in the time interval $(\tau_i, \tau_{i+1}]$. It includes the maintenance costs at time τ_{i+1}, that is to say c_{13} if the item is in state 1, or c_{03} if the item is in state 0. Because CM and PM are performed only at specified inspection dates, a single failure can occur between two inspections. The cumulative cost in $(\tau_i, \tau_{i+1}]$ is

$$\begin{aligned} C((\tau_i, \tau_{i+1}]) &= \frac{c_{03} \Pr[X(\tau_{i+1}) = 0] + c_{13} \Pr[X(\tau_{i+1}) = 1]}{\tau_{i+1} - \tau_i} \\ &= \frac{c_{03} P_0(\tau_{i+1}) + c_{13} P_1(\tau_{i+1})}{\tau_{i+1} - \tau_i} \end{aligned}$$

It is possible to consider a penalty cost due to the sojourn time in degraded or failed states. It is denoted γ_j, for $j = 1, 2, 0$. In this case, failures and degraded states may be detected by a loss of production, but PM and CM tasks are still performed at inspection dates. Then we have

$$C((\tau_i, \tau_{i+1}]) = \frac{c_{03}P_0(\tau_{i+1}) + c_{13}P_1(\tau_{i+1})}{\tau_{i+1} - \tau_i} + \frac{\sum_{j=0}^{2} \gamma_j \int_{\tau_i}^{\tau_{i+1}} \mathrm{sf}_{\tilde{T}_j}(\mathrm{s})\mathrm{ds}}{\tau_{i+1} - \tau_i}. \tag{12.48}$$

If there is an inspection or a maintenance duration that can be considered as deterministic and the item is stopped during maintenance tasks, the same model can be used with a time lag at inspection dates, see Section 11.11. If the inspection or maintenance durations are random with exponential density, a multiphase Markov process can be still used but with additional states and additional phases. The two last cases are studied in Problem 12.10.

12.5.5 Inspection-Based Monitoring and Continuous State Space

Consider a continuous degradation model $X(t)$ that is known only at inspection dates and that takes values in a continuous state space. The same assumptions as in Section 12.5.4 are made for the maintenance/inspection tasks, by considering a more generic framework where the inspection dates may be condition-based, such that the next inspection date is updated according to the item condition observed at the current inspection date.

Maintenance/Inspection Strategy

Consider the following PM strategy, which is realistic in case of inspection-based monitoring and a continuous degradation state space: a preventive renewal is started at the first inspection date upon which the degradation level is observed above a degradation level m, where $m < \ell$. The cost is denoted c_m. In addition, a corrective renewal is performed at the first inspection date upon which the degradation level is observed above the degradation level ℓ. The cost is denoted c_ℓ.The cost of the downtime per time unit is denoted γ when the item failed between two inspections. The next inspection is rescheduled after each inspection according to the current degradation level (periodic inspection strategy or calendar-based inspection strategy are special cases). Formally, the time of the next inspection T_{n+1} is random and is defined by $T_{n+1} = T_n + g(X_{T_n})$ where $g(\cdot)$ is a decreasing function from $[0, m)$ to R_+. Periodic inspections may be modeled by setting $g(X_{T_n}) = \tau$, where τ is a constant value that is independent of the item state and the time. This gives an optimization problem for m and the date of the next inspection.

Maintenance Cost

To calculate the maintenance cost requires studying the process $\{X(t), t \geq 0\}$ and the maintenance effects at inspection dates. Analytical developments of the maintenance cost are not straightforward and require knowledge that is outside the scope of this book. Consequently, we delimit the presentation to a brief formulation of the cost function and propose to evaluate it by providing a solution with Monte Carlo simulation. A Python script is available on the `book companion site`. More details on maintenance cost calculations are available in Grall et al. (2002) and Omshi et al. (2019).

Let $C(t)$ be the cost function taking into account the cost of each type of task as well as the cost of inactivity of the system between 0 an t,

$$C(t) = c_m N_m(t) + c_\ell N_\ell(t) + \gamma d(t),$$

where $N_\ell(t)$ is the number of corrective tasks, c_ℓ is the cost of one corrective task, $N_m(t)$ is the number of preventive tasks between 0 and t,c_m is the cost of one preventive task, $d(t)$ is the time of inactivity of the system between 0 and t, and γ is the downtime cost per time unit. Then we have

$$C_\infty = \frac{E[C(T_R)]}{E(T_R)} = \frac{E[c_m N_m(T_R) + c_\ell N_\ell(T_R)] + \gamma\ E[d(T_R)]}{E(T_R)},$$

where T_R is the renewal cycle. A renewal occurs at the first inspection time when the degradation level of the item is observed above level m. The optimization consists in finding the preventive repair threshold m and the inspection function $g(\cdot)$ for which C_∞ is minimum. For instance, if we choose a linear maintenance function $g(x) = m_{\max} - x\left(\frac{m_{\max}-m_{\min}}{m}\right)$, we must find the three numbers m, $m_{\max}$, and $m_{\min}$ for which C_∞ is minimum.

12.6 Maintenance of Multi-Item Systems

When addressing maintenance optimization problems at the system level, it is required first to model the system behavior and then to integrate maintenance effects into the system model. These two stages are reviewed in this section.

12.6.1 System Model

As described in Chapters 4–6, a system model must include:

- The system structure

- The stochastic behavior of each item
- The interactions between the items

For all these items, we refer to Chapters 6 and 11 and see them as an application of reliability theory to maintenance modeling.

Models for the System Structure

The system structure is a representation of how the items are combined to fulfill a main function at the system level, as defined and formalized in Chapters 2 and 4.

Stochastic Models for Single Items

A single item is modeled by a state space and by probability distributions to describe the sojourn time in each state. The two main classes of item models are as follows:

Time-to-failure models. The state space of the items is reduced to two states (functioning or failed). These models are presented in Chapters 3 and 5. They are widely used in practice because they often rely on a reasonable amount of data (failure dates). In this class, the exponential distribution has a particular place because it is used for items that do not experience any wear (i.e. having constant failure rate). For degrading items, it is possible to use other distributions, such as the Weibull distribution, with increasing failure rate.
With time-to-failure models for each item, the degradation at the system level may be interpreted as the number of failed items or the number of functioning states with some failed items.

Degradation models. Each item is described by a degradation model, either a discrete state space, a continuous one, or a mixture of both. The degradation at the system level may be defined by a multi-dimensional degradation process or by a scalar function of the degradation of the individual items.

Interactions Between Items

We usually distinguish three types of interactions between items.

Economic dependencies imply that the maintenance cost of a group of items is not equal to the sum of the maintenance cost of the individual items. There are two main cases:

- *Positive economic dependencies.* The maintenance cost of several items at the same time is lower than the sum of individual maintenance costs. This, for example, the case for a series structure where the maintenance cost may be reduced by sharing the maintenance preparation costs and by reducing the downtime of the system during the maintenance.
- *Negative economic dependencies.* The maintenance cost of several items at the same time is greater than the sum of individual maintenance costs. This is,

for example, the case of parallel structures where grouping the maintenance tasks can lead to increased system downtime.

These dependencies are modeled with the maintenance strategy and are part of the maintenance model.

Statistical dependencies (also called stochastic dependencies) occur when the stochastic behavior of some of the items are dependent, which means that their failure dates are dependent in the sense given by probability theory. These dependencies are studied in Chapter 8. In practice, this corresponds to situations where the items are subjected to the same harsh environment or shocks (e.g. common-cause failures), when the items share the same load and the failure of an item implies a redistribution of the load (load sharing), or when the failure of one item can trigger the failure of other items (i.e. cascading failures). Such dependencies are modeled with the stochastic behavior of each and every item.

Structural dependencies occur when an item cannot be maintained without impacting other items. This is a major challenge, for example, for systems that are very difficult to access and should be designed as compact as possible such as in subsea industry, aerospace industry, or nuclear industry. When designing such systems, it is important to split them into suitable modules that are stacked on top of each other and to place the least reliable modules on the top of the stack. This process is called stacking and is among the most challenging parts of a system development project. Observe also that the smaller the modules, the more connectors are needed and the higher the total failure rate. The way the stack is designed will have a major influence on the way the system can be maintained and the maintenance of one module can imply to retrieve some others. These dependencies have to be part of the maintenance model.

12.6.2 Maintenance Models

PM at the system level means that maintenance tasks are performed before the whole system is failed. In some ways, the number of failed items, their importance factors, and their own degradation levels (if any) can be used to define the state of the system and to decide on PM tasks.

Opportunistic Maintenance and Grouping

Opportunistic maintenance is only relevant in cases where the PM task requires the system or some subsystem to be shut down. Opportunistic maintenance is also called *opportunity maintenance*. Opportunities occur when the system is shut down due to production or administrative reasons, or when failures occur that either shut down the system or that require system shutdown during the corrective repair task.

Definition 12.2 (Opportunistic maintenance)
Maintenance task that is deferred or advanced in time and is performed when an unplanned opportunity occurs. □

One of the most developed areas for multi-item maintenance models is dedicated to economic dependencies and grouping strategies to optimize opportunistic maintenance. They are not developed here. The main issue is to optimize groups of items that are maintained preventively or correctively at the same time to save setup costs. This is, most of the time, a discrete optimization problem with time-to-failure models for each item. When the PM tasks on some items are performed at a failure date, we speak about an opportunistic maintenance strategy. This is of interest in series structures, when the repair of one failed item stops the whole system production. It may then be worthwhile to perform PM on some other items at the same time (named *grouping*). Such strategies may give setup-cost savings.

For a series structure, theoretical results show that the optimal grouping strategy is among a limited set of possible ones. Considering that all the items are stochastically independent, the PM date is first optimized individually for each item. Then, at a given possible time of maintenance, the optimal group for PM tasks is among those with the nearest optimal PM dates obtained individually. What has to be optimized is the number of items to be grouped. But if there is redundancy in the system structure, these results are not true anymore because grouping PM tasks can have a bad side effect by reducing the whole system's reliability or availability. In this case, it is required to list all the possible ways to group maintenance tasks. With a high number of items, exact solutions are not tractable within a reasonable time and have to be replaced by heuristics. Maintenance grouping optimization with a fixed grouping schedule is discussed by Cho and Parlar (1991), Dekker and Scarf (1997), and Nicolai and Dekker (2008), and dynamic grouping schedules are discussed by Do and Barros (2017) and Vu et al. (2018). For further information on opportunistic maintenance, see Bouvard et al. (2011) and Shafiee et al. (2015).

Condition-Based Maintenance

When the items are independent and the CBM tasks are decided at the item's state level, the tools described in Chapter 6 may be used in combination with item models. Each item may, for example, be modeled by an isolated gamma process or Markov process representing *independent* degradation processes and isolated maintenance strategies. Next, the system availability is calculated as described in Chapter 6, by combining the item availabilities with a structure function. On the other hand, when the items are dependent or when the maintenance tasks are decided at the system level (meaning that the tasks are decided according to the system state and not for each item) such an approach is not adequate. A review of CBM for systems with multiple dependent components is provided in Keizer et al. (2017). There is an extensive literature dedicated to prognostics and CBM for multi-item systems. Several examples are provided by Castanier et al. (2005), Deloux et al. (2016), and Zhang et al. (2019).

The current literature distinguishes three main modeling frameworks to address CBM at the system level:

Scenario-based approaches rely on describing all the possible scenarios that may occur between two renewals of the system. This approach may be used when the number of scenarios is very limited (often 2) with stochastic dependencies, or when the number of items is high but where the items are independent such that the scenarios can be captured by the structure function. Such models are used to identify if we have a renewal process, a semi-regenerative renewal process, or a Markov renewal process. The scenarios are built case by case, and may require a theoretical basis that is beyond the scope of the book. Section 12.6.3 describes what can be done with simple tools when the items are independent.

State-transition approaches as described in Chapter 6 with Monte Carlo simulation, or with Markov processes and their extensions (e.g. piecewise deterministic Markov processes) as described in Chapter 11. They rely on a description of the system states and the possible transitions between them. This may be a solution when the number of items is still limited (it should be possible to list "by hand" all the system states required to calculate the cost function) and can be useful for stochastic or structural dependencies. The approach may be used when the number of scenarios is too high to be exhaustively listed in a scenario-based approach. Such models can be a basis for numerical computation of Kolmogorov equations when a Markov process is used or for building a discrete event simulation algorithm with Monte Carlo simulation.

Dedicated modeling languages do not require the analyst to give a hand-made descriptions of all the system states. The idea is to structure the modeling work to handle a higher number of system states than with the two previous approaches. The true system states are automatically computed from a higher level description of the system or at least from a structured description of subsystems. The state probabilities are usually estimated with Monte Carlo simulation. Many modeling languages and associated software programs are available for this approach. A brief list is provided on the `book companion site`.

12.6.3 An Illustrative Example

Consider a safety-instrumented system (SIS) that is represented by the RBD in Figure 12.12. A thorough introduction to SISs is given in Chapter 13.

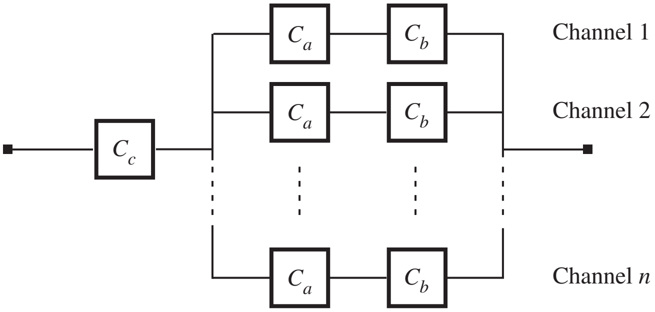


Figure 12.12 RBD of a safety-instrumented system (SIS).

- The redundant structure corresponds to the mechanical part of the system (actuators). In this structure, each path can be seen as a channel with several items. There are n channels and every channel i $(1 \leq i \leq n)$ may have two types of failure modes that can be modeled by two items in series (with indices a and b, respectively).
- The item with index c, in series with the redundant actuators, corresponds to a logic solver.

The following assumptions are made:

- Partial inspections (partial tests) are performed after time intervals of length Δ, at times $\tau_1, \tau_2, \dots, \tau_{m-1}$, such that $\tau_k = k\Delta$. The whole system is renewed at the end of regular intervals of length $\tau = m\Delta$.
- During the partial tests, failure modes associated to index a are detected, and appropriate maintenance tasks are planned. The tasks can correspond to complete renewal, renewal in case of an item failure, or imperfectly maintained (preventively or correctively). Failures of item b remain undetectable.
- The item with index c is continuously monitored. Its failure is supposed to be immediately detected due to embedded self-diagnostic functions.
- The item with index c is not exposed to degradation. It is modeled by a time-to-failure model with constant failure rate λ_c.
- The items with indices a and b are exposed to degradation. Items a are modeled by a discrete state degradation models and items b are modeled by a time-to-failure model with time-dependent failure rate function.
- The time to repair item c can be taken into account as a constant value m_c.
- The time to repair items of type a can be neglected or be a constant equal to m_a.
- All the items are supposed to be independent regarding structural or stochastic dependencies.

Degradation Model for Items of Type *a*

Assume that failures of items of type a are caused by degradation. The degradation is modeled by a discrete state Markov process with $\kappa + 1$ states. State κ is the new state at time $t = 0$, states $1, \dots, \kappa - 1$ are functioning states where the degradation increases with the decreasing state number, and state 0 is the failed state. The degradation of a type a item between two renewals is modeled by a discrete state Markov process with transition $\lambda_{a,0} = 0$ from state 0 to any other state, and transition rates $\lambda_{a,k}$ from state k to state $k - 1$, for $k = 1, \dots, \kappa$. The transition matrix is given by:

$$\mathbb{A} = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ \lambda_{a,1} & -\lambda_{a,1} & 0 & \dots & 0 & 0 \\ 0 & \lambda_{a,2} & -\lambda_{a,2} & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_{a,\kappa-1} & -\lambda_{a,\kappa-1} & 0 \\ 0 & 0 & 0 & \dots & \lambda_{a,\kappa} & -\lambda_{a,\kappa} \end{pmatrix}. \tag{12.49}$$

CBM Strategy

During one partial inspection, items of type a can be renewed systematically, renewed if they are failed, or imperfectly maintained (preventively or correctively). To model these strategies, define a matrix $\mathbb{B}$ of entries $\mathbb{B}_{k,j}$, where $\mathbb{B}_{k,j}$ is the probability that an item in state k, just before the maintenance will be in state j after the maintenance task. Observe that $\sum_{j=1}^{K} \mathbb{B}_{k,j} = 1$.

If item a is renewed at each inspection, then for any k, $\mathbb{B}_{k,\kappa+1} = 1$, and $\mathbb{B}_{k,j} = 0$ for $j \neq \kappa + 1$. If the item is renewed only when a failure occurs, then $\mathbb{B}_{1,\kappa+1} = 1$ and $\mathbb{B}_{k,k} = 1$, for $k \neq 1$. For imperfect PM, all the cases can be considered with, for example, $\mathbb{B}_{k,j} = 1$ for any $k \geq m$ and for a given $j < k$. Then $m - 1$ is the PM threshold above which a PM has to be performed and $j - 1$ is the degradation level after imperfect maintenance.

Maintenance Cost

Consider the system performance within the renewal time interval $[0, \tau)$ and let $\mathcal{F}$ be the set of the system functioning states. The system states are denoted $\eta = (\eta_1, \dots, \eta_n)$ with $\eta_i = 1$ if channel i is functioning and $\eta_i = 0$ if it is failed.

We calculate the system availability $A_S(t)$ in the interval $[0, \tau)$ when the time to repair an item of type a is negligible. Let $A_c(t)$ be the availability at time t of item c and $A_i(t)$ the availability of channel i. The availability $A(t)$ of the whole system is given by:

$$A_S(t) = A_c(t) \sum_{\eta \in \mathcal{F}} \prod_{i=1}^{n} [A_i(t)]^{\eta_i} [1 - A_i(t)]^{1-\eta_i}. \tag{12.50}$$

If item c is not available at time t it means that a self-diagnosed failure occurred in the time interval $[t - m_c, t)$. Because the occurrences of these failures are modeled by an HPP with rate λ_c, this event occurs with a probability $1 - e^{-\lambda_c m_c}$, such that

$$A_c(t) = e^{-\lambda_c m_c}. \tag{12.51}$$

The availability $A_i(t)$ for channel i is

$$A_i(t) = R_b(t) A_a(t), \tag{12.52}$$

where $A_a(t)$ is the availability of item a at time t and $R_b(t)$ is the survivor function of item b (recall that failures of item b are not detectable).

To calculate $A_a(t)$, we describe the degradation of a type a item. If $t \in [\tau_k, \tau_{k+1})$, its state probability vector $\boldsymbol{P}(t) = [P_0(t), \dots, P_\kappa(t)]$ equals:

$$\boldsymbol{P}(t) = \boldsymbol{P}(0)(e^{\Delta \mathbb{A}} \mathbb{B})^k e^{(t-\tau_k)\mathbb{A}},$$

where $P_j(t)$ is the probability that a type a item is in state j. Then $A_a(t)$ is obtained by summing the $P_i(t)$'s corresponding to functioning states.

We can now calculate the availability. By denoting $N(\eta)$, the number of channels in a functioning state in configuration η, the availability is given by:

$$\begin{aligned} A_S(t) &= A_c(t) \sum_{\eta \in \mathcal{F}} [R_b(t) A_a(t)]^{N(\eta)} [1 - R_b(t) A_a(t)]^{n-N(\eta)} \\ &= A_c(t) \sum_{\eta \in \mathcal{F}} \sum_{j=N(\eta)}^{n} (-1)^{j-N(\eta)} \, C_{n-N(\eta)}^{j-N(\eta)} [R_b(t) A_a(t)]^j \end{aligned}$$

The average availability in $(0, \tau)$ is

$$\begin{aligned} A_{\text{av}}(0, \tau) &= \frac{1}{\tau} \int_0^\tau A_S(s) \, ds \\ &= \frac{e^{-\lambda_c m_c}}{\tau} \sum_{i=1}^{m} \sum_{\eta \in \mathcal{F}} \sum_{j=N(\eta)}^{n} (-1)^{j-N(\eta)} \, C_{n-N(\eta)}^{j-N(\eta)} \int_{t_{i-1}}^{t_i} [R_b(s) A_a(s)]^j \, ds \end{aligned}$$

This formulation can be easily extended with a time lag, when the time to repair type a items is constant and equal to m_a. This means that after an inspection and a maintenance task, the system is restarted at time $\tau_k + m_a$.

Various maintenance costs can be calculated by using $A_S(t)$. For example, for a given matrix $\mathbb{B}$ and the corresponding costs related to the maintenance/inspection tasks, the benefits in terms of availability can be evaluated. For safety instrumented systems, what is important is that the system is available when required. A cost function is usually not considered at such, but the average availability in a given period of time $A_{\text{av}}(0, \tau)$ should be kept above a specified safety limit, at the lowest possible maintenance cost.

12.7 Problems

12.1 Consider an item that is replaced with a new item of the same type after regular intervals of length τ. If the item fails within a replacement interval, it is repaired to an as-good-as-new state. Show that the limiting availability A of the item does not exist.

12.2 An item has constant failure rate $\lambda = 5 \times 10^{-4}\ \mathrm{h}^{-1}$. When the item fails, it is repaired to an as-good-as-new state. The associated mean downtime is six hours. The item is supposed to be in continuous operation.

(a) Find the average availability A_{av} of the item.

(b) How many hours per year will the item on average be out of operation?

12.3 A machine with constant failure rate $\lambda = 2 \times 10^{-3}\ \mathrm{h}^{-1}$ is operated 8 h/d, 230 d/yr. The mean downtime required to repair the machine and bring it back into operation is MDT = 5 hours. The machine can only fail during active operation. If a repair task cannot be completed within normal working hours, overtime will be used to complete the repair, such that the machine is available next morning.

(a) Determine the average availability of the machine (during the planned working hours)

(b) Determine the average availability of the machine if the use of overtime were not allowed

12.4 An item is exposed to wear and has failure rate function $z_1(t) = \beta t$.

(a) Determine the survival probability $R(t)$ of the item at time $t = 2000$ hours, when $\beta = 5 \times 10^{-8}\ \mathrm{h}^{-2}$.

The item is to be overhauled after regular intervals of length τ. Assume that the overhaul will reduce the failure rate and that we may use the following model:

$$z(t) = \beta t - \alpha k\tau \qquad \text{for } k\tau < t \leq (k+1)\tau,$$

where k is the number of overhauls after time $t = 0$.

(a) Draw a sketch of $z(t)$. Explain what is meant by the term $\alpha k\tau$. Do you consider this model to be realistic?

(b) Determine the survivor function $R(t)$ at time $t = k\tau$, which is just before overhaul number k. Draw a sketch of $R(t)$ as a function of t.

(c) Find the conditional probability that the item is functioning just before overhaul $k+1$, when you know that it was functioning just before overhaul k.

12.5 Consider the age replacement strategy, and find the mean time between actual item failures, $E(Y_i)$ when the distribution of the time-to-failure T of the item has

(a) An exponential distribution with failure rate λ. Give a "physical" interpretation of the result you get.

(b) A gamma distribution with parameters $(2, \lambda)$.

12.6 Consider the block replacement strategy that is described in Section 12.3.2, and find the optimal number of spares when the cost of a spare, c_s per spare item and per time unit is included.

(a) Determine the optimal average maintenance cost including the average spare cost as a function of the block replacement interval t_0 and the number m of spare items.

(b) Plot the curve of the optimal average maintenance cost as a function of m. Assume that the time-to-failure T has a gamma distribution with parameters (α, β). Select realistic values for the necessary input parameters and generate the plot.

12.7 Consider an item that is inspected every 15 months ($p = 15$) and the total number of inspections is 35 ($n = 35$) when the first inspection is "fake" because the system is new and therefore in a perfect state. Assume that the system is experiencing degradation, that this degradation phenomenon depends on time and is, in essence, deterministic and monotonically increasing. The randomness is only due to measurement noise inherent to the inspection. We first simulate the degradation (i.e. we generate a "toy" dataset), and then estimate the parameters of the degradation model on this dataset. We are in the ideal situation where the model we use for the simulation is the one we use for the estimation.The degradation X depends on the time according to the following equation:

$$X(t) = 0.001\ t + 0.001\ t^2.$$

The observation $Y(t)$ of $X(t)$ is defined as $Y(t) = X(t) + \epsilon(t)$ where $\epsilon(t)$ is a Gaussian noise (i.e. a normally distributed noise with mean and standard deviation equal to 0 and 100, respectively).

(a) The first step is to create a script that generates one history (i.e. the samples from $t = 0$ to $t = p\ n$) of the degradation. To do so:

- Define a vector of times for inspections;

- Calculate for each time the corresponding actual degradation;
- By using Monte Carlo simulation, simulate the noise associated to the degradation measurement.

(b) The second step is to estimate the degradation parameters by using the dataset. To do so:

- Create a script that estimates parameters of polynomial time depending on degradation model with least square method.
- Evaluate, visually, the prediction quality by plotting several paths of the theoretical model and the ones of the estimated model beyond the last inspection, until a failure level $\ell = 5000$.
- By using these prediction paths, compare the empirical probability functions (the histograms) of the hitting time of degradation level ℓ obtained with the theoretical model and the estimated one. They correspond to the empirical probability functions of the RUL at the time of the last inspection. Plot the two empirical cumulative distributions.

(c) Modify the parameters to study the impact of the inspection number and/or the inspection period on the estimation quality.

12.8 Consider an item that is inspected every $p = 15$ months where the total number of inspections is $n = 6$. Assume that it is experiencing degradation between two inspections and that the phenomenon is, in essence, stochastic and monotonically increasing: The increments of degradation between two inspections have some randomness but they are always positive.

We first simulate the degradation (i.e. we generate a "toy" dataset), and then estimate the parameters of the degradation model based on this dataset. We are in the ideal situation where the model used for the simulation is the one we use for the estimation. We want a script that allows having a gamma or an exponential distribution of the increments.

(a) The first step is to create a script that generates $m = 50$ samples (named also histories or paths) of a degradation process with degradation increments following an exponential density or a gamma probability density function between two inspections. To do so:

- Use Monte Carlo simulation to simulate the increments: $n - 1$ increments for every path, such that we get $(n - 1)$ $x(m)$ increments. Store these increments in a matrix of size $(n - 1,m)$. Plot your dataset.
- Build the paths by summing the increments related to each path.

(b) The second step is to estimate the parameters used for simulating the dataset, by using the dataset itself. To do so:

- Consider that you observed the n simulated paths at inspection dates. Transform these paths into increments.
- Estimate the parameters of the probability density function of the increments by maximizing the likelihood function.
- In order to evaluate the quality of the estimates, plot on the same graph, the probability density function obtained by the estimated parameters, with the true parameters (the one used to simulate data), and the histogram corresponding to the dataset (i.e. the histogram of degradation increments).

(c) Modify the parameter values to study the impact of the number of path, number of inspections, inspection period on the estimation quality.

(d) For a true dataset (not simulated), how can we decide whether the degradation increments are following an exponential or a gamma distribution?

12.9 The degradation of an item can be discretized in four levels (level 1 is new, level 4 is failed) and the degradation level is continuously known. Assume that a maintenance action can begin without any delay.

(a) List all the possible maintenance strategies, preventive, and corrective.

(b) What assumption(s) do you need to make to model the maintained item as a Markov process?

(c) Assume that you choose a maintenance strategy for which corrective maintenance and preventive maintenance return the item to an as-good-as-new state, and a PM is only performed when the item is in the degradation level 3. Determine the state transition diagram and the corresponding transition rate matrix when the transition rates for the degradation phenomenon are all equal to 10^{-4} h^{-1}, the preventive repair rate equals 2×10^{-2} h^{-1} and the corrective repair rate equals 10^{-2} h^{-1}.

(d) Calculate the item availability as a function of t and make a plot of the availability.

(e) Calculate the MDT (the item is considered to be failed only when degradation level 4 is reached)

(f) Discuss (without doing any calculation) how you could choose the best maintenance policy among all the possible ones, if the criteria to optimize is the asymptotic cost per time unit. List all the assumptions and the parameters you need.

12.10 Consider an item that experiences a degradation phenomenon modeled by a gamma process with parameters α and β. The degradation is then modeled by a scalar indicator that is continuously increasing. The monitoring is continuous and ℓ is the degradation level that is considered to be the failure level. We want to introduce a PM strategy such that a PM task is planned when the degradation reaches a level m, such that $m < \ell$. The optimization problem is then: “What is the optimal value of m given that there is a delay τ between the time when the maintenance is planned and the time when the maintenance is actually started? Such delays can be due to time to prepare, gather the maintenance team, come to the spot, etc.” The assumptions are:

- The cost of item replacement (whether corrective of preventive) is c,
- The cost of downtime per time unit is γ,
- The maintenance duration, once the maintenance is started, is negligible,
- The repair (preventive or corrective) brings the item to an as-good-as-new state.

(a) Recall the definition of the gamma process.

(b) Derive the formula that gives the mean asymptotic cost per time unit, identify the renewal cycle and the quantities you need to assess.

(c) Make a script that simulates the gamma process and the maintenance strategy.

(d) Use the corresponding script to approximate the quantities of interest by running a sufficient number of Monte Carlo simulations.

(e) Use the script to “play around” with different parameters. The initial values you should return after every question are as follows: $\alpha = 9$, $\beta = 0.5$, $\ell = 500$, $m = 400$, $\tau = 2.5$ hours, $c = 1000$, $\gamma = 10\ 000$.

   i. Make variations of α and discuss the impact on the histograms of $T_m^{(h)}$ (hitting time of level m) and of $T_\ell^{(h)}$ (hitting time of level ℓ).

   ii. Make variations of m below and above 400 and discuss the value of the cost, the downtime. Can you identify an optimal “region” for the value of m?

   iii. Make variations of τ below and above 2.5 and discuss the value of the cost, the downtime.

   iv. Make variations of γ below and above 10 000 and discuss the “optimal region” for m.

(f) Check you results with analytical solutions in the special case when the gamma distribution is replaced by an exponential distribution for the degradation increments.

12.11 Consider an item with a single degraded state (state 2 is as-good-as-new, state 1 is degraded, state 0 is failed). The degradation is gradual, meaning that the item moves from state 2 to state 1 (with transition rate λ_{21}) and then from state 1 to state 0 (with transition rate λ_{10}). Numerical values per hour: $\lambda_{21} = 10^{-4}$, $\lambda_{10} = 10^{-3}$.

(a) Assuming that the monitoring is continuous:

(i) List all the possible maintenance strategies we can implement.

(ii) Choose one maintenance strategy and explain which model you can use to calculate the probability to be in each state at any time. Numerical values per hour: $\mu_{01} = 1$, $\mu_{12} = 1$, $\mu_{02} = 0.1$.

(iii) Provide some performance indicators we can derive from the probability to be in each state at any time.

(iv) Assume that two of these items are put in a parallel structure and are independent, and calculate the availability of the system for your maintenance strategy at times 1000, 10 000, 50 000 hours.

(b) Assume that the monitoring is inspection-based and that an inspection is performed every month:

(i) Explain which model you can use in order to give the probability for the item to be in each state at any time between two inspections. Apply it for the following maintenance strategy: (i) a PM task is performed at inspection date if the item is found in state 1 and the item is put back to the new state immediately, (ii) a corrective maintenance is performed at inspection date if the item is found in state 0 and the item is put back to the new state immediately. Assume that all the repair durations are negligible. Give the item availability at times 1000, 10 000, 50 000 hours.

(ii) Calculate the availability of the item at steady state by using Monte Carlo simulation.

(iii) Modify the previous model to take into account repair durations at inspection dates, assuming that they are random with constant repair rates. Numerical values per hour: $\mu_{01} = 1$, $\mu_{12} = 1$, $\mu_{02} = 0.1$.

(iv) Modify the previous model to take into account repair durations at inspection dates, assuming that they are constant. Numerical values in hours: $r_{01} = 1$, $r_{12} = 1$, $r_{02} = 10$.

12.12 Assume that you have two items in a redundant structure and that each item is experiencing a continuous degradation which is discretized according to four levels (level 1 is new, level 4 is failed). We want to implement a preventive CBM given that the items are periodically inspected at the same time, their degradation level is only known at inspection time and there is no delay before intervention.

(a) Which maintenance strategies can we consider? List at least two strategies.

(b) Describe the modeling process you would follow to optimize the choice of one of the maintenance strategy.

References

Barlow, R.E. and Hunter, L. (1960). Optimal preventive maintenance policies. *Operations Research* 8 (1): 90–100.

Bérenguer, C., Grall, A., Dieulle, L., and Roussignol, M. (2003). Maintenance policy for a continuously monitored deteriorating system. *Probability in the Engineering and Informational Sciences* 17 (2): 235–250.

Bouvard, K., Artus, S., Bérenguer, C., and Cocquempot, V. (2011). Condition-based dynamic maintenance operations planning & grouping. Application to commercial heavy vehicles. *Reliability Engineering & System Safety* 96 (6): 601–610. https://doi.org/10.1016/j.ress.2010.11.009.

Castanier, B., Grall, A., and Bérenguer, C. (2005). A condition-based maintenance policy with non-periodic inspections for a two-unit series system. *Reliability Engineering & System Safety* 87 (1): 109–120.

Cho, D.I. and Parlar, M. (1991). A survey of maintenance models for multi-unit systems. *European Journal of Operational Research* 51 (1): 1–23.

Christer, A.H. and Waller, W.M. (1984). Delay-time models of industrial inspection maintenance problems. *Journal of Operational Research* 35 (5): 401–406.

Christer, A.H. and Wang, W. (1992). Model of condition monitoring of a production plant. *International Journal of Production Research* 30: 2199–2211.

Dekker, R. (1996). Application of maintenance optimization models: a review and analysis. *Reliability Engineering & System Safety* 51: 229–240.

Dekker, R. and Scarf, P.A. (1997). On the impact of optimisation models in maintenance decision making: the state of the art. *Reliability Engineering & System Safety* 60 (2): 111–119.

Deloux, E., Fouladirad, M., and Bérenguer, C. (2016). Health and usage based maintenance policies for a partially observable deteriorating system. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability* 230 (1): 120–129.

Deng, Y., Barros, A., and Grall, A. (2016). Degradation modeling based on a time-dependent Ornstein-Uhlenbeck process and residual useful lifetime estimation. *IEEE Transactions on Reliability* 65 (1): 126–140.

Do, P. and Barros, A. (2017). Maintenance grouping models for multicomponent systems. In: *Mathematics Applied to Engineering* (ed. M. Ram and J.P. Davim), 147–170. Academic Press.

Ghamlouch, H., Fouladirad, M., and Grall, A. (2018). Prognostics for non-monotonous health indicator data with jump diffusion process. *Computers & Industrial Engineering* 126: 1–15.

Grall, A., Dieulle, L., Bérenguer, C., and Roussignol, M. (2002). Continuous-time predictive-maintenance scheduling for a deteriorating system. *IEEE Transactions on Reliability* 51 (2): 141–150.

Grall, A., Dieulle, L., Bérenguer, C., and Roussignol, M. (2006). Asymptotic failure rate of a continuously monitored system. *Reliability Engineering & System Safety* 91 (2): 126–130.

Huynh, K.T., Castro, I.T., Barros, A., and Bérenguer, C. (2014). On the use of mean residual life as a condition index for condition-based maintenance decision-making. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 44 (7): 877–893.

ISO 14224 (2016). *Petroleum, petrochemical, and natural gas industries: collection and exchange of reliability and maintenance data for equipment*, *International standard*. Geneva: International Organization for Standardization.

Kahle, W., Mercier, S., and Paroissin, C. (2016). *Degradation Processes in Reliability*. Hoboken, NJ: Wiley.

Keizer, M.C.A.O., Flapper, S.D.P., and Teunter, R.H. (2017). Condition-based maintenance policies for systems with multiple dependent components: a review. *European Journal of Operational Research* 261 (2): 405–420.

Langeron, Y., Grall, A., and Barros, A. (2015). A modeling framework for deteriorating control system and predictive maintenance of actuators. *Reliability Engineering & System Safety* 140: 22–36.

Le Son, K., Fouladirad, M., Barros, A. et al. (2013). Remaining useful life estimation based on stochastic deterioration models: a comparative study. *Reliability Engineering & System Safety* 112: 165–175.

Lee, J., Ghaffari, M., and Elmeligy, S. (2011). Self-maintenance and engineering immune systems: towards smarter machines and manufacturing systems. *Annual Reviews in Control* 35 (1): 111–122.

Nakagawa, T. (2007). *Shock and Damage Models in Reliability Theory*. London: Springer.

Nguyen, K.T.P., Do, P., Huynh, K.T. et al. (2019). Joint optimization of monitoring quality and replacement decisions in condition-based maintenance. *Reliability Engineering & System Safety* 189: 177–195.

Nicolai, R.P. and Dekker, R. (2008). Optimal maintenance of multi-component systems: a review. In: *Complex System Maintenance Handbook*, Chapter 11 (ed. K.A.H. Kobbachy and D.N.P. Murthy), 263–286. London: Springer.

Omshi, E.M., Grall, A., and Shemehsavar, S. (2019). A dynamic auto-adaptive predictive maintenance policy for degradation with unknown parameters. *European Journal of Operational Research* 282 (1): 81–92.

Rafiee, K., Feng, Q., and Coit, D.W. (2015). Condition-based maintenance for repairable deteriorating systems subject to a generalized mixed shock model. *IEEE Transactions on Reliability* 64 (4): 1164–1174.

Scarf, P.A. (1997). On the application of mathematical models in maintenance. *European Journal of Operational Research* 99 (3): 493–506.

Shafiee, M., Finkelstein, M., and Bérenguer, C. (2015). An opportunistic condition-based maintenance policy for offshore wind turbine blades subject to degradation and shocks. *Reliability Engineering & System Safety* 142: 463–471.

Si, X.S., Wang, W., Hu, C.H., and Zhou, D.H. (2011). Remaining useful life estimation – a review on the statistical data driven approaches. *European Journal of Operational Research* 213 (1): 1–14.

Van Noortwijk, J.M. (2009). A survey of the application of gamma processes in maintenance. *Reliability Engineering & System Safety* 94 (1): 2–21.

Vatn, J. and Svee, H. (2002). A risk-based approach to determine ultrasonic inspection frequencies in railway applications. *Proceedings of the 22nd ESReDA Seminar*, Madrid, Spain.

Vu, H.C., Do, P., and Barros, A. (2018). A study on the impacts of maintenance duration on dynamic grouping modeling and optimization of multicomponent systems. *IEEE Transactions on Reliability* 67 (3): 1377–1392.

Wang, W. (2008). Delay time modelling. In: *Complex System maintenance Handbook*, Chapter 14 (ed. K.A.H. Kobbacy and D.N.P. Murthy), 345–370. London: Springer.

Zhang, N., Fouladirad, M., and Barros, A. (2019). Reliability-based measures and prognostic analysis of a K-out-of-N system in a random environment. *European Journal of Operational Research* 272 (3): 1120–1131.

Zhu, W., Fouladirad, M., and Bérenguer, C. (2015). Condition-based maintenance policies for a combined wear and shock deterioration model with covariates. *Computers & Industrial Engineering* 85: 268–283.

Zio, E. and Compare, M. (2013). Evaluating maintenance policies by quantitative modeling and analysis. *Reliability Engineering & System Safety* 109: 53–65.

13

Reliability of Safety Systems

13.1 Introduction

This chapter deals with reliability aspects of safety systems that are designed to be activated upon hazardous system or process deviations (*system or process demands*) to protect people, the environment, and material assets. In Example 4.2, we discuss the safety systems of a gas/oil separator. The safety system has three *protection layers:*

(1) An inlet shutdown system comprising pressure sensors, a logic solver, and shutdown valves.
(2) A pressure relief system comprising two pressure relief valves.
(3) A rupture disc.

In Example 4.2, the process demand is a blockage of the gas outlet line. If safety systems were not available, the process demand would cause a rapid increase of the pressure in the separator and the separator might rupture. The system the protection layers are installed to protect is called the *equipment under control* (EUC). In this example, the EUC is the separator. An EUC may experience several hazardous process demands that require their own safety systems. In the process industry, the potential process demands are usually identified by a hazard and operability (HAZOP) study (e.g. see IEC 61882 2016).

Process demands may be classified according to their frequency of occurrence. Some process demands occur so frequently that the safety system is operated almost continuously. An example of such a safety system is the brakes of a car. "Process" demands for the brakes occur several times each time we drive the car, and brake failures and malfunctions may therefore be detected almost immediately. The brakes are said to be a safety system with a *high demand mode of operation*.

System Reliability Theory: Models, Statistical Methods, and Applications, Third Edition.
Marvin Rausand, Anne Barros, and Arnljot Høyland.

Companion Website: www.wiley.com/go/SystemReliabilityTheory3e

Other process demands occur very infrequently, and the safety system is therefore in a passive state for long periods of time. An example of such a system is the airbag system in a car. The airbag system remains passive until a “process” demand occurs and is said to be a safety system with a *low demand mode of operation*. Such a safety system may fail in passive state, and the failure may remain *hidden* until a process demand occurs or until the system is tested. To reveal hidden failures, safety systems with low demand mode of operation are normally proof tested at regular intervals.

A safety system composed of sensors, logic solvers, and final elements is called a *safety-instrumented system* (SIS). A brief introduction to SISs is given in Section 13.2. Several standards have been issued setting requirements to SISs. The most important of these standards is IEC 61508 (2010) “Functional safety of electrical/electronic/programmable electronic safety-related systems” that is briefly introduced in Section 13.7. For a thorough discussion of SIS reliability assessment, see Rausand (2014).

Section 13.3 introduces the main reliability models for the elements of safety systems and discuss various issues related to the analysis of such systems. The discussion is mainly limited to systems with a low demand of operation that are periodically tested. Problems related to common-cause failures (CCFs) and spurious activation of the systems are discussed. A Markov approach to analyzing safety systems is introduced in Section 13.9.

13.2 Safety-Instrumented Systems

A SIS is an independent protection layer that is installed to mitigate the risk associated with the operation of a specified hazardous system, EUC. The EUC may be various types of equipment, machinery, apparatus, or plant used for manufacturing, process, transportation, medical, or other activities. A SIS is composed of *sensors*, *logic solvers*, and *final elements*. The final elements may, for example, be shutdown valves or brakes. A sketch of a simple SIS is shown in Figure 13.1. SISs are used in many sectors of our society, for example, as emergency shutdown (ESD) systems in hazardous chemical plants, fire and

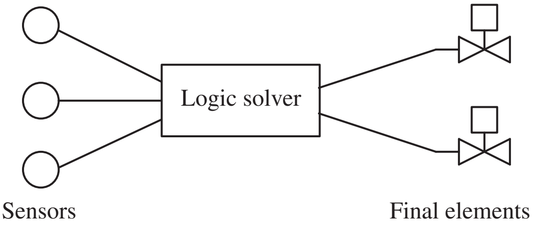


Figure 13.1 Sketch of a simple SIS.

gas detection (FGD) and alarm systems, pressure protection systems, dynamic positioning systems for ships and offshore platforms, automatic train stop (ATS) systems, fly-by-wire operation of aircraft flight control surfaces, anti-lock brakes, and airbag systems in cars, and systems for interlocking and controlling the exposure dose of medical radiotherapy machines. Recent developments include network-based safety-related systems, often facilitated by Internet technology.

A *safety-instrumented function* (SIF) is a function that is implemented by a SIS, and that is intended to achieve or maintain a safe state for the EUC with respect to a specific process demand. A SIS may perform one or more SIFs.

In addition to the elements shown in Figure 13.1 (sensors, logic solver, and final elements), an SIS usually comprises: electric power supply, user interface, pneumatic and/or hydraulic system, electrical connections, and various process connections.

IEC 61508 refers to a SIS as an “electrical/electronic/programmable electronic (E/E/PE) safety-related system.”

13.2.1 Main SIS Functions

A SIS has two main system functions:

(1) When a predefined process demand (deviation) occurs in the EUC, the deviation shall be detected by the SIS *sensors*, and the required *final elements* shall be activated and fulfill their intended functions.
(2) The SIS shall not be activated spuriously, that is, without the presence of a predefined process demand (deviation) in the EUC.

A failure to perform the first system function is called a *fail to function* (FTF), and a failure of the second function is called a *spurious trip* (ST).

Example 13.1 (Safety systems on offshore oil and gas platforms)
The safety systems on an offshore oil and gas platform are usually grouped into three categories:

(1) Process control (PC) system
(2) Process shutdown (PSD) system
(3) Fire and gas detection (FGD) and emergency shutdown (ESD) system

The objective of the PC system is to keep an EUC process within preset limits. Various PC valves and regulators are used to control the process, based on signals from temperature, pressure, level, and other types of transmitters. When the process deviates from normal values, the PSD system is activated and closes down the EUC. The required actions for each type of deviation/demand is programmed into the logic solver. The actions may involve activation of alarms, closure of shutdown valves, and opening of relief valves. The PC and PSD systems are *local* systems that are related to a specific EUC. For process demands that have a potential for a major accident, the ESD system is activated. Relevant process demands include fires, gas leaks, and loss of main power. The required ESD actions are usually grouped into several levels, depending on the type of deviation/demand that is detected and *where* it is detected. The top ESD level usually involves shutdown of the whole platform and evacuation of the personnel. □

13.2.2 Testing of SIS Functions

Many SISs are passive systems that are only activated when a specified process demand occurs in the EUC. A fire detection and extinguishing system should, for example only be activated when a fire occurs. Such a system may fail in the passive position and the failure may remain undetected (hidden) until the system is activated or tested.

Diagnostic Self-Testing

In modern SISs the logic solver is often programmable and may carry out *diagnostic self-testing* during online operation. The logic solver may send frequent signals to the detectors and to the final elements and compare the responses with predefined values. The diagnostic testing can reveal failures of input and output devices, and to an increasing degree, also failures of detectors and final elements. In many cases, the logic solver consists of two or more redundant computers that can carry out diagnostic self-testing of each other. The fraction of failures that can be revealed by diagnostic self-testing is called the *diagnostic coverage*. The self-testing may be carried out so often that failures are detected almost immediately.

Proof Testing

The diagnostic self-testing cannot reveal all failure modes and failure causes, and the various parts of the SIS are therefore proof tested at regular intervals. The objective of a proof test is to reveal hidden failures/faults and to verify that the system is (still) able to perform the required functions if a process demand should occur. It is sometimes not feasible to carry out a fully realistic proof test because it may not be technically feasible, or very time-consuming. Another reason may be that the test itself leads to unacceptable hazards. It is, for example not realistic to fill a room with toxic gases to test a gas detector. The gas detector is rather tested with a nontoxic test gas that is directly input to the gas detector through a test pipe.

Consider a safety valve that is installed in a pipeline. During normal operation, the valve is kept in open position. If a specified process demand occurs, the valve should close and stop the flow in the pipeline. A realistic test of the safety valve would imply to close the valve and apply a pressure to the upstream side of the valve that is equal to the maximum expected shut-in pressure in a demand situation. This may not be possible, and we may have to suffice with only checking that the valve is able to close on demand, and perhaps to check the valve for leakage with normal shut-in pressure. In some cases, it may be possible to pressure test the valve from the downstream side. In this case, we may be able to test the valve to maximum shut-in pressure, but the wrong side of the seals is tested. In some situations, it may be hazardous to shut down a flow, and the closure of the valve should therefore be avoided. Some valve functions may be tested by partly closing the valve (the gate of a gate valve may be moved some few millimeters, and a ball valve may be rotated some degrees). This type of testing is called *partial stroke testing* and is discussed further by Lundteigen and Rausand (2008).

Some final elements employ an actuating principle that is not possible to proof test without destroying the item. This is, for example, the case for the pyrotechnic seat belt tensioners in cars. The reader can refer to Brissaud et al. (2012), Srivastav et al. (2018), and Wu et al. (2018) for further work and more examples on this topic.

13.2.3 Failure Classification

A general introduction to failures and failure classification is given in Chapter 3. For a SIS and the SIS subsystems, we may use the following failure mode classification (see IEC 61508):

(1) *Dangerous (D).* The SIS does not fulfill its required safety-related functions upon demand. These failures may further be split into
  (a) *Dangerous undetected (*DU*).* Dangerous failures are preventing activation on demand and are revealed only when tested or when a demand occurs. DU failures are sometimes called *dormant* failures.
  (b) *Dangerous detected (*DD*).* Dangerous failures that are detected immediately when they occur, for example by an automatic, built-in self-test. The average period of unavailability due to a DD failure is equal to the mean downtime, MDT, that is, the mean time elapsing from the failure is detected by the built-in self-test until the function is restored.
(2) *Safe failures (S).* The SIS has a nondangerous failure. These failures may further be split into
  (a) *Safe undetected (*SU*).* Nondangerous failures that are not detected by automatic self-testing.
  (b) *Safe detected (*SD*).* Nondangerous failures that are detected by automatic self-testing. In some configurations, early detection of failures may prevent an actual ST of the system.

The failure mode classification is shown in Figure 13.2.

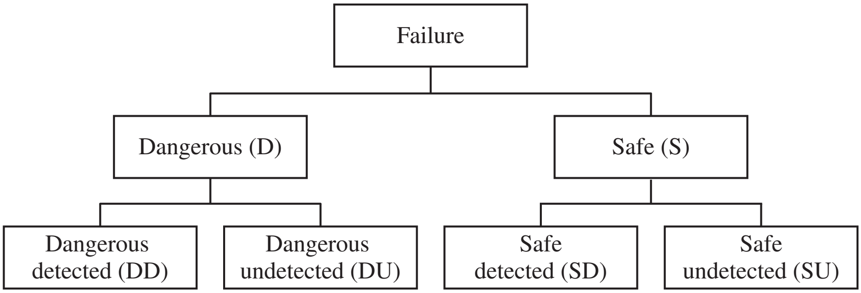


Figure 13.2 Failure mode classification.

Example 13.2 (Safety shutdown valve)
A safety shutdown valve is installed in a gas pipeline feeding a production system. If an emergency occurs in the production system, the valve should close and stop the gas flow. The valve is a hydraulically operated gate valve. The actual open/close function is performed by sliding a rectangular gate, having a bore equal to the bore of the conduct. The gate is moved by a hydraulic piston connected to the gate by a stem. The gate valve has a *fail-safe* actuator. The valve is opened and kept open by hydraulic control pressure on the piston. The fail-safe function is achieved by a steel spring that is compressed by hydraulic pressure. The valve is automatically closed by spring force when the hydraulic pressure is bled off.

The valve is connected to an ESD system. When an emergency situation is detected in the production system, an electric signal is sent to the valve control system and the pressure is bled off. In this example, we only consider the valve but will come back to the rest of the ESD system in Sections 13.4 and 13.5.

The main failure modes of the valve are

- *Fail to close (FTC) on command*. This failure mode may be caused by a broken spring, blocked return line for the hydraulic fluid, too high friction between the stem and the stem seal, too high friction between the gate and the seats, or by sand, debris, or hydrates in the valve cavity.
- *Leakage (through the valve) in closed position* (LCP). This failure mode is mainly caused by corrosion and/or erosion on the gate or the seat. It may also be caused by misalignment between the gate and the seat.
- *Spurious trip* (ST). This failure mode occurs when the valve closes without a signal from the ESD system. It is caused by a failure in the hydraulic system or a leakage in the supply line from the control system to the valve.
- *Fail to open* (*FTO*) on command. When the valve is closed, it may fail to reopen. Possible causes may be leakage in the control line, too high friction between the stem seals and the stem, too high friction between the gate and the seats, and sand, debris, or hydrates in the valve cavity.

The valve has been installed to close the flow (and keep tight) following a demand. The failure modes FTC and LCP prevent this function and are *dangerous* failure modes with respect to safety. ST and FTO failures are normally not dangerous with respect to safety, but may cause production shutdown and lost income.

Because the valve is normally in open position, we are not able to detect the dangerous failure modes, FTC and LCP, unless we try to close the valve. These dangerous failure modes are *hidden* during normal operation and are therefore called *dangerous undetected* (DU) failure modes. To reveal, and repair, DU failures, the valve is tested periodically, with test interval τ. This means that the valve is tested at times $0, \tau, 2\tau, \ldots$. A typical test interval may be 3–12 months. During a standard test, the valve is closed and tested for leakage. The cause of a DU failure may occur at a random point of time within a test interval and is not manifested (revealed) until the valve is tested, or attempted closed due to operational reasons. The safety unavailability (SU) of the valve is obviously lower with a short test interval than with a long test interval. The gas flow has to be closed down during the test, and the test will usually lead to a production loss. In some situations, the shutdown and startup procedure may have safety implications. The length of the test interval τ must therefore be a compromise between safety and economic considerations.

In some situations, it may be impractical and even dangerous to close the valve, and we have to suffice with *partial stroke* testing. In this case, we move the gate slightly and monitor the movement of the valve stem. The test reveals some of the DU failure causes, but not all. A hidden LCP failure will, for example not be revealed.

The ST failure stops the flow and is usually detected immediately. An ST failure is therefore called an *evident* failure. In some systems, an ST failure may also have significant safety implications.

The FTO failure may occur after a test and is an evident failure. The FTO failure will cause a repair intervention, but this has no extra safety implications because the gas flow is shut down when the failure occurs. The FTO failure is therefore called a *noncritical* or *safe* failure. □

13.3 Probability of Failure on Demand

Consider a safety item (component or system) that is tested periodically, in the same way as the safety valve in Example 13.2. We assume that no diagnostic self-testing is carried out, and that *all* hidden failures are revealed by the proof testing. Some of the main concepts that are used in this section are introduced in Example 13.2. The reader should therefore study the example carefully before reading this section.

The safety item is put into operation at time $t = 0$. The item may be a safety valve (e.g. shutdown valve, or relief valve), a sensor (e.g. fire/gas detector, pressure sensor, or level sensor), or a logic solver. The item is tested and, if necessary, repaired or replaced after regular time intervals of length τ. The time required to test and repair the item is considered to be negligible. After a test (repair), the item is considered to be as-good-as-new. We say that the item is functioning as a safety *barrier* if a DU failure mode is not present.

The state variable $X(t)$ of an item with respect to DU failures is

$$X(t) = \begin{cases} 1 & \text{if the item is able to function as a safety barrier} \\ & \text{(i.e. no DU failure is present)} \\ 0 & \text{if the item is not able to function as a safety barrier} \\ & \text{(i.e. a DU failure is present)} \end{cases}.$$

The state variable $X(t)$ is shown in Figure 13.3.

13.3.1 Probability of Failure on Demand

Let T be the time to DU failure of the item, with distribution function $F(t)$. The safety *unavailability* $A^*(t)$ of the item in the *first* test interval $(0, \tau]$ is

$$\begin{aligned} A^*(t) &= \Pr(\text{a DU failure has occurred at, or before, time } t) \\ &= \Pr(T \leq t) = F(t). \end{aligned} \tag{13.1}$$

Because the item is assumed to be as-good-as-new after each test, the test intervals $(0, \tau], (\tau, 2\tau], \ldots$ are all equal from a stochastic point of view. Hence, the safety unavailability $A^*(t)$ of the item is as shown in Figure 13.4. Observe that $A^*(t)$ is discontinuous for $t = n\tau$, for $n = 1, 2, \ldots$. If a demand for the safety item occurs at time t, the safety unavailability $A^*(t)$ is the probability that the item will fail

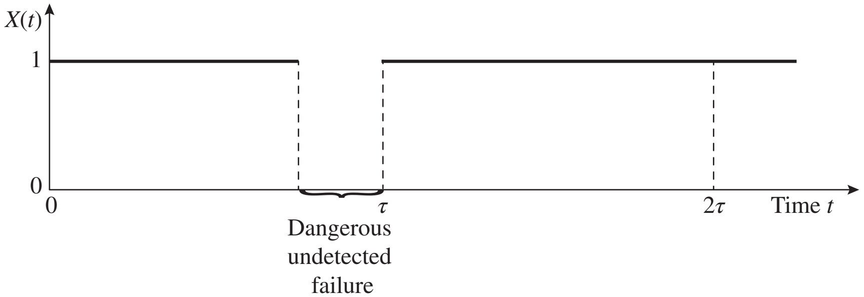


Figure 13.3 The state $X(t)$ of a periodically tested item with respect to DU failures.

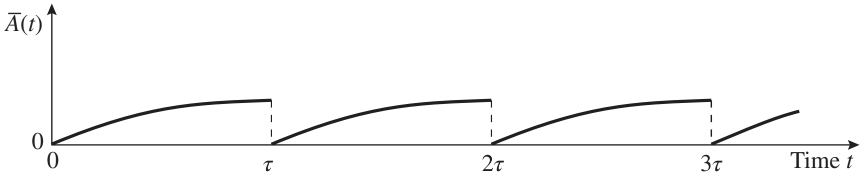


Figure 13.4 The safety unavailability $A^*(t)$ of a periodically tested item.

to respond adequately to the demand. The safety unavailability $A^*(t)$ is therefore often called the *probability of failure on demand* (PFD) at time t.

In most applications, we are not interested in the PFD as a function of time. It is sufficient to know the long-run average value of PFD. The average value is denoted PFD, without reference to the time t. Because of the periodicity of $A^*(t)$, the long-run average PFD is equal to the average value of $A^*(t)$ in the first test interval $(0, \tau]$,

$$\text{PFD} = \frac{1}{\tau}\int_0^\tau A^*(t)\,dt = \frac{1}{\tau}\int_0^\tau F(t)\,dt. \tag{13.2}$$

Let $R(t)$ be the survivor function of the item with respect to DU failure. Because $R(t) = 1 - F(t)$, (13.2) may alternatively be written as

$$\text{PFD} = 1 - \frac{1}{\tau}\int_0^\tau R(t)\,dt. \tag{13.3}$$

Consider a test interval, and let T_1 be the part of this test interval where the item is able to function as a safety barrier. Let D_1 be the part of the interval where the item is in a failed state (i.e. a DU failure is present but has not been detected), such that $T_1 + D_1 = \tau$.

The PFD in (13.2) is the average safety unavailability in a test interval. Because the average safety unavailability is the mean proportion of time the item is not functioning as a safety barrier, the PFD may be written as

$$\text{PFD} = \frac{\text{E}(D_1)}{\tau}. \tag{13.4}$$

The MDT in a test interval is therefore

$$E(D_1) = \int_0^\tau F(t)\,dt, \tag{13.5}$$

and the mean uptime in a test interval is

$$E(T_1) = \tau - \int_0^\tau F(t)\,dt = \int_0^\tau R(t)\,dt. \tag{13.6}$$

The PFD may from (13.4) be interpreted as the mean proportion of time the item is not functioning as a safety barrier upon demand. The PFD is therefore also referred to as the *mean fractional deadtime* (MFDT) of the item.

Example 13.3 (Single item)

A sensor is tested at regular intervals of length τ and has constant failure rate λ_{DU} with respect to DU failures. The survivor function of the sensor is $R(t) = e^{-\lambda_{\text{DU}} t}$ and the PFD is from (13.3)

$$\begin{aligned} \text{PFD} &= 1 - \frac{1}{\tau} \int_0^\tau R(t)\, dt = 1 - \frac{1}{\tau} \int_0^\tau e^{-\lambda_{\text{DU}} t}\, dt \\ &= 1 - \frac{1}{\lambda_{\text{DU}}\tau}\, (1 - e^{-\lambda_{\text{DU}}\tau}). \end{aligned} \tag{13.7}$$

By replacing $e^{-\lambda_{\text{DU}}\tau}$ in (13.7) with its Maclaurin series, we get

$$\begin{aligned} \text{PFD} &= 1 - \frac{1}{\lambda_{\text{DU}}\tau} \left(\lambda_{\text{DU}}\tau - \frac{(\lambda_{\text{DU}}\tau)^2}{2} + \frac{(\lambda_{\text{DU}}\tau)^3}{3!} - \frac{(\lambda_{\text{DU}}\tau)^4}{4!} + \cdots \right) \\ &= 1 - \left(1 - \frac{\lambda_{\text{DU}}\tau}{2} + \frac{(\lambda_{\text{DU}}\tau)^2}{3!} - \frac{(\lambda_{\text{DU}}\tau)^3}{4!} + \cdots \right). \end{aligned}$$

When $\lambda_{\text{DU}}\tau$ is small, then

$$\text{PFD} \approx \frac{\lambda_{\text{DU}}\tau}{2}. \tag{13.8}$$

This approximation is often used in practical calculation. The approximation is always conservative, meaning that the approximated value in (13.8) is slightly greater than the correct value in (13.7).

According to OREDA (2015) the failure rate of a specific type of fire detectors is $\lambda_{\text{DU}} = 0.21 \times 10^{-6}$ DU failures/h. If we use a test interval $\tau = 3$ months ≈ 2190 hours, the PFD is

$$\text{PFD} \approx \frac{\lambda_{\text{DU}}\tau}{2} = \frac{0.21 \times 10^{-6} \times 2190}{2} \approx 0.000\,23 = 2.30 \times 10^{-4}.$$

If a demand for the fire detector occurs, the (average) probability that the detector will not be able to detect the fire is: PFD $\approx$ 0.000 23. This means that approximately one out of 4350 fires will not be detected by the fire detector.

The mean proportion of time the detector is not able to detect a fire is PFD $\approx$ 0.000 23. This means that the fire detector is not able to detect a fire in 0.023% of the time, or approximately 2h/yr, when we assume that the detector is in continuous operation, and that a year is 8760 hours. We also say that we are *unprotected* by the fire detector in 0.023% of the time. □

Example 13.4 (Parallel structure)

Assume that we have two independent fire detectors of the same type with failure rate λ_{DU} with respect to DU failures, that are tested at the same time with test interval τ. The fire detectors are operated as a 1oo2:G structure, where it is sufficient that one detector is functioning for the structure to function. The survivor function for the structure is

$$R(t) = 2e^{-\lambda_{\text{DU}} t} - e^{-2\lambda_{\text{DU}} t},$$

and the PFD is from (13.3)

$$\begin{aligned} \text{PFD} &= 1 - \frac{1}{\tau} \int_0^{\tau} 2e^{-\lambda_{\text{DU}} t} - e^{-2\lambda_{\text{DU}} t}\, dt \\ &= 1 - \frac{2}{\lambda_{\text{DU}}\tau}(1 - e^{-\lambda_{\text{DU}}\tau}) + \frac{1}{2\lambda_{\text{DU}}\tau}(1 - e^{-2\lambda_{\text{DU}}\tau}). \end{aligned} \tag{13.9}$$

If we replace $e^{-\lambda_{\text{DU}}\tau}$ by its Maclaurin series, we may use the following approximation:

$$\text{PFD} \approx \frac{1}{3}\, (\lambda_{\text{DU}}\tau)^2, \tag{13.10}$$

when $\lambda_{\text{DU}}\tau$ is small.

Let us now introduce the same data as we used for one single fire detector in Example 13.3, $\lambda_{\text{DU}} = 0.21 \times 10^{-6}\ \text{h}^{-1}$ and $\tau = 3$ months. The average unavailability of the parallel structure is then

$$A^*_{\text{avg}} \approx \frac{1}{3}\, (\lambda_{\text{DU}}\tau)^2 = \frac{1}{3}(0.21 \times 10^{-6} \times 2190)^2 \approx 7.1 \times 10^{-8}.$$

If a demand for the fire detector system occurs, the (average) probability that the system will not be able to detect the fire is: PFD $\approx 7.1 \times 10^{-8}$, that is, a very high reliability. □

Remark 13.1 (The average of a product is not the product of the averages)
Because the parallel structure fails only when both of its components fail, the probability, $Q_S(t)$, that the structure is in a failed state at time t, is equal to $q_1(t)\ q_2(t)$, where $q_i(t)$ is the probability that component i is in a failed state at time t, for $i = 1, 2$. Because the (average) probability that component i is in a failed state is $\text{PFD}_i \approx \lambda_{\text{DU}}\tau/2$, we should expect that the average unavailability (PFD) of the system would be approximately $(\lambda_{\text{DU}}\tau/2)^2 = (\lambda_{\text{DU}}\tau)^2/4$ instead of $(\lambda_{\text{DU}}\tau)^2/3$ as we found in (13.10). The result in (13.10) is the correct result. The reason for this difference is the fact that the average of a product is not the same as the product of averages. Several computer programs for fault tree analysis make this error. A bad effect is that the wrong approach produces a nonconservative result. □

Example 13.5 (2oo3 structure)
Assume that we have three independent fire detectors of the same type with failure rate λ_{DU} with respect to DU failures, that are tested at the same time with test interval τ. The fire detectors are operated as a 2oo3:G structure, where two detectors have to function for the structure to function. The survivor function for the structure is

$$R(t) = 3e^{-2\lambda_{\text{DU}} t} - 2e^{-3\lambda_{\text{DU}} t},$$

and the PFD is from (13.3)

$$\begin{aligned} \text{PFD} &= 1 - \frac{1}{\tau}\int_0^\tau (3e^{-2\lambda_{\text{DU}}t} - 2e^{-3\lambda_{\text{DU}}t})\,dt \\ &= 1 - \frac{3}{2\lambda_{\text{DU}}\tau}(1 - e^{-2\lambda_{\text{DU}}\tau}) + \frac{2}{3\lambda_{\text{DU}}\tau}(1 - e^{-3\lambda_{\text{DU}}\tau}). \end{aligned} \tag{13.11}$$

If we replace $e^{-\lambda_{\text{DU}}\tau}$ by its Maclaurin series, we may use the following approximation:

$$\text{PFD} \approx (\lambda_{\text{DU}}\tau)^2, \tag{13.12}$$

when $\lambda_{\text{DU}}\tau$ is small.

Let us now introduce the same data as we used for one single fire detector in Example 13.3, $\lambda_{\text{DU}} = 0.21 \times 10^{-6}\ \text{h}^{-1}$ and $\tau = 3$ months. The average unavailability of the parallel structure is then

$$\text{PFD} \approx (\lambda_{\text{DU}}\tau)^2 = (0.21 \times 10^{-6} \times 2190)^2 \approx 2.1 \times 10^{-7}.$$

If a demand for the fire detector system occurs, the (average) probability that the system will not be able to detect the fire is: PFD $\approx 2.1 \times 10^{-7}$. □

The PFD of a 2oo3:G structure is seen to be approximately three times as high as for a parallel structure. Chapter 6 shows that a 2oo3:G structure may be represented as a series structure of three 1oo2:G, parallel structures. Each of these parallel structures has an average unavailability $(\lambda_{\text{DU}}\tau)^2/3$. When $\lambda_{\text{DU}}\tau$ is small, the probability of two parallel structures being in a failed state at the same time is negligible, and the average unavailability of the 2oo3:G structure is then approximately the sum of the average availabilities of the three parallel structures, which is the case.

Example 13.6 (Series structure)

Assume that we have two independent items with failure rate $\lambda_{\text{DU},1}$ and $\lambda_{\text{DU},2}$, respectively, with respect to DU failures. The items are tested at the same time with test interval τ. The items are operated as a 2oo2:G structure, where both items have to function for the structure to function. The survivor function for the structure is

$$R(t) = e^{-(\lambda_{\text{DU},1}+\lambda_{\text{DU},2})t},$$

and the PFD is from (13.3)

$$\begin{aligned} \text{PFD} &= 1 - \frac{1}{\tau}\int_0^\tau e^{-(\lambda_{\text{DU},1}+\lambda_{\text{DU},2})t}\,dt \\ &\approx \frac{(\lambda_{\text{DU},1} + \lambda_{\text{DU},2})\tau}{2} = \frac{\lambda_{\text{DU},1}\tau}{2} + \frac{\lambda_{\text{DU},2}\tau}{2}, \end{aligned} \tag{13.13}$$

when $\lambda_{\mathrm{DU},i}\tau$ is small, for $i = 1, 2$. When we have a series structure, the PFD of the structure is hence approximately the sum of the PFDs of the individual items. □

13.3.2 Approximation Formulas

Assume that we have a system of n independent components with constant failure rates $\lambda_{\mathrm{DU},i}$, for $i = 1, 2, \dots, n$. The distribution function $F_{T_i}(t)$ of item i is approximated by

$$F_{T_i}(t) = 1 - e^{-\lambda_{\mathrm{DU},i}t} \approx \lambda_{\mathrm{DU},i}t.$$

By using fault tree terminology, the unavailability of component i in the first test interval is

$$\begin{aligned} q_i(t) &= \Pr(\text{Component } i \text{ is in a failed state at time } t) \\ &= F_{T_i}(t) \approx \lambda_{\mathrm{DU}_i}t. \end{aligned}$$

Let $K_1, K_2, \dots, K_k$ be the k minimal cut sets of the system. The probability that the minimal cut parallel structure corresponding to the minimal cut set K_j is failed at time t is

$$\check{Q}_j(t) = \prod_{i \in K_j} q_i(t) \approx \prod_{i \in K_j} \lambda_{\mathrm{DU},i}t \qquad \text{for } j = 1, 2, \dots, k.$$

The probability that the system is failed (has a hidden failure) at time t is

$$\begin{aligned} Q_0(t) = F_S(t) &\approx \sum_{j=1}^{k} \check{Q}_j(t) \approx \sum_{j=1}^{k} \prod_{i \in K_j} \lambda_{\mathrm{DU},i}t \\ &= \sum_{j=1}^{k} \left(\prod_{i \in K_j} \lambda_{\mathrm{DU},i} \right) t^{|K_j|}, \end{aligned} \tag{13.14}$$

where $|K_j|$ denotes the *order* of the minimal cut set K_j, $j = 1, 2, \dots, k$.

The PFD of the system that is tested periodically with test interval τ is, by combining (13.2) and (13.14), approximately

$$\text{PFD} = \frac{1}{\tau} \int_0^\tau F_s(t)\, dt \approx \sum_{j=1}^{k} \prod_{i \in K_j} \lambda_{\mathrm{DU},i} \frac{1}{\tau} \int_0^\tau t^{|K_j|}\, dt. \tag{13.15}$$

Hence,

$$\text{PFD} \approx \sum_{j=1}^{k} \frac{1}{|K_j| + 1} \prod_{i \in K_j} \lambda_{\mathrm{DU},i}\tau. \tag{13.16}$$

Table 13.1 PFD of some *k*oo*n* structures of identical and independent components with failure rate λ_{DU} and test interval τ.

k\n	1	2	3	4
1	$\frac{\lambda_{\text{DU}}\tau}{2}$	$\frac{(\lambda_{\text{DU}}\tau)^2}{3}$	$\frac{(\lambda_{\text{DU}}\tau)^3}{4}$	$\frac{(\lambda_{\text{DU}}\tau)^4}{5}$
2	—	$\lambda_{\text{DU}}\tau$	$(\lambda_{\text{DU}}\tau)^2$	$(\lambda_{\text{DU}}\tau)^3$
3	—	—	$\frac{3\lambda_{\text{DU}}\tau}{2}$	$2(\lambda_{\text{DU}}\tau)^2$
4	—	—	—	$2\lambda_{\text{DU}}\tau$

Assume now that we have a *k*oo*n* structure of identical and independent components with failure rate λ_{DU}. A *k*oo*n*:G structure has $\binom{n}{n-k+1}$ minimal cut sets of order $(n-k+1)$. The PFD of the *k*oo*n*:G structure is thus

$$\begin{aligned} \text{PFD} &\approx \frac{1}{\tau}\int_0^\tau \binom{n}{n-k+1} (\lambda_{\text{DU}}t)^{n-k+1}\,dt \\ &= \binom{n}{n-k+1} \frac{(\lambda_{\text{DU}}\tau)^{n-k+1}}{n-k+2}. \end{aligned} \tag{13.17}$$

The PFD of some simple *k*oo*n*:G structures are listed in Table 13.1.

13.3.3 Mean Downtime in a Test Interval

The MDT in a test interval was found in (13.4) to be

$$E(D_1) = \int_0^\tau F(t)\,dt.$$

Suppose that we test an item at time τ and find that the item is in a failed state [i.e. $X(\tau) = 0$]. What is the (conditional) MDT in the interval $(0, \tau]$ when the item is found in a failed state at time τ?

By using double expectation, $E(D_1)$ may be written as

$$\begin{aligned} E(D_1) &= E(E[D_1 \mid X(\tau)]) \\ &= E(D_1 \mid X(\tau) = 0)\ \Pr(X(\tau) = 0) \\ &\quad + E(D_1 \mid X(\tau) = 1)\ \Pr(X(\tau) = 1). \end{aligned}$$

If the component is functioning at time τ, the downtime D_1 is equal to 0. Therefore, $E(D_1 \mid X(\tau) = 1) = 0$. Furthermore,

$$\Pr(X(\tau) = 0) = \Pr(T \le \tau) = F(\tau).$$

Hence,

$$E(D_1) = E(D_1 \mid X(\tau) = 0)\ F(\tau).$$

By using (13.6) and (13.3)

$$E(D_1 \mid X(\tau) = 0) = \frac{E(D_1)}{F(\tau)} = \frac{1}{F(\tau)} \int_0^\tau F(t)\,dt$$
$$= \frac{\tau}{F(\tau)}\,\frac{1}{\tau} \int_0^\tau F(t)\,dt = \frac{\tau}{F(\tau)}\,\text{PFD}. \tag{13.18}$$

Example 13.7 (Example 13.3 (Cont.))
With a single item, the conditional MDT in (13.18) is approximately

$$E(D_1 \mid X(\tau) = 0) = \frac{\tau}{F(\tau)}\,\text{PFD} \approx \frac{\tau}{1 - e^{-\lambda_{\text{DU}}\tau}}\,\frac{\lambda_{\text{DU}}\,\tau}{2} \approx \frac{\tau}{2},$$

which is an intuitive result. □

Example 13.8 (Example 13.4 (Cont.))
With a parallel structure of two independent, and identical items, the conditional MDT in (13.18) is

$$E(D_1 \mid X(\tau) = 0) = \frac{\tau}{F(\tau)}\,\text{PFD} \approx \frac{\tau}{1 - 2e^{-\lambda_{\text{DU}}\tau} + e^{-2\lambda_{\text{DU}}\tau}}\,\frac{(\lambda_{\text{DU}}\,\tau)^2}{3} \approx \frac{\tau}{3}.$$

The last approximation follows because the distribution function of the parallel structure $1 - 2e^{-\lambda_{\text{DU}}\tau} + e^{-2\lambda_{\text{DU}}\tau}$ can be approximated by $(\lambda_{\text{DU}}\,\tau)^2$ by using Maclaurin series. □

13.3.4 Mean Number of Test Intervals Until First Failure

Let us next determine the mean number of test intervals *until* the first failure occurs. Let C_i be the event that the component does *not* fail in test interval i for $i = 1, 2, \ldots$. Then

$$\Pr(C_i) = \Pr(T > \tau) = R(\tau).$$

Because the events $C_1, C_2, \ldots$ are independent with the same probability $p = R(\tau)$, the number of test intervals, Z, *until* the component fails for the first time, has a geometric distribution with point probability

$$\Pr(Z = z) = \Pr(C_1 \cap C_2 \cap \cdots \cap C_z \cap C_{z+1}^c) = p^z(1-p) \quad \text{for} \quad z = 0, 1, \ldots.$$

The mean number of test intervals *until* the component fails is then

$$E(Z) = \sum_{z=0}^{\infty} \text{z}\Pr(\text{Z} = \text{z}) = \frac{p}{1-p} = \frac{R(\tau)}{F(\tau)}. \tag{13.19}$$

Let T' be the time from the component is put into operation until its first failure. Then

$$E(T') = \tau E(Z) + (\tau - E(D_1 \mid X(\tau) = 0))$$

$$= \tau \frac{R(\tau)}{F(\tau)} + \tau - \frac{1}{F(\tau)} \left(\tau - \int_0^\tau R(t)\, dt \right)$$
$$= \frac{1}{F(\tau)} \int_0^\tau R(t)\, dt. \tag{13.20}$$

Example 13.9 If in particular the component has constant failure rate λ_{DU}, then

$$E(T') = \frac{1}{F(\tau)} \int_0^\tau R(t)\, dt = \frac{1}{1 - e^{-\lambda_{\text{DU}}\tau}} \int_0^\tau e^{-\lambda_{\text{DU}} t}\, dt = \frac{1}{\lambda_{\text{DU}}},$$

a result that follows directly from the properties of the exponential distribution.□

13.3.5 Staggered Testing

When we have two items in parallel, we may reduce the system PFD by testing the items with the same test interval, but at different times. Assume that we have two independent items with constant failure rates $\lambda_{\text{DU},1}$ and $\lambda_{\text{DU},2}$, respectively, with respect to DU failures. Item 1 is tested at times $0, \tau, 2\tau, \ldots$, whereas item 2 is tested at times $t_0, \tau + t_0, 2\tau + t_0, \ldots$. This testing is called *staggered testing* with interval t_0. Assume that the time necessary for testing and repair is so short that it can be neglected. Further assume that the process has been running some time and that time 0 is the time for a test of item 1.

The PFD of the two items as a function of time is shown in Figure 13.5. In the first test interval $(0, \tau]$ the items have the following unavailabilities:

$$\begin{aligned} q_1(t) &= 1 - e^{-\lambda_{\text{DU},1} t} && \text{for} \quad 0 < t \le \tau \\ q_2(t) &= 1 - e^{-\lambda_{\text{DU},2}(t+\tau-t_0)} && \text{for} \quad 0 \le t \le t_0\,. \\ &= 1 - e^{-\lambda_{\text{DU},2}(t-t_0)} && \text{for} \quad t_0 < t \le \tau \end{aligned}$$

The unavailability of item 1, $q_1(t)$, is shown by a short-dashed line in Figure 13.5, whereas the unavailability of item 2, $q_2(t)$, is shown by a long-dashed line.

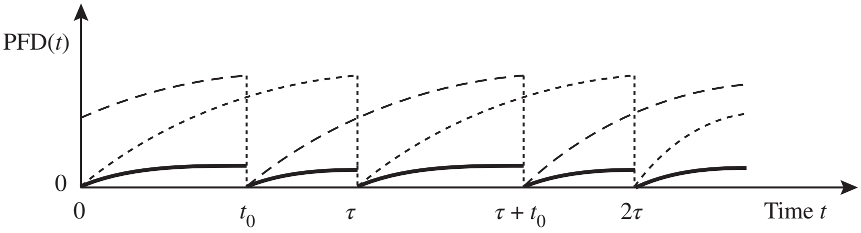


Figure 13.5 PFD(t) of a parallel structure of two items with staggered testing. Item 1 (short dash) is tested at times $0, \tau, 2\tau, \ldots$, whereas item 2 (long dash) is tested at times $t_0, \tau + t_0, 2\tau + t_0, \ldots$. The system PFD($t$) is the fully drawn curve.

The system unavailability $q_s(t) = q_1(t)\, q_2(t)$ is shown by a fully drawn line in Figure 13.5.

$$q_s(t) = \begin{cases} (1 - e^{-\lambda_{\mathrm{DU},1} t})(1 - e^{-\lambda_{\mathrm{DU},2}(t+\tau-t_0)}) & \text{for} \quad 0 < t \le t_0 \\ (1 - e^{-\lambda_{\mathrm{DU},1} t})(1 - e^{-\lambda_{\mathrm{DU},2}(t-t_0)}) & \text{for} \quad t_0 < t \le \tau \end{cases}.$$

The average unavailability in $(0, \tau]$ is equal to the PFD and is a function of t_0

$$\mathrm{PFD}(t_0) = \frac{1}{\tau} \int_0^{\tau} q_s(t)\, dt.$$

The integration is straightforward but requires several steps and is left to the reader as Problem 13.7. For details and extensions, see Liu (2014).

13.3.6 Nonnegligible Repair Time

In some situations, the repair time after a failure is so long that it cannot be neglected. This is, for example illustrated by the following example:

Example 13.10 (Downhole safety valve)
A downhole safety valve (DHSV) is located in the oil/gas production tubing in subsea production wells. The DHSV is an integral part of the tubing approximately 100 m below the sea bottom. The valve has a spring-loaded hydraulic fail-safe actuator and is held open by hydraulic pressure. The operation of the DHSV is comparable to the gate valve described in Example 13.2, and the DHSV have the same failure modes as the gate valve. The DHSV is tested periodically, with a test interval of 6–12 months. To repair a failed valve is a long, hazardous, and extremely costly operation. A semisubmersible intervention rig has to be moved from its permanent location out to the offshore field. The tubing string has to be pulled and the well pressure has to be controlled during the intervention. The operation may last several weeks, depending on the system and the weather conditions. In addition, we may have to wait months before an intervention rig becomes available. In this case, the repair time is far from negligible. □

Example 13.10 illustrates that the item may sometimes be unavailable as a safety barrier during the repair action, and while waiting for repair. This unavailability may be different from the unavailability in the test interval because we know that the item is in a failed state and may take precautions to reduce the risk. The time from a failure is detected until the function is restored and is sometimes called the *restoration time*. The risk associated to the restoration time may depend on

- The *failure mode*. The various failure modes of the item may require different repair actions and the risk during waiting for repair may also be different.

- The various *phases* of the restoration time may have different risk levels. The risk during waiting for repair may, for example be different from the risk during actual repair.

It may therefore be necessary to find the unavailability for each failure mode and for the various phases of the restoration time.

13.4 Safety Unavailability

The *safety unavailability* $A^*(t)$, of a safety system is the probability that the system is *not* able to perform its required function upon a demand. The safety unavailability may be split in four categories, as shown in Figure 13.6. The categories of the safety unavailability are discussed in Hauge et al. (2013), where also more detailed categories are defined.

NSU. Noncritical safety unavailability of the item, mainly caused by functional testing. In this case, it is known that the item is unavailable, and other preventive actions may be taken.

PFD: The (unknown) safety unavailability due to DU failures during the test interval when it is not known that the function is unavailable.

PFD$_K$: Safety unavailability of the item due to restoration actions after a failure has been revealed. In this case, we know that the item is unavailable. The various phases of the restoration actions may give rise to different levels of risk.

PSF: The probability that a systematic failure prevents the item from performing its intended function. Systematic failures are not revealed by periodic testing.

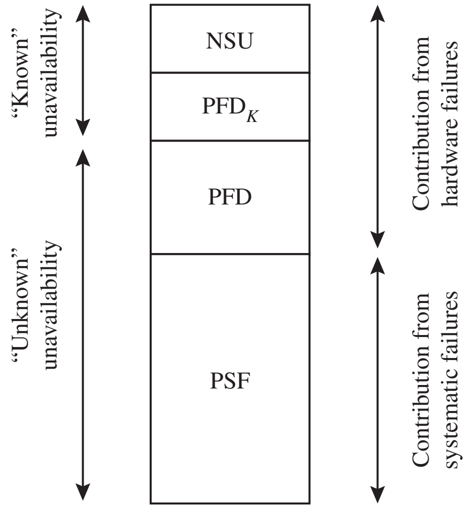

"Known" unavailability
"Unknown" unavailability
NSU
PFD$_K$
PFD
PSF
Contribution from hardware failures
Contribution from systematic failures

Figure 13.6 Contributions to safety unavailability.

The PSF is approximately equal to the probability that an item that has just been proof tested fails on demand. Unavailability due to imperfect testing, such as partial stoke testing of valves, may adequately be included in the PSF.

13.4.1 Probability of Critical Situation

Consider a safety system that has been installed as a barrier against a specific type of hazardous events. We may, for example assume that the safety system is a fire detector system, and that the hazardous events are fires (in an early phase). Assume that fires occur randomly according to a homogeneous Poisson process (HPP) with intensity β. The parameter β denotes the mean number of fires per time unit and is sometimes called the *process demand rate*.

A *critical situation* occurs if a fire occurs while the fire detector system is in a failed state. This situation is shown in Figure 13.7 .

Each time a fire occurs, there is a probability SU that the fire detector system is not able to detect the fire. In Section 10.2, we show how to combine an HPP with Bernoulli trials, such that critical situations will occur as an HPP with intensity β SU.

Let $N_C(t)$ be the number of critical situations in the interval $(0, t)$. The probability of having n critical situations in the interval is

$$\Pr(N_C(t) = n) = \frac{(\beta \text{ SU } t)^n}{n!} e^{-\beta \text{ SU } t} \quad \text{for} \quad n = 0, 1, \dots . \tag{13.21}$$

The mean number of critical situations in the time interval $(0, t)$ is

$$E[N_C(t)] = \beta \text{ SU } t. \tag{13.22}$$

13.4.2 Spurious Trips

For many safety items, the rate of STs may be comparable, and even higher, than the rate of DU failures. STs usually imply significant costs and also reduce the confidence in the system.

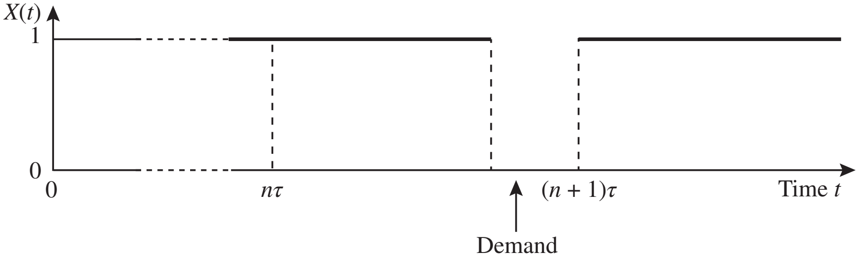


Figure 13.7 Critical situation – fire detector system. $X(t)$ is the state of the fire detector system.

Consider a safety system comprising m independent subsystems. The system may, for example comprise a flame detector subsystem, a heat detector subsystem, a smoke detector subsystem, a logic solver subsystem, and safety shutdown valves. Each subsystem may comprise several items. The system is considered to be a series structure of the subsystems with respect to ST failures. A subsystem ST failure will therefore give a system ST failure. Let $\lambda_{\text{ST}}^{(j)}$ be the rate of STs of the safety subsystem j, and let $\text{MDT}_{\text{ST}}^{(j)}$ denote the mean system downtime associated with the ST, for $j = 1, 2, \dots, m$. The safety unavailability of the system caused by STs is approximately

$$A_{\text{ST}}^{*} \approx \sum_{j=1}^{m} \lambda_{\text{ST}}^{(j)}\, \text{MDT}_{\text{ST}}^{(j)}. \tag{13.23}$$

Example 13.11 (Parallel structure)
Consider a sensor subsystem of n independent sensors. Sensor i has constant failure rate $\lambda_{\text{ST},i}$ with respect to STs, for $i = 1, 2, \dots, n$. The subsystem is a parallel structure with respect to safety, meaning that if one of the sensors is activated, the subsystem raises an alarm. The subsystem is therefore a 1oo$n$:G structure with respect to safety. With this configuration, a spurious signal (a false alarm) from any of the sensors will raise alarm. The subsystem is therefore a series ($n$oo$n$:G) structure with respect to STs, and the ST rate from the subsystem is

$$\lambda_{\text{ST}}^{1\text{oo}n} = \sum_{i=1}^{n} \lambda_{\text{ST},i}. \tag{13.24}$$

A high degree of redundancy may therefore lead to many STs. □

Example 13.12 (2oo3:G structure)
Consider a subsystem of three independent sensors of the same type, and let λ_{ST} be the constant failure rate with respect to STs from one sensor. The sensors are connected to a logic solver with a 2oo3:G voting logic. The system is illustrated in Figure 13.8. Two sensors have to send a signal to the logic solver to raise alarm. We assume that the logic solver is so reliable that failures may be neglected. Because the sensors are independent, STs (false alarms) occur as single failures. When a

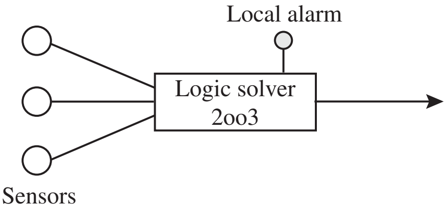


Figure 13.8 A 2oo3:G sensor system.

Table 13.2 PFD and spurious trip rate for three simple structures.

System	PFD	Rank	Spurious trip rate	Rank
Single item (1oo1)	$\frac{\lambda_{\text{DU}}\tau}{2}$	(3)	λ_{ST}	(2)
Parallel structure (1oo2)	$\frac{(\lambda_{\text{ST}}\tau)^2}{3}$	(1)	$2\lambda_{\text{ST}}$	(3)
2oo3 structure (2oo3:G)	$(\lambda_{\text{ST}}\tau)^2$	(2)	≈ 0	(1)

sensor gives a false alarm, the system gives a false alarm only if a second sensor gives a false alarm before the first false alarm is detected and repaired. Let us assume that when the logic solver receives a signal from a sensor, a local alarm is raised. The operators may therefore check the status and repair the sensor that has given the false alarm. Assume that the restoration time is t_r. If a second alarm is not received by the logic solver before the first failure is repaired, there will be no system false alarm. The ST (false alarm) rate from the 2oo3:G subsystem is, therefore,

$$\begin{aligned}\lambda_{\text{ST}}^{2\text{oo}3} &= 3\lambda_{\text{ST}} \int_0^{t_r} (1 - e^{-2\lambda_{\text{ST}}t})\, dt \\ &= 3\lambda_{\text{ST}}\ (1 - e^{-2\lambda_{\text{ST}}t_r}).\end{aligned} \tag{13.25}$$

Let $\lambda_{\text{ST}} = 5 \times 10^{-5}$ ST failures/h, and $t_r = 2$ hours. In this case, we get $\lambda_{\text{ST}}^{2\text{oo}3} \approx 1.5 \times 10^{-8}\ \text{h}^{-1}$, that is, a very low ST rate. □

Table 13.2 gives a brief comparison of three simple structures with independent items of the same type, with constant failure rate λ_{DU} with respect to DU failures and constant failure rate λ_{ST} with respect to ST. The test interval is τ. The 2oo3:G structure is often chosen as the best configuration for sensor systems because it has a PFD in the same order of magnitude as a parallel structure and because it can be made much more reliable than a parallel structure when it comes to STs.

13.4.3 Failures Detected by Diagnostic Self-Testing

Many failures of a modern SIS may be revealed by diagnostic self-testing. This applies both for dangerous failures and safe failures. The diagnostic testing is assumed to be carried out so frequently that the failures are revealed immediately. In subsystems with redundant items, a failure may sometimes be repaired, while the subsystem is online and is able to perform its safety function. In other cases, the subsystem has to be taken offline to repair the failure. Let $\lambda_{\text{DT},i}^{(j)}$ be the rate of failures of item i in subsystem j that are revealed by diagnostic self-testing, for $i = 1, 2, \dots, n_j$ and $j = 1, 2, \dots, m$. If we assume that all items are independent, then the rate of failures of subsystem j that are revealed by diagnostic self-testing is

$$\lambda_{\text{DT}}^{(j)} = \sum_{i=1}^{n_j} \lambda_{\text{DT},i}^{(j)}.$$

Let $\text{MDT}_{\text{DT}}^{(j)}$ be the MDT of subsystem j to repair a failure of an item in subsystem j that has been revealed by diagnostic self-testing. (For some configurations, the MDT may be zero.) The system unavailability caused by failures that are revealed by diagnostic self-testing is therefore

$$A_{\text{DT}}^{*} \approx \sum_{j=1}^{m} \lambda_{\text{DT}}^{(j)}\ \text{MDT}_{\text{DT}}^{(j)}. \tag{13.26}$$

In (13.26), the MDT is given for each subsystem. For subsystems with different types of items, it may be more appropriate to give the MDT associated to repair of each type of items.

The *diagnostic coverage* of the diagnostic self-test of item i is defined by

$$c_{\text{DT},i} = \frac{\lambda_{\text{DT},i}}{\lambda_i},$$

where λ_i is the total failure rate (of a specified category) of item i, for $i = 1, 2, \dots, n$. A diagnostic self-testing with test coverage 70%, hence, reveal 70% of all the failures of the item. The term “diagnostic coverage” is mainly used for dangerous failures, and is then the percentage of dangerous failures that can be detected by self-testing. The term may, however, also be used for safe failures.

Example 13.13 (Process shutdown valve)

Consider a PSD valve, as illustrated by the sketch in Figure 13.9. The valve has a fail-safe actuator and is held open by hydraulic pressure. When a process demand occurs, the logic solver sends an electrical signal to the solenoid valve to open and bleed off the hydraulic pressure. Diagnostic self-testing may be carried out by sending on/off electric signals to the solenoid valve. The solenoid valve will start to open and bleed off hydraulic pressure, and the shutdown valve will start to close. The movement of the valve actuator may be monitored by the logic solver. When the valve actuator has moved some few millimeters, full hydraulic pressure is again applied to the actuator and the valve will fully open. By this testing, we can reveal failures of the electrical cables, the solenoid valve, and the PSD valve. The test coverage for the electrical cables is 100%. The test coverage of the solenoid valve and the hydraulic flow depend on the design of the system and may be made close to 100%. This type of testing of the shutdown valve is called *partial stroke testing* and

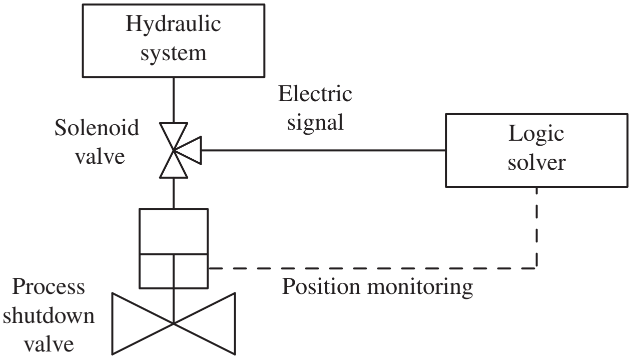


Figure 13.9 A process shutdown valve with fail-safe hydraulic actuator.

can only reveal some of the failure modes of the valve. The partial stroke testing will reveal some main causes of FTC failures, but it cannot reveal LCP failures.

In most applications, only the electrical cables will be tested by very frequent diagnostic testing. To avoid excessive wear of the valve seals, the diagnostic testing of the solenoid valve and the shutdown valve will be less frequent. □

13.5 Common Cause Failures

So far in this chapter, we have assumed that all items are independent. This is not always realistic in practice. Safety systems often have a high degree of redundancy, and the system reliability is therefore strongly influenced by potential CCFs. It is therefore important to identify potential CCFs and take the necessary precautions to prevent such failures.

Checklists that may be used to identify CCF problems of a SIS during its life cycle have been developed (e.g. see Summers and Raney 1999).

When we are able to identify the causes of CCFs, these should always be explicitly modeled, as illustrated in Example 13.14. In most cases, we are not able to find high quality input data for the explicitly modeled common causes. Even with low quality input data, or guesstimates, the result is usually more accurate than results obtained by including the explicit common causes into one of the general (implicit) dependent failure models that were introduced in Chapter 8.

Example 13.14 (Pressure sensors CCF)
Consider a parallel structure of two pressure sensors that are installed in a pressure vessel. Based on a search for potential causes for CCFs, we have identified two possible causes: (i) the common tap to the sensors is plugged with solids, and

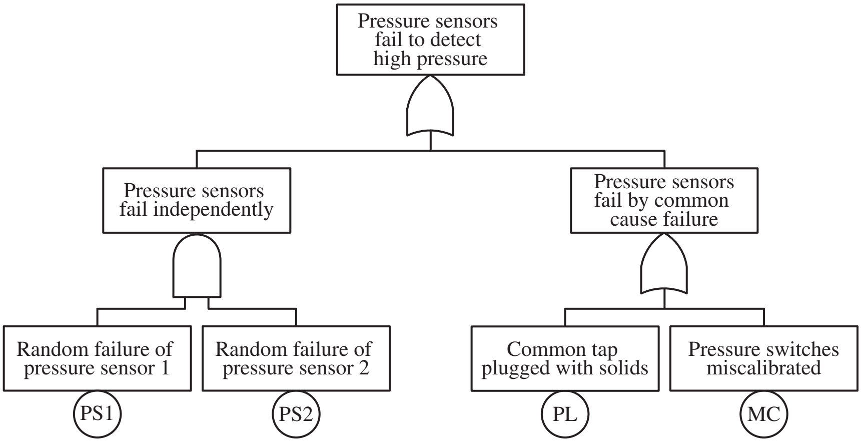

Pressure sensors fail to detect high pressure
Pressure sensors fail independently
Pressure sensors fail by common cause failure
Random failure of pressure sensor 1
Random failure of pressure sensor 2
Common tap plugged with solids
Pressure switches miscalibrated
PS1
PS2
PL
MC

Figure 13.10 Explicit modeling of a CCF for a system with two pressure sensors.

(ii) the sensors are miscalibrated. Other specific causes have not been identified. The two causes for CCFs may be modeled explicitly as illustrated by the fault tree in Figure 13.10. In the fault tree, the remaining failures of the sensors are said to be independent. If we believe that there are some implicit causes of dependency, in addition to the two explicit causes, this dependency may be modeled by one of the models discussed in Chapter 8, for example the β-factor model. □

The most commonly used (implicit) model for CCFs of safety systems is the β-factor model. In the β-factor model, we assume that a certain percentage of all failures are CCFs that cause all the items to fail at the same time (or within a very short time interval). The failure rate λ_{DU} with respect to DU failures may therefore be written as

$$\lambda_{\text{DU}} = \lambda_{\text{DU}}^{(i)} + \lambda_{\text{DU}}^{(c)},$$

where $\lambda_{\text{DU}}^{(i)}$ is the rate of independent DU failures that only affects one component, and $\lambda_{\text{DU}}^{(c)}$ is the rate of common cause DU failures that will cause failure of all the system components at the same time. The common cause factor

$$\beta_{\text{DU}} = \frac{\lambda_{\text{DU}}^{(c)}}{\lambda_{\text{DU}}}$$

is the percentage of common cause DU failures among all DU failures of a component.

Similarly, the ST rate λ_{ST} may be written as

$$\lambda_{\text{ST}} = \lambda_{\text{ST}}^{(i)} + \lambda_{\text{ST}}^{(c)},$$

where $\lambda_{\text{ST}}^{(i)}$ is the rate of independent ST failures that only affects one component, and $\lambda_{\text{DU}}^{(c)}$ is the rate of common cause ST failures that will cause failure of all the system components at the same time. The common cause factor

$$\beta_{\text{ST}} = \frac{\lambda_{\text{ST}}^{(c)}}{\lambda_{\text{ST}}}$$

is the percentage of common cause ST failures among all ST failures of a component. Because there may be different failure mechanisms leading to DU and ST failures, β_{DU} and β_{ST} need not be equal.

13.5.1 Diagnostic Self-Testing and CCFs

CCFs may be classified in two main types:

(1) Multiple failures that occur at the same time due to a common cause.
(2) Multiple failures that occur due to a common cause, but not necessarily at the same time.

As an example of type 2, consider a redundant structure of electronic components that are exposed to a common cause: increased temperature. The components will fail due to the common cause, but usually not at the same time. If we have an SIS with an adequate diagnostic coverage with respect to this type of failures, we may be able to detect the first CCF and take action before the system fails. A system failure due to the common cause may therefore be avoided.

Remark 13.2
If the common cause, increased temperature, is due to a cooling fan failure, this should be explicitly modeled as illustrated in Example 13.14. Monitoring the condition of the cooling fan would in this case give an earlier warning than diagnostic testing of the electronic components and a higher probability of successful shutdown before a system CCF occurs. A similar example is discussed in IEC61508-6 without mentioning any explicit modeling of the cooling fan. □

When we have identified the causes of potential CCFs (e.g. by applying a checklist), we should carefully split the potential CCFs in the two types (1 and 2) above. For each cause leading to failures of type 2, we should evaluate the ability of the diagnostic self-testing to reveal the failure (or the failure cause), the time required to take action, and the probability that this action will prevent a system failure.

It seems obvious that the common cause factor β for an SIS with good diagnostic coverage should be lower than for a system with no, or a poor, diagnostic coverage. We should therefore be careful and not use estimates for β from old-fashioned systems when analyzing a modern SIS with good diagnostic coverage.

Example 13.15 (Parallel structure)
Reconsider the parallel structure of two sensors in Example 13.4 and assume that DU failures occur with a common cause factor β_{DU}. The PFD of the parallel structure is from (13.10) and (13.13) approximately

$$\mathrm{PFD}(\beta_{\mathrm{DU}}) \approx \frac{[(1-\beta_{\mathrm{DU}})\lambda_{\mathrm{DU}}\tau]^2}{3} + \frac{\beta_{\mathrm{DU}}\lambda_{\mathrm{DU}}\tau}{2}. \tag{13.27}$$

With respect to STs, the system is a series structure, and the trip rate is therefore

$$\lambda_{\mathrm{ST}}^{1\mathrm{oo}2}(\beta_{\mathrm{ST}}) = (2-\beta_{\mathrm{ST}})\lambda_{\mathrm{ST}}. \tag{13.28}$$

The rate of STs will therefore decrease when β_{ST} increases.

By using the same data as in Example 13.4, $\lambda_{\mathrm{DU}} = 0.21 \times 10^{-6}\ \mathrm{h}^{-1}$ and $\tau = 2190$ hours, and $\beta_{\mathrm{DU}} = \beta_{\mathrm{ST}} = 0.10$, we get from (13.27)

$$\mathrm{PFD}(\beta_{\mathrm{DU}}) \approx 5.71 \times 10^{-8} + 2.30 \times 10^{-5} \approx 2.31 \times 10^{-5}.$$

Observe that with realistic estimates of λ_{DU} and τ, $\mathrm{PFD}_{\mathrm{DU}}$ is dominated by the common cause term in (13.27). We may therefore use the approximation

$$\mathrm{PFD}(\beta_{\mathrm{DU}}) \approx \frac{\beta_{\mathrm{DU}}\lambda_{\mathrm{DU}}\tau}{2},$$

when $\lambda_{\mathrm{DU}}\tau$ is small. □

Example 13.16 (2oo3 structure)
The PFD for a 2oo3:G structure is from (13.11) and (13.12)

$$\mathrm{PFD}(\beta_{\mathrm{DU}}) \approx [(1-\beta_{\mathrm{DU}})\lambda_{\mathrm{DU}}\tau]^2 + \frac{\beta_{\mathrm{DU}}\lambda_{\mathrm{DU}}\tau}{2}. \tag{13.29}$$

With a local alarm on the logic solver, we may avoid almost all independent STs. All CCFs will, on the other hand, result in a system ST, and we therefore have

$$\lambda_{\mathrm{ST}}^{2\mathrm{oo}3}(\beta_{\mathrm{ST}}) = \beta_{\mathrm{ST}}\lambda_{\mathrm{ST}}. \tag{13.30}$$

With the same data as in Example 13.15, we get from (13.29)

$$\mathrm{PFD}(\beta_{\mathrm{DU}}) \approx 1.71 \times 10^{-7} + 2.30 \times 10^{-5} \approx 2.32 \times 10^{-5}.$$

As in Example 13.15, we observe that with realistic estimates of λ_{DU} and τ, $\mathrm{PFD}_{\mathrm{DU}}$ is dominated by the common cause term in (13.29). We may therefore use the approximation

$$\mathrm{PFD}(\beta_{\mathrm{DU}}) \approx \frac{\beta_{\mathrm{DU}}\lambda_{\mathrm{DU}}\tau}{2},$$

when $\lambda_{\mathrm{DU}}\tau$ is small. □

In Examples 13.15 and 13.16, $\text{PFD}_{\text{DU}}(\beta_{\text{DU}})$ was dominated by the common cause term of the expressions (13.27) and (13.29), respectively, when $\lambda_{\text{DU}}\tau$ is small. It is straightforward to show that the same applies to all $k$oo$n$:G structures, where $n \geq 2$, and $k \leq n$. Therefore, we have

$$\text{PFD}^{k\text{oo}n}(\beta_{\text{DU}}) \approx \frac{\beta_{\text{DU}}\lambda_{\text{DU}}\tau}{2}, \tag{13.31}$$

when $\lambda_{\text{DU}}\tau$ is small. When $\beta_{\text{DU}} > 0$, we therefore get approximately the same result for all types of $k$oo$n$:G configurations, and the result is nearly independent of the number n of components, as long as $n \geq 2$. This may not be a realistic feature of the β-factor model. A more realistic alternative to the β-factor model has been proposed as part of the PDS approach that is described in Section 13.8.

IEC 61508 recommends using the β-factor model with a single “plant specific” β that is determined by using a checklist for all voting configurations (see IEC 61508-6, appendix D). This makes a comparison between different voting logics rather meaningless. Hokstad and Corneliussen (2004) criticize the β-factor model and introduced a multiple β-factor (MBF) model, that is a generalization of the β-factor model.

Remark 13.3

- Some reliability data sources (see Chapter 16) present the total failure rates, whereas other data sources only present the independent failure rates. The data in OREDA (2015) are collected from maintenance reports and contain all failures, both independent and CCFs. The data in MIL-HDBK-217F (1995) mainly come from laboratory testing of single components and therefore only presents the failure rate of independent failures. When using data from reliability data sources in CCF models, we should be aware of this difference.
- Some causes of CCFs, such as miscalibration of sensors, are equally likely for a single component as it is for a system of several components. If we include miscalibration as a cause of CCFs of n redundant sensors, it should also be included for a single sensor. This problem is further discussed by Summers and Raney (1999).

□

13.6 CCFs Between Groups and Subsystems

A voted group is a set of identical (or similar) components. Examples of voted groups are (1) a 2oo3:G voted group of pressure transmitters and (2) a 1oo2:G voted group of level transmitters. The methods described in this chapter are mainly focused on CCFs within a single group.

13.6.1 CCFs Between Voted Groups

A subsystem of a safety loop (or a SIS) may sometimes have more than one voted group. An example is a shutdown function (SIF) on a pressure vessel, with a sensor subsystem of both pressure transmitters [group 1] and level transmitters [group 2]. These two groups may be configured either with 1oo2:G voting or with 2oo2:G voting.

An intuitive approach would be to use the β-factor model (or the PDS model, see Section 13.8) for each voted group and determine factors β_1 and β_2 for group 1 and group 2, respectively, and thereafter, to determine a β-factor β_{12} to model possible CCFs between the two voted groups. The two types of β-factors are sometimes referred to as "inner" (i.e. within voted groups) and "outer" (i.e. between voted groups) β-factors.

A problem with this approach is that even if all components have constant failure rates, the voted groups will generally not have constant failure rates. This means that a main assumption of the β-factor model is not fulfilled.

13.6.2 CCFs Between Subsystems

The three main subsystems of a safety loop (or a SIS) may also be exposed to CCFs. The three subsystems are generally set up as a series structure.

Consider a series structure of two identical components with constant failure rate λ. The components are exposed to CCF that is modeled by a β-factor model with factor β. Because the failure rate of a series structure is the sum of the failure rates, the failure rate of the series structure is

$$\lambda_S = 2(1-\beta)\lambda + \beta\lambda = 2\lambda - \beta\lambda.$$

This means that a series structure that is exposed to CCFs has a lower failure rate, and a higher reliability, than a series structure of independent components, when using the β-factor model. This also means that assuming independence gives a conservative result for series structures.

This argument cannot be directly transferred to a series of subsystems, because the β-factor model does not easily apply to nonidentical subsystems with nonconstant and different failure rate functions.

13.7 IEC 61508

The international standard, IEC 61508 is the main standard for SISs. IEC 61508 is a generic, performance-based standard that covers most safety aspects of a SIS. As such, many topics covered in IEC 61508 are outside the scope of this book.

This section gives a brief presentation of some main aspects of IEC 61508 that are relevant for the theory and methods presented in this book.

IEC 61508 has seven parts:

Part 1	General requirements
Part 2	Requirements for E/E/PE safety-related systems
Part 3	Software requirements
Part 4	Definitions and abbreviations
Part 5	Examples of methods for the determination of safety integrity levels
Part 6	Guidelines on the application of IEC 61508-2 and IEC 61508-3
Part 7	Overview of techniques and measures

IEC 61508 gives safety requirements to SISs and provides guidance to validation and verification of such systems. The first three parts are normative parts and deal with the assessment of industrial process risk and the SIS hardware and software reliability. The remaining four parts deal with definitions and provide informative annexes to the standard.

Part 1 defines the overall performance-based criteria for an industrial process. It mandates the use of an overall safety lifecycle model (see Figure 13.11). Part 2 is directed toward manufacturers and integrators of SISs and presents methods and techniques that can be used to design, evaluate, and certify the hardware reliability of an SIS, and thus its contribution to process risk reduction.

IEC 61508 is a generic standard that is common to several industries. Application specific standards and guidelines are therefore developed, giving more specific requirements. Among these standards and guidelines are

IEC 61511 (2003)	Functional safety – Safety instrumented systems for the process industry
IEC 62061 (2005)	Safety of machinery – Functional safety of electrical, electronic, and programmable electronic systems
IEC 61513 (2011)	Nuclear power plants – Instrumentation and control important to safety – General requirements for systems
NOG (2018)	Guideline for the application of IEC 61508 and IEC 61511 in the petroleum activities on the Norwegian Continental Shelf

13.7.1 Safety Lifecycle

The requirements in IEC 61508 are related to an overall *safety lifecycle* outlining the main steps of the life cycle, similar to – but more detailed than – the overview of the reliability engineering process in Figure 1.8. Main steps of the safety lifecycle include

(a) Concept definition

(b) Overall scope definition
(c) Hazard and risk analysis
(d) Safety requirements specification
(e) Safety requirements allocation
(f) SIS design and development (with several substeps)
(g) Installation and commissioning
(h) Safety validation
(i) Operation and maintenance
(j) Decommissioning or disposal

Each step is thoroughly described in Part 1 of IEC 61508.

13.7.2 Safety Integrity Level

Safety integrity is a fundamental concept in IEC 61508 and may be defined as

Definition 13.1 (Safety integrity)
The ability of a safety-related system to achieve its required safety functions under all the stated conditions within a stated operational environment and within a stated duration (IEV 821-12-54). □

The safety integrity is classified into four discrete levels called *safety integrity levels* (SILs).

The SIL is in turn defined by the PFD. The relation between the SIL and the PFD is shown in Table 13.3.

- *Low demand mode* means that the frequency of demands for operation of the SIS is not greater that once per year, and not greater than twice the proof-test frequency

Table 13.3 Safety integrity levels for safety functions.

Safety integrity level (SIL)	Low demand mode of operation (average probability of failure to perform its design function on demand)	High demand mode or continuous mode of operation (probability of a dangerous failure per hour)
4	$\geq 10^{-5}$ to $< 10^{-4}$	$\geq 10^{-9}$ to $< 10^{-8}$
3	$\geq 10^{-4}$ to $< 10^{-3}$	$\geq 10^{-8}$ to $< 10^{-7}$
2	$\geq 10^{-3}$ to $< 10^{-2}$	$\geq 10^{-7}$ to $< 10^{-6}$
1	$\geq 10^{-2}$ to $< 10^{-1}$	$\geq 10^{-6}$ to $< 10^{-5}$

- *High demand or continuous mode* means that the frequency of demands for operation of the SIS is greater than once per year or greater than twice the proof-test frequency

ANSI/ISA-84.01 (1996) uses the same safety integrity levels as shown in Table 13.3 but clearly states that SIL 4 is not relevant in the process industry.

A SIL has to be assigned to each SIF. Observe that the SIL is assigned to the SIF, and not to the SIS, because a single SIS may perform several SIFs.

Assume that process demands for a SIF with low demand mode occur according to a HPP with rate ν demands per hour. For each demand, the SIF fails to perform the required function with probability PFD. A *critical situation* occurs if a process demand occurs and the SIF fails. Let $N_c(t)$ be the number of critical situations in the time interval $(0, t)$. The process $\{N_c(t), t > 0\}$ is therefore an HPP with rate $\nu_c = \nu$ PFD. The probability that n critical situations will occur in the interval $(0, t)$ is

$$\Pr(N_c(t) = n) = \frac{(\nu \text{ PFD } t)^n}{n!} \exp(-\nu \text{ PFD } t) \quad \text{for} \quad n = 0, 1, 2, \dots . \tag{13.32}$$

The mean time between critical situations is

$$\text{MTBF} = \frac{1}{\nu \text{ PFD}}. \tag{13.33}$$

When the mean time between demands is 10^4 hours (≈1.15 years), we observe that the mean time between critical situations is the same for a SIF with low demand mode as for a SIF with high demand mode, with the same SIL. The demand rate ν is usually defined as the net demand rate for the SIF, excluding the demands that are effectively taken care of by non-SIS protection layers and other risk reduction facilities.

The risk related to a specified critical event for a SIF with low demand mode is a function of (i) the potential consequences of a critical event, and (ii) the frequency of the critical event. To select an appropriate SIL, we therefore need to assess

(1) The frequency ν of demands for the SIS.
(2) The potential consequences following an occurrence of the critical event.

13.7.3 Compliance with IEC 61508

The overall objective of IEC 61508 is to identify the required SIFs, to establish the required SIL for each SIF, and to implement the safety functions in a SIS in order to achieve the desired safety level for the process. IEC 61508 is *risk based* and decisions taken shall be based on criteria related to risk reduction and tolerability of risk.

The objective and the requirements related to each of the lifecycle phases are described in detail in Section 7 of IEC 61508, Part 1. The actions that have to be carried out, and the extent of these actions, will vary with the type and complexity of the system (process). We have proposed a sequence of actions in the following. The described actions should be regarded as a supplement to the detailed requirements in the standard. Our proposed actions do not replace the requirements in the standard, but may hopefully give additional insight. When developing this sequence of actions, we have had the process section of an offshore oil/gas platform in mind. For other processes/applications, some of the actions might be reduced or omitted.

(1) *System definition*. We start with a conceptual design of the system. The conceptual design is assumed to be a basic design where no SIFs are implemented. The conceptual design is a (close to) final design that is described by process and instrument diagrams (P&IDs), other flow diagrams, and calculation results.

(2) *Definition of EUCs*. The system (process) must be broken down into suitable subsystems. The subsystems are the EUCs. Guidance on how to define EUCs is given in NOG (2018). Examples of suitable EUCs are pressure vessels, pumping stations, and compressors.

(3) *Risk acceptance criteria*. We have to define risk acceptance criteria, or *tolerable risk* criteria for each EUC. In some industries, such as the Norwegian oil and gas industry, risk acceptance criteria have to be defined on the plant (platform) level in the initial phases of a development project. The risk acceptance criteria are qualitative or quantitative criteria related to the risk to humans, the environment, and sometimes also related to material assets and production regularity. Risk acceptance criteria may, for example be formulated as "the fatal accident rate (FAR)[1] shall be less than nine," and "no release of toxic gas to the atmosphere with a probability of occurrence greater than 10^{-4} in one year.

The plant risk acceptance criteria have to be broken down and allocated to the various EUCs. The allocation of requirements must be based on criteria related to feasibility, fairness, and cost, and is generally not a straightforward task.

(4) *Hazard analysis*. A hazard analysis has to be carried out to identify all potential hazards and process demands[2] of each EUC. The hazard analysis may be carried out using methods, such as:

[1] FAR = Expected number of fatalities per 10^8 hours of exposure.

[2] A process demand is significant deviation from normal operation that can lead to adverse consequences for humans, the environment, material assets, or production regularity.

- Preliminary hazard analysis
- Hazard and operability (HAZOP) analysis (e.g. see IEC 61882 2016)
- Failure modes, effects, and criticality analysis (FMECA)
- Checklists

The hazard analysis provides

(a) A list of all potential process demands that may occur in the EUC.
(b) The direct causes of each process demand.
(c) Rough estimates of the frequency of each project demand.
(d) A rough assessment of the potential consequences of each process demand.
(e) Identification of non-SIS protection layers for each process demand.

The hazard analysis shall consider all reasonable, foreseeable circumstances including possible fault conditions, misuse and extreme environmental conditions. The hazard and risk analysis shall also consider possible human errors and abnormal or infrequent modes of operation of the EUC.

(5) *Quantitative risk assessment*. A quantified risk assessment is carried out to quantify the risk caused by the various process demands for the EUC and for the system (process). The risk assessment is carried out by methods, such as

- Fault tree analysis
- Event tree analysis
- Consequence analysis (e.g. fire and explosion loads)
- Simulation (e.g. accident escalation)

The quantitative risk assessment provides

(a) Estimates of the frequency of the process demands that were identified in step 4.
(b) Identification of potential consequences of each process demand and assessment of these consequences.
(c) Risk estimates related to each process demand and for the EUC.
(d) Requirements for risk reduction to meet the tolerable risk criteria for the EUC.

Note 1: The traditional quantitative risk analyses (QRAs) that are carried out for Norwegian offshore installations (NORSOK Z-013) do generally not meet all the requirements for risk assessment in IEC 61508.

Note 2: The QRA may partly be replaced with a layer of protection analysis (LOPA) (CCPS 2001).

(6) *Non-SIS layers of protection*. The required risk reduction may in some cases be obtained by non-SIS layers of protection. In this step, various non-SIS layers of protection (e.g. mechanical devices, fire walls) are identified and evaluated with respect to EUC risk reduction. Based on this step, we can decide whether or not a SIF is required to meet the risk acceptance criteria.

(7) *Determination of SIL*. The required SIL for each SIF is determined such that the risk reduction for the EUC may be obtained. Qualitative and quantitative approaches to the determination of SIL are provided in IEC 61508-5.

Note 3: The Norwegian offshore industry has proposed an alternative approach, where the risk assessments and the SIL determinations are carried out for a generic system. Based on these analyses, a minimum SIL is specified for each category of EUCs (NOG 2018).

(8) *Specifications and reliability requirements*. The specifications and reliability requirements of the SIFs have to be defined.

(9) *SIS design*. The SIS has to be designed according to the specifications. IEC 61511 give guidance on building an SIS with specific SIFs that meet a desired SIL.

(10) *PFD calculation*. Reliability models are established and the PFD calculated for the proposed SIS design.

(11) *Spurious trip assessment*. The frequency of ST failures of the proposed SIS design has to be estimated. Other potential, negative effects of the proposed SIS design should be evaluated. (This step is not required in IEC 61508).

(12) *Iteration*. We must now check that the proposed SIS design fulfills the criteria in step 9 and that the frequency of ST failures is acceptable. If not, the design has to be modified. Several iterations may be necessary.

(13) *System risk evaluation*. The system (process) risk reduction due to the proposed SIS is now assessed.

(14) *Verification*. The required modifications and analysis are made to ascertain that the proposed SIS meets the risk reduction (SIL) requirements.

Interested readers may find more information in Rausand (2014) and van Beurden and Goble (2018).

13.8 The PDS Method

The safety unavailability of a SIS with low demand mode may be assessed by the methods described in Sections 13.2 and 13.3. A more comprehensive approach has, however, been developed by SINTEF as part of the PDS[3] project. The PDS method (Hauge et al. 2013) is used to quantify both the reliability (the safety unavailability and the ST rate) and the life cycle cost of a SIS. The PDS method is compatible with the requirements in IEC 61508 and can be used to verify whether or not a specific SIL requirement is met.

[3] PDS is a Norwegian abbreviation for “Reliability of computer-based safety systems.”

13.9 Markov Approach

Consider a safety system that is tested periodically with test interval τ. When a failure is detected during a test, the system is repaired. The time required for testing and repair is considered to be negligible.

Let $X(t)$ be the state of the safety system at time t, and let $\mathcal{X} = \{0, 1, \dots, r\}$ be the (finite) set of all possible states. Assume that we can split the state space $\mathcal{X}$ into two parts, a set B of functioning states, and a set F of failed states, such that $F = \mathcal{X} - B$. The average PFD(n) of the system in test interval n is

$$\text{PFD}(n) = \frac{1}{\tau} \int_{(n-1)\tau}^{n\tau} \Pr(X(t) \in F)\, dt, \tag{13.34}$$

for $n = 1, 2, \dots$. If a demand for the safety system occurs in interval n, the (average) probability that the safety system is not able to shut down the EUC is PFD(n). The following approach is mainly based on Lindqvist and Amundrustad (1998).

We assume that $\{X(t)\}$ behaves like a time homogeneous continuous–time Markov chain (see Chapter 11) with transition rate matrix $\mathbb{A}$ as long as time runs inside a test interval, that is, inside intervals $(n-1)\tau \le t < n\tau$, for $n = 1, 2, \dots$. Let $P_{jk}(t) = \Pr(X(t) = k \mid X(0) = j)$ denote the transition probabilities for $j, k \in \mathcal{X}$, and let $\mathbb{P}(t)$ denote the corresponding matrix. Failures detected by diagnostic self-testing, and ST failures may occur and be repaired within the test interval.

Let $Y_n = X(n\tau-)$ be the state of the system immediately before time $n\tau$, that is, immediately before test n. If a malfunctioning state is detected during a test, a repair action is initiated, and changes the state from Y_n to a state Z_n, where Z_n is the state of the system just after the test (and possible repair) n. When Y_n is given, we assume that Z_n is independent of all transitions of the system before time $n\tau$. Let

$$\Pr(Z_n = j \mid Y_n = i) = R_{ij} \qquad \text{for all } i, j \in \mathcal{X} \tag{13.35}$$

denote the transition probabilities, and let $\mathbb{R}$ denote the corresponding transition matrix. If the state of the system is $Y_n = i$ just before test n, the matrix $\mathbb{R}$ tells us the probability that the system is in state $Z_n = j$ just after test/repair n. The matrix $\mathbb{R}$ depends on the repair strategy and also on the quality of the repair actions. Probabilities of maintenance-induced failures and imperfect repair may be included in $\mathbb{R}$. The matrix $\mathbb{R}$ is called the *repair matrix* of the system.

Example 13.17 (Safety valve)
Consider a safety valve that is located in the production tubing in an oil/gas production well. The valve is closed and tested for leakage at regular intervals. When the valve is closed, it may fail to reopen. That is, the failure mode FTO may occur. Experience has shown that a specific type of valves fails to reopen

approximately once every 200 tests. The probability of FTO-failure can easily be taken into account in the repair matrix $\mathbb{R}$. □

Let the distribution of the state of the safety system at time $t = 0$, $Z_0 \equiv X(0)$ be denoted by $\boldsymbol{\rho} = [\rho_0, \rho_1, \dots, \rho_r]$, where $\rho_i = \Pr(Z_0 = i)$, and $\sum_{i=0}^{r} \rho_i = 1$. The distribution of the state of the system just before the first test, at time τ, is

$$\begin{aligned}\Pr(Y_1 = k) &= \Pr(X(\tau-) = k)\\ &= \sum_{j=0}^{r} \Pr(X(\tau-) = k \mid X(0) = j)\ \Pr(X(0) = j)\\ &= \sum_{j=0}^{r} \rho_j\ P_{jk}(\tau) = [\rho\ \mathbb{P}(\tau)]_k, \qquad (13.36)\end{aligned}$$

for any $k \in \mathcal{X}$, where $[\mathbf{B}]_k$ denotes the kth entry of the vector $\mathbf{B}$.

Let us now consider a test interval n (≥ 1). Just after test interval n the state of the system is Z_n. We assume that the continuous-time Markov chain in $n\tau \leq t < (n+1)\tau$, given its initial state Z_n, is independent of all transitions that have taken place before time $n\tau$.

$$\begin{aligned}&\Pr(Y_{n+1} = k \mid Y_n = j)\\ &\quad= \sum_{i=0}^{r} \Pr(Y_{n+1} = k \mid Z_n = i, Y_n = j)\ \Pr(Z_n = i \mid Y_n = j)\\ &\quad= \sum_{i=0}^{r} P_{ik}(\tau) R_{ji} = [\mathbb{R}\ \mathbb{P}(\tau)]_{jk}, \qquad (13.37)\end{aligned}$$

where $[\mathbb{B}]_{jk}$ denotes the (jk)th entry of the matrix $\mathbb{B}$. It follows that $\{Y_n, n = 0, 1, \dots\}$ is a discrete-time Markov chain with transition matrix

$$\mathbb{Q} = \mathbb{R}\ \mathbb{P}(\tau). \qquad (13.38)$$

In the same way,

$$\begin{aligned}&\Pr(Z_{n+1} = k \mid Z_n = j)\\ &\quad= \sum_{i=0}^{r} \Pr(Z_{n+1} = k \mid Y_{n+1} = i, Z_n = j)\ \Pr(Y_{n+1} = i \mid Z_n = j)\\ &\quad= \sum_{i=0}^{r} P_{ji}(\tau)\ R_{ik} = [\mathbb{P}(\tau)\ \mathbb{R}]_{jk}, \qquad (13.39)\end{aligned}$$

and $\{Z_n, n = 0, 1, \dots\}$ is a discrete-time Markov chain with transition matrix

$$\mathbb{T} = \mathbb{P}(\tau)\ \mathbb{R}. \qquad (13.40)$$

Let $\pi = [\pi_0, \pi_1, \dots, \pi_r]$ denote the stationary distribution of the Markov chain $\{Y_n, n = 0, 1, \dots\}$. Then π is the unique probability vector satisfying the equation

$$\pi\ \mathbb{Q} \equiv \pi\ \mathbb{R}\ \mathbb{P}(\tau) = \pi, \tag{13.41}$$

where π_i is the long-term proportion of times the system is in state i just before a test.

In the same way, let $\gamma = [\gamma_0, \gamma_1, \dots, \gamma_r]$ denote the stationary distribution of the Markov chain $\{Z_n, n = 0, 1, \dots\}$. Then γ is the unique probability vector satisfying the equation

$$\gamma\ \mathbb{T} \equiv \gamma\ \mathbb{P}(\tau)\ \mathbb{R} = \gamma, \tag{13.42}$$

where γ_i is the long-term proportion of times the system is in state i just after a test/repair.

Let F denote the set of all states representing a DU failure in $\mathcal{X}$, and define $\pi_F = \sum_{i \in F} \pi_i$. Then, π_F denotes the long-run proportion of times the system is in a dangerously failed state immediately before a test. If, for example, $\pi_F = 5 \times 10^{-3}$, the system will have a critical failure, on the average, in one out of 200 tests. Moreover, $1/\pi_F$ is the mean time, in the long run, between visits to F (measured with time unit τ). The mean time between DU failures is hence

$$\text{MTBF}_{\text{DU}} = \frac{\tau}{\pi_F}, \tag{13.43}$$

and the average rate of DU failures is

$$\lambda_{\text{DU}} = \frac{1}{\text{MTBF}_{\text{DU}}} = \frac{\pi_F}{\tau}. \tag{13.44}$$

The average PFD in interval n, PFD(n) may now be expressed as

$$\begin{aligned} \text{PFD}(n) &= \frac{1}{\tau} \int_{(n-1)\tau}^{n\tau} \Pr(X(t) \in F)\ dt \\ &= \frac{1}{\tau} \int_0^{\tau} \sum_{j=0}^{r} \sum_{k \in F} P_{jk}(t)\ \Pr(Z_n = j)\ dt. \end{aligned} \tag{13.45}$$

Because $\Pr(Z_n = j) \to \gamma_j$ when $n \to \infty$, we get the long-term average PFD as

$$\text{PFD} = \lim_{n \to \infty} \text{PFD}(n) = \frac{1}{\tau} \int_0^{\tau} \sum_{j=0}^{r} \sum_{k \in F} P_{jk}(t)\ \gamma_j\ dt = \sum_{j=0}^{r} \gamma_j Q_j, \tag{13.46}$$

where

$$Q_j = \frac{1}{\tau} \int_0^{\tau} \sum_{k \in F} P_{jk}(t)\ dt$$

is the PFD given that the system is in state j at the beginning of the test interval.

Example 13.18 Hokstad and Frøvig (1996) consider a single component that is subject to various types of failure mechanisms. In one of their examples, they study a component with the following states:

State	Description
3	Component as-good-as-new
2	Degraded (noncritical) failure
1	Critical failure caused by sudden shock
0	Critical failure caused by degradation

The component is able to perform its intended function when it is in state 3 or state 2 and has a critical failure if it is in state 1 or state 0. State 1 is produced by a random shock, whereas state 0 is produced by degradation. In state 2, the component is able to perform its intended function, but has a specified level of degradation.

It is assumed that the continuous-time Markov chain is defined by the state transition diagram in Figure 13.11 and the transition rate matrix

$$\mathbb{A} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \lambda_{\text{dc}} & \lambda_s & -(\lambda_{\text{dc}} + \lambda_s) & 0 \\ 0 & \lambda_s & \lambda_d & -(\lambda_s + \lambda_d) \end{pmatrix},$$

where λ_s is the rate of failures caused by a random shock, λ_d is the rate of degradation failures, and λ_{dc} is the rate of degraded failures that become critical.

Because no repair is performed within the test interval, the failed states 0 and 1 are absorbing states. Let us assume that we know that the system is in state 3 at time 0, such that $\rho = [1, 0, 0, 0]$. We may now use the methods outlined in Section 11.9 to solve the forward Kolmogorov equations $\mathbf{P}(t)\ \mathbb{A} = \dot{\mathbf{P}}(t)$ and find the distribution $\mathbb{P}(t)$. It is clear that $\mathbb{P}(t)$ can be written as

$$\mathbb{P}(t) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ P_{20}(t) & P_{21}(t) & P_{22}(t) & 0 \\ P_{30}(t) & P_{31}(t) & P_{32}(t) & P_{33}(t) \end{pmatrix}.$$

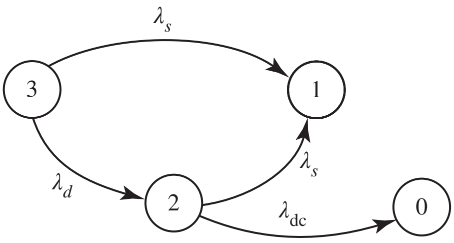


Figure 13.11 State transition diagram for the failure process described by Hokstad and Frøvig (1996).

The first two rows of $\mathbb{P}(t)$ are obvious because state 0 and state 1 are absorbing. The entry $P_{23}(t) = 0$ because it is impossible to have a transition from state 2 to state 3. From the state transition diagram, the diagonal entries are seen to be

$$P_{22}(t) = e^{-(\lambda_s+\lambda_{\text{dc}})t}$$
$$P_{33}(t) = e^{-(\lambda_s+\lambda_d)t}.$$

The remaining entries were shown by Lindqvist and Amundrustad (1998) to be

$$P_{20}(t) = \frac{\lambda_{\text{dc}}}{\lambda_s + \lambda_{\text{dc}}}(1 - e^{-(\lambda_s+\lambda_{\text{dc}})t})$$

$$P_{21}(t) = \frac{\lambda_s}{\lambda_s + \lambda_{\text{dc}}}(1 - e^{-(\lambda_s+\lambda_{\text{dc}})t})$$

$$P_{30}(t) = \frac{\lambda_d\lambda_{\text{dc}}}{(\lambda_d + \lambda_s)(\lambda_s + \lambda_{\text{dc}})} + \frac{\lambda_d\lambda_{\text{dc}}}{(\lambda_d - \lambda_{\text{dc}})(\lambda_d + \lambda_s)}\ e^{-(\lambda_s+\lambda_d)t} + \frac{\lambda_d\lambda_{\text{dc}}}{(\lambda_{\text{dc}} - \lambda_d)(\lambda_s + \lambda_{\text{dc}})}\ e^{-(\lambda_s+\lambda_{\text{dc}})t}$$

$$P_{31}(t) = \frac{\lambda_s(\lambda_d + \lambda_s + \lambda_{\text{dc}})}{(\lambda_d + \lambda_s)(\lambda_s + \lambda_{\text{dc}})} + \frac{\lambda_s\lambda_{\text{dc}}}{(\lambda_d - \lambda_{\text{dc}})(\lambda_d + \lambda_s)}\ e^{-(\lambda_s+\lambda_d)t} + \frac{\lambda_s\lambda_d}{(\lambda_{\text{dc}} - \lambda_d)(\lambda_s + \lambda_{\text{dc}})}\ e^{-(\lambda_s+\lambda_{\text{dc}})t}$$

$$P_{32}(t) = \frac{\lambda_d}{\lambda_d - \lambda_{\text{dc}}}(e^{-(\lambda_s+\lambda_{\text{dc}})t} - e^{-(\lambda_s+\lambda_d)t}).$$

Several repair policies may be adopted,

(1) All failures are repaired after each test, such that system always starts in state 3 after each test.
(2) All critical failures are repaired after each test. In this case, the system may have a degraded failure when it starts up after the test.
(3) The repair action may be imperfect, meaning that there is a probability that the failure will not be repaired.

13.9.1 All Failures are Repaired After Each Test

In this case all failures are repaired, and we assume that the repair is perfect, such that the system will be in state 3 after each test. The corresponding repair matrix $\mathbb{R}_1$ is therefore

$$\mathbb{R}_1 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

With this policy, all test intervals have the same stochastic properties. The average PFD is therefore given by

$$\text{PFD} = \frac{1}{\tau}\int_0^{\tau} (P_{31}(t) + P_{30}(t))\, dt.$$

13.9.2 All Critical Failures Are Repaired after Each Test

In this case, the $\mathbb{R}$ matrix is

$$\mathbb{R}_2 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

13.9.3 Imperfect Repair after Each Test

In this case, the $\mathbb{R}$ matrix is

$$\mathbb{R}_3 = \begin{pmatrix} r_0 & 0 & 0 & 1 - r_0 \\ 0 & r_1 & 0 & 1 - r_1 \\ 0 & 0 & r_2 & 1 - r_2 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The PFD may be found from (13.46). The calculation is straightforward, but the expressions become rather complicated and are not included here. Further results are given by Lindqvist and Amundrustad (1998). □

13.10 Problems

13.1 Figure 13.12 shows a part of a smoke detection system. The system comprises two optical smoke detectors (with separate batteries) and a start relay. All components are assumed to be independent with constant failure rates:

Smoke detectors 1 and 2:	$\lambda_{\text{SD}} = 2 \times 10^{-4}$ failures/h
Start relay:	$\lambda_{\text{SR}} = 5 \times 10^{-5}$ failures/h

The system is tested and, if necessary, repaired after time intervals of equal length $\tau = 1$ month. After each test (repair), the system is considered to be as-good-as-new. The repair time is assumed to be negligible. Dangerous undetected (DU) failures are only detected during tests.

(a) Find the PFD for the system.

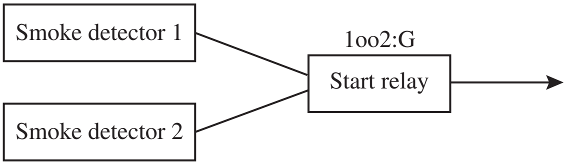


Figure 13.12 Smoke detector system (simplified).

(b) Find the mean number of test intervals the system passes from $t = 0$ until the first DU failure.

(c) Assume that you in a specific test find that the system has a DU failure. Determine the mean time the system has been in a failed state.

(d) Assume that fires occur as a homogenous Poisson process with intensity $\lambda = 1$ fire per 10 years. Find the probability that a fire occurs while a DU failure of the smoke detection system is present, during a period of 50 years.

13.2 Reconsider the 1oo2:G structure of independent fire detectors in Example 13.4, but assume that the two fire detectors are different and have failure rates $\lambda_{\mathrm{DU},1}$ and $\lambda_{\mathrm{DU},2}$ respectively, with respect to DU failures. The fire detectors are tested at the same time with test interval τ.

(a) Find the PFD for the fire detector system.

(b) Find an approximation to the PFD when $\lambda_{\mathrm{DU},i}\ \tau$ is "small," for $i = 1, 2$.

13.3 The β-factor model is often the preferred way of including CCFs due to its simplicity, but the model has some questionable properties:

(a) Comment on the effect on the independent failure rate when introducing measures to reduce the value of β. Why is this effect questionable?

(b) A 1oo4:G voted structure and a 2oo4:G voted structure would obtain approximately the same value for PFD, assuming identical components and the same β. Why is this the case, and what is the realism in having this effect on the PFD? In what situations would the effect be a realistic scenario, and in what situations would it be less realistic?

13.4 Reconsider the 2oo3:G structure of independent fire detectors in Example 13.5 but assume that the three fire detectors are different and have failure rates $\lambda_{\mathrm{DU},1}$, $\lambda_{\mathrm{DU},2}$, and $\lambda_{\mathrm{DU},3}$ respectively, with respect to DU failures. The fire detectors are tested at the same time with test interval τ.

(a) Find the PFD for the fire detector system.

(b) Find an approximation to the PFD when $\lambda_{\mathrm{DU},i}\ \tau$ is "small" for $i = 1, 2, 3$.

13.5 Will a 2oo4:G structure of identical and independent items give more or less spurious trips than a 2oo3:G structure of the same type of items? Justify your answer.

13.6 You are planning to install a pressure sensor system on a pressure vessel. From past experience, you know that the pressure sensors you are planning to use, have the following constant failure rates with respect to the actual failure modes:

No signal when the pressure increases beyond the pressure setting:	$\lambda_{\text{FTF}} = 3.10 \times 10^{-6}$ failures/h
False high pressure signal:	$\lambda_{\text{FA}} = 3.60 \times 10^{-6}$ failures/h

The pressure sensors will be connected to a logic unit (LU). The LU transforms the incoming signals and transmits them to the emergency shutdown (ESD) system. The failure rates of the LU are estimated to be

Does not transmit correct signal:	$\lambda_{\text{A}} = 0.10 \times 10^{-6}$ failures/h per input
False high pressure signal out:	$\lambda_{\text{B}} = 0.05 \times 10^{-6}$ failures/h

Four different system configurations are considered:

- One single pressure sensor (with LU)
- Two pressure sensors in parallel
- Three pressure sensors as a 2oo3:G structure
- Four pressure sensors as a 2oo4:G structure

The pressure sensors and the logic unit are tested and, if necessary, repaired at the same time once a month. Dangerous undetected (DU) failures are only detected during tests. After a test (repair), all items are assumed to be as-good-as-new. The time required for testing and repair is assumed to be negligible.

(a) Determine the PFD with respect to DU failures for each of the four system configurations when you assume that all items are independent, and the failure rates of cables, and so on, are negligible.

(b) Determine the probability of getting at least one false alarm (FA) from each of the four system configurations during a period of one year.

(c) Which of the four system configurations would you install?

13.7 Consider staggered testing as introduced in Section 13.3.5 for a parallel structure of two items with DU failure rates $\lambda_{\text{DU},1}$ and $\lambda_{\text{DU},2}$, respectively. The test interval is τ and the staggered delay is $t_0 < \tau$.

(a) Develop the formula for PFD(t_0). All steps in the development shall be shown.

(b) Find the formula for the optimal staggered delay t_0 as a function of the DU failure rates and the test interval.

(c) Show that if the two items have the same failure rate, the optimal staggered delay is $t_0 = \tau/2$ and provide intuitive arguments for this result.

13.8 Consider a parallel structure of n identical components with constant failure rates λ. The system is put into operation at time $t = 0$. The system is tested and if necessary repaired after regular time intervals of length τ. After a test (repair) the system is considered to be as-good-as-new. The system is exposed to CCFs that may be modeled by a β-factor model. Let PFD_n denote the PFD of a parallel structure of order n.

(a) Determine PFD_n as a function of λ, τ, and β.

(b) Let $\lambda = 5 \times 10^{-5}$ failures/h, and $\tau = 3$ months, and make a sketch of PFD_n as a function of β for $n = 2$ and $n = 3$.

(c) With the same data as in question (b), determine the difference between PFD_2 and PFD_3 when $\beta = 0$, and $\beta = 0.20$, respectively.

13.9 List the main pros and cons related to using a Markov model to model the reliability of a SIS.

13.10 Figure 13.13 shows a part of a shutdown system of a process plant. There are two process sections, A, and B. If a fire occurs in one of the process sections, the emergency shutdown (ESD) system is installed to close the emergency shutdown valve, ESDV. The ESD valve has a failsafe hydraulic actuator. The valve is held open by hydraulic pressure. When the hydraulic pressure is bled off, the valve closes.

Each process section has two redundant detector circuits (circuits 1 and 2). Each detector circuit is connected to the ESDV actuator by a pilot valve, which by signal from the detectors opens and bleeds off the hydraulic pressure in the ESDV actuator, and thereby closes the ESD valve. Further, each circuit comprises an input card, a central processing unit (CPU), an output card, and two fire detectors in each process section. When a fire detector is activated, the current in that circuit is broken. When the current to the input card is broken, a "message" is sent to the CPU via the output card to open the pilot valve. It is assumed that minor fires in one of the process sections cannot be detected by the fire detectors in the other process section.

It is assumed that all the components are independent with constant failure rates. Each component has two different failure modes:

- Fail to function (FTF) (i.e. no reaction when a signal is received)
- False alarm

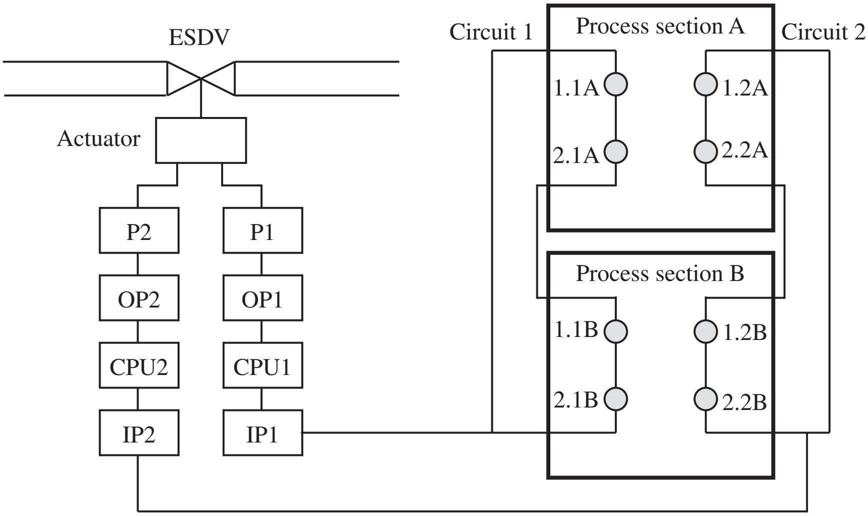


Figure 13.13 Sketch of an emergency shutdown system.

Table 13.4 Failure rates for the "fail to function" mode.

Component	Symbol	FTF-failure rate λ (failures/h)
ESD-valve	ESDV	3.0×10^{-6}
Actuator	Actuator	5.0×10^{-6}
Pilot valve	P1, P2	2.0×10^{-6}
Output card	OP1, OP2	0.1×10^{-7}
Input card	IP1, IP2	0.1×10^{-7}
CPU	CPU1, CPU2	0.1×10^{-7}
Fire detector	1.1A, 1.2A, 2.1A, 2.2A 1.1B, 1.2B, 2.1B, 2.2B	4.0×10^{-6}

The system components, their symbols, and FTF failure rates are listed in Table 13.4.

(a) Construct a fault tree with respect to the TOP event: "The ESD valve does not close when a fire occurs in process section A."
Write down the extra assumptions you have to make during the fault tree construction. As seen from Table 13.4, the failure rates of the input card, the CPU, and the output card are negligible compared to the failure rates of the other components. To simplify the fault tree

construction, you may therefore disregard the input/output cards and the CPU.

Show that the fault tree has the following minimal cut sets:

{Actuator}
{ESDV}
{P1, P2}
{P2, 1.1A, 2.1A}
{P1, 1.2A, 2.2A }
{1.1A, 1.2A, 2.1A, 2.2A }

All the components are tested once a month. FTF failures are normally only detected during tests. The time required for testing and, if necessary, repair is assumed to be negligible compared to the length of the testing interval. In question (b) we shall assume that the testing of the various components are carried out at different, and for us unknown, times.

(b) A. Determine the PFD for each of the relevant components.
   B. Determine the TOP event probability by the "upper bound approximation," when the basic events of the fault tree are assumed to be independent.
   C. Discuss the accuracy of the "upper bound approximation" in this case.
   D. Describe other, and more exact methods, to compute the TOP event probability. Discuss pros and cons for each of these methods.

(c) Minor fires are assumed to occur in process section A on the average two times a year, according to a HPP. A *critical situation* occurs when a fire occurs at the same time as the ESD system has FTF failure (i.e. when the TOP event is present). Find the probability of at least one such a *critical situation* during a period of 10 years.

(d) Next consider the subsystem comprising the two fire detectors 1.1A and 2.1A. Determine the PFD of this subsystem when the detectors are tested:
   (i) Once every third month at different and, for us, unknown time points.
   (ii) At the same time once every third month.
   (iii) By staggered testing, where detector 1.1A is tested once every third month and detector 2.1A is also tested once every third month, but always one month later than detector 1.1A.

   Which of these testing regimes would you prefer (give pros and cons). Explain why the PFD in case (i) is different from the PFD in case (ii).

(e) Do you consider the suggested system structure to be optimal with respect to avoid “False alarm” failures? Suggest an improved structure and discuss possible positive and negative properties of this structure.

13.11 A downhole safety valve (DHSV) is placed in the oil/gas production tubing on offshore production platforms, approximately 50–100 m below the sea floor. The valve is held open by hydraulic pressure through a 1/16″ hydraulic pipeline from the platform. When the hydraulic pressure is bled off, the valve closes by spring force. The valve is thus *failsafe close*. The valve is the last barrier against blowouts in case of an emergency situation on the platform. It is very important that the valve is functioning as a safety barrier, and the valve is therefore tested at regular intervals.

There are two main types of DHSVs; wireline retrievable (WR) valves, and tubing retrievable (TR) valves. WR valves are locked in a landing nipple in the tubing, and may be installed and retrieved by a wireline operation from the platform. A TR valve is an integrated part of the tubing. To retrieve a TR valve, the tubing has to be pulled. Here we shall consider a WR valve. When the WR valve fails, it is retrieved by a wireline operation and a new valve of the same type is installed in the same nipple.

The DHSV is tested once a month. During the testing, which requires approximately 1.5 hours, the production has to be closed down. The mean time to repair a failure is estimated to be nine hours.

The DHSV has four main failure modes:

FTC	Fail to close on command
LCP	Leakage in closed position
FTO	Fail to open on command
PC	Premature closure

The failure modes FTC and LCP are critical with respect to safety. The failure modes FTO and PC are noncritical with respect to safety, but will stop the production. The three failure modes FTC, LCP, and FTO may only be detected during testing, whereas PC failures are detected at once because the production from the well closes down.

The following failure mode distribution has been discovered:

FTC	15%
LCP	20%
FTO	15%
PC	50%

The failure rates are assumed to be constant with respect to all failure modes. The mean time between valve failures (with respect to all failure modes) has been estimated to 44 months.

If a critical failure is detected during a test, the well will be unsafe during approximately 1/3 of the repair time. If a noncritical failure is detected, the well will be safe during these operations.

(a) Determine the mean time between FTC failures of a valve.

(b) Determine the probability that a valve survives a test interval without any failure.

(c) Find the PFD. The time required for testing and repair shall be taken into account. Discuss the complications encountered in this calculation due to PC failures.

(d) Find the mean proportion of time the production is shut down due to DHSV testing and failures.

(e) Assume now that an emergency situation occurs on the platform on the average once every 50 platform years, which requires that the DHSV must be closed. A *critical situation* occurs when such an emergency situation occurs when the DHSV is not functioning as a safety barrier. Compute the mean time between this types of *critical situations*.

(f) Consider a platform with 20 production wells, with a DHSV in each well. In an emergency situation, all the wells have to be closed down. With the same assumptions as above, determine the mean time between *critical situations* on the platform.

13.12 A gas detector has constant failure rate $\lambda_{\text{DU}} = 1.8 \times 10^{-6}\ \text{h}^{-1}$ with respect to the critical (DU) failure mode "gas detector does not raise alarm when gas is present." Please record any extra assumptions you have to make to answer the questions below.

(a) Find the mean time-to-failure, MTTF, of the gas detector (with respect to DU failures).

(b) The critical failure mode is a so-called hidden failure. The gas detector is therefore proof tested after regular intervals of length $\tau = 6$ months (where 1 month = 730 hours). The time required to test and repair a failed detector is so short that it may be neglected. After a test/repair, the gas detector is assumed to be "as-good-as-new."

- Explain what is meant by a "hidden failure."
- Determine the PFD ("probability of failure on demand") for the gas detector.

(c) Assume now that we have three gas detectors of the same type. The three detectors are connected to a logic solver with a 2oo3 logic. The gas detectors are proof tested at the same time every six months. Otherwise, the same assumptions as in (b) apply. The logic solver is assumed to be so reliable that its failure rate may be set to zero. In this question, we assume that the three detectors are independent.
  - Find the probability that the 2oo3:G structure survives 12 months without a critical system failure.
  - Find the PFD for the 2oo3:G structure.
  - How many hours per year are we, on the average, unprotected by the gas detector system when we assume that the system shall be functioning continuously?

(d) Assume that the gas detectors are exposed to common cause failures that can be modeled by a β-factor model with $\beta = 0.08$.
  - Find the PFD of the 2oo3:G structure in this case. Specify the proportion of the PFD that is caused by independent failures and the proportion caused by common cause failures.
  - When a critical failure of the gas detector system is revealed in a proof test, how long time can we expect that the system has been unable to function?

References

ANSI/ISA-84.01 (1996). *Application of safety instrumented systems for the process industries*, *American National Standard 84.01*, ANSI/ISA. Research Triangle Park, NC 27709: American National Standard.

van Beurden, I. and Goble, W.M. (2018). *Safety Instrumented System Design: Techniques and Design Verification*. Research Triangle Park, NC: International Society of Automation (ISA).

Brissaud, F., Barros, A., and Berenguer, C. (2012). Probability of failure on demand of safety systems: impact of partial test distribution. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability* 226 (4): 426–436.

CCPS (2001). *Layer of Protection Analysis: Simplified Process Risk Assessment*. New York: American Institute of Chemical Engineers, Center for Chemical Process Safety.

Hauge, S., Kråknes, T., Håbrekke, S., and Jin, H. (2013). Reliability prediction methods for safety instrumented systems, PDS method handbook. In: *Handbook*. Trondheim: SINTEF.

Hokstad, P.R. and Corneliussen, K. (2004). Loss of safety assessment and the IEC 61508 standard. *Reliability Engineering & System Safety* 83: 111–120.

Hokstad, P.R. and Frøvig, A.T. (1996). The modelling of degraded and critical failures for components with dormant failures. *Reliability Engineering & System Safety* 51: 189–199.

IEC 61508 (2010). *Functional safety of electrical/electronic/programmable electronic safety-related systems.* Parts 1-7, *International standard.* Geneva: International Electrotechnical Commission.

IEC 61511 (2003). *Functional safety – safety instrumented systems for the process industry*, *International standard.* Geneva: International Electrotechnical Commission.

IEC 61513 (2011). *Nuclear power plants – instrumentation and control important to safety – general requirements for systems*, *International standard*. Geneva: International Electrotechnical Commission.

IEC 61882 (2016). *Hazard and operability studies (HAZOP studies) – application guide*, *International standard.* Geneva: International Electrotechnical Commission.

IEC 62061 (2005). *Safety of machinery – functional safety of safety-related electrical, electronic and programmable electronic control systems*, *International standard.* Geneva: International Electrotechnical Commission.

Lindqvist, B.H. and Amundrustad, H. (1998). Markov model for periodically tested components. In: *Proceedings of the European Conference on Safety and Reliability – ESREL'98* (ed. S. Lydersen, G.K. Hansen, and H.A. Sandtorv). Boston, MA: Balkema. 191–197.

Liu, Y. (2014). Optimal staggered testing strategies for heterogeneously redundant safety systems. *Reliability Engineering & System Safety* 126: 65–71.

Lundteigen, M.A. and Rausand, M. (2008). Partial stroke testing of process shutdown valves: how to determine the test coverage. *Journal of Loss Prevention in the Process Industries* 21: 579–588.

MIL-HDBK-217F (1995). Reliability Prediction of Electronic Equipment. *Military Handbook*. Washington, DC: U.S. Department of Defense.

NOG (2018). *Application of IEC61508 and IEC61511 in the Norwegian Petroleum Industry. Guideline 070*. Stavanger, Norway: Norwegian Oil and Gas.

OREDA (2015). *Offshore and Onshore Reliability Data*, 6e. 1322 Høvik, Norway: OREDA Participants, DNV GL.

Rausand, M. (2014). *Reliability of Safety-Critical Systems: Theory and Applications.* Hoboken, NJ: Wiley.

Srivastav, H., De Azevedo Vale, G., Barros, A. et al. (2018). Optimization of periodic inspection time of sis subject to a regular proof testing. *Proceedings of ESREL 2018*, Trondheim, Norway, pp. 1125–1131.

Summers, A.E. and Raney, G. (1999). Common cause and common sense, designing failure out of your safety instrumented system (SIS). *ISA Transactions* 38: 291–299.

Wu, S., Zhang, L., Barros, A. et al. (2018). Performance analysis for subsea blind shear ram preventers subject to testing strategies. *Reliability Engineering & System Safety* 169: 281–298.

14

Reliability Data Analysis

14.1 Introduction

This chapter gives an introduction to *reliability data analysis*, also known as *survival analysis* and *lifetime analysis*. The dataset to be analyzed consists of lifetimes that are measured from a starting time to an endpoint of interest. The starting time is usually the time when the item is put into operation for the first time, but may also be the time when we start observing the item. The endpoint of interest is usually a failure event, sometimes restricted to a specific failure mode.

For many datasets, the data collection is stopped before all the items fail. This means that some items are still in a functioning state when the data collection stops. The recorded times for such items are therefore not times-to-failure, but the time measured from the starting time until the data collection stopped. These times are said to be *censored* times, and a dataset with one or more censored times is called a *censored dataset*.

Some datasets include one or more explanatory variables such as pressure, temperature, flow-rate, and vibration. These variables are called *covariates* and help explain why there are differences between the times-to-failure of the same type of items.

Reliability data analysis is a loosely defined term that encompasses a variety of statistical methods for analyzing positive-valued datasets. The methods presented in this chapter are also extensively used in biostatistics and medical research under the heading survival analysis.

A high number of books have been published on this topic, but to recommend one of these for further study is difficult and will depend on your particular application.

To analyze reliability data, we need to use a suitable computer program. Many programs are available, and it is difficult to claim that one is better than all the others. In this book, we have chosen the program R because it covers most of the

System Reliability Theory: Models, Statistical Methods, and Applications, Third Edition.
Marvin Rausand, Anne Barros, and Arnljot Høyland.

Companion Website: www.wiley.com/go/SystemReliabilityTheory3e

techniques dealt with in this chapter, because it is used by many universities, and because it is free software that can be run on all major computer platforms.

By searching the Internet for "survival analysis," "survival models," and similar terms, you find an almost endless number of presentations, lecture notes, and slides. Most of these are made for medical applications, but they are in most cases still relevant for the content of this chapter.

14.1.1 Purpose of the Chapter

The purpose of this chapter is to give an *introduction* to reliability data analysis that can be understood based on the material presented in the previous chapters. We focus on explaining the basic concepts and how the various methods can be used and do not dig deeply into the theoretical problems. We do, however, give references to where interested readers may find more extensive information. We assume that the reader has installed the program R on an available computer and has become a bit familiar with how it is used. We illustrate how R may be used in the analyses and present simple R scripts, such that the reader may repeat the analyses for other datasets. More complete R scripts may be found on the `book companion site`.

14.2 Some Basic Concepts

Before discussing the various analysis methods, we need to introduce the main terminology to be used.

Population. A population is a set of similar items or events that are of interest for some question or experiment. The population may, for example be

- All the valves of the same type in a plant.
- All the mobile phones of a particular brand.
- All the brakes of the same type used in the railway rolling stock within a country.

Model. To study an aspect of a population, we define a random variable X that may give us information about this aspect. To be able to use statistical methods in our study, we establish a *probabilistic model*, M, related to the random variable X. The model may be parametric, nonparametric, or semiparametric. As a starting point, we assume that the model M is parametric with some parameter θ. The parameter is fixed but unknown, applies for the population, and is sometimes called a *population parameter*.

If X is a discrete variable, the model is formulated by a conditional probability mass function $\Pr(X = x \mid \theta)$, where θ is fixed, but unknown. If X is a continuous variable, the model is formulated by a probability density function $f(x \mid \theta)$.

Sample. To study the entire population is usually too time-consuming and expensive, and we therefore suffice by studying a sample from the population. A sample is a subset of a population, collected or selected by a defined *sampling procedure*. When the sampling is random, the sample is said to be a *random sample*. In reliability studies, the sample is not always random, and we have to suffice with the sample that is possible to get.

Experiment. To get information about the random variable X, we carry out independent and identical experiments of the n items in the sample. When the n experiments are completed, we have the *dataset* $x_1, x_2, \dots, x_n$. The joint distribution of obtaining this dataset is because of independence:

$$f(x_1, x_2, \dots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta), \tag{14.1}$$

for a continuous variable. The expression for a discrete variable is left to the reader.

Inference. Inference is a procedure to use information gathered from a sample to make statements about the population from which the sample was taken. The main concepts involved in statistical inference are shown in Figure 14.1.

14.2.1 Datasets

The starting point in this chapter is a *dataset* containing a *random sample* from a *population* of independent items. Throughout this chapter, we assume that the time-to-failure of an item is a nonnegative random variable T. In most applications, we assume that we observe n identical items with random times-to-failure

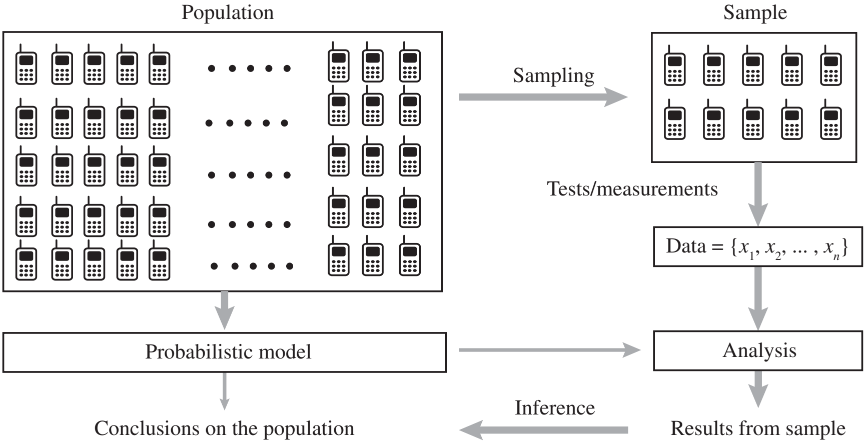


Figure 14.1 Main concepts of statistical inference.

$T_i, T_2, \dots, T_n$ that are independent and identically distributed with distribution function $F_{\mathrm{T}}(t)$ and probability density function $f_{\mathrm{T}}(t)$. The corresponding *observed* sample survival times are denoted $t_1, t_2, \dots, t_n$.

For many datasets, the observation of an item is stopped before the item has failed, and we say that the time-to-failure is *censored*. There are many reasons for censoring, including that the test equipment breaks down, the item is taken out of service due to operational causes, or that the allocated test or observation period is over. For each item, we assume that censoring occurs at time C that may be deterministic or random.

14.2.2 Survival Times

When censoring is present, we cannot always observe the true time-to-failure T. We only observe the *survival time*, the time until a failure or a censoring occurs, as shown in Figure 14.2.

We may assume that two independent processes are competing to terminate item i, a failure process and a censoring process. With no censoring, the time-to-failure T_i would be observed, and with no failure, the censoring time C_i would be observed. With both processes active, we observe the minimum of T_i and C_i, that is min $\{T_i, C_i\}$, for $i = 1, 2, \dots, n$.

We still denote the dataset $t_1, t_2, \dots, t_n$, but to each observation t_i, we associate an indicator δ_i, defined by

$$\delta_i = \begin{cases} 1 & \text{if } t_i \text{ ends with a failure (i.e. } T_i < C_i) \\ 0 & \text{if } t_i \text{ ends with censoring (i.e. } T_i > C_i) \end{cases} \quad \text{for} \quad i = 1, 2, \dots, n.$$

We call the indicator δ_i the *status* of survival time t_i. The dataset therefore consists of n duplets (t_i, δ_i), for $i = 1, 2, \dots, n$, telling how long time the item survived and whether the observation stopped with a failure (F) or a censoring (C).

In this chapter, we assume that survival time t_i is measured from when item i was new. In many practical applications, item i has a certain age $t_i^{(0)}$ when the observation starts. Here, we assume that $t_i^{(0)} = 0$ for all $i = 1, 2, \dots, n$, and we further assume that all survival times can be shifted to a common starting point, without loss of information. This is shown in Figure 14.3.

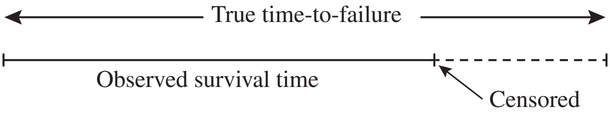


Figure 14.2 Time-to-failure and Observed Survival Time.

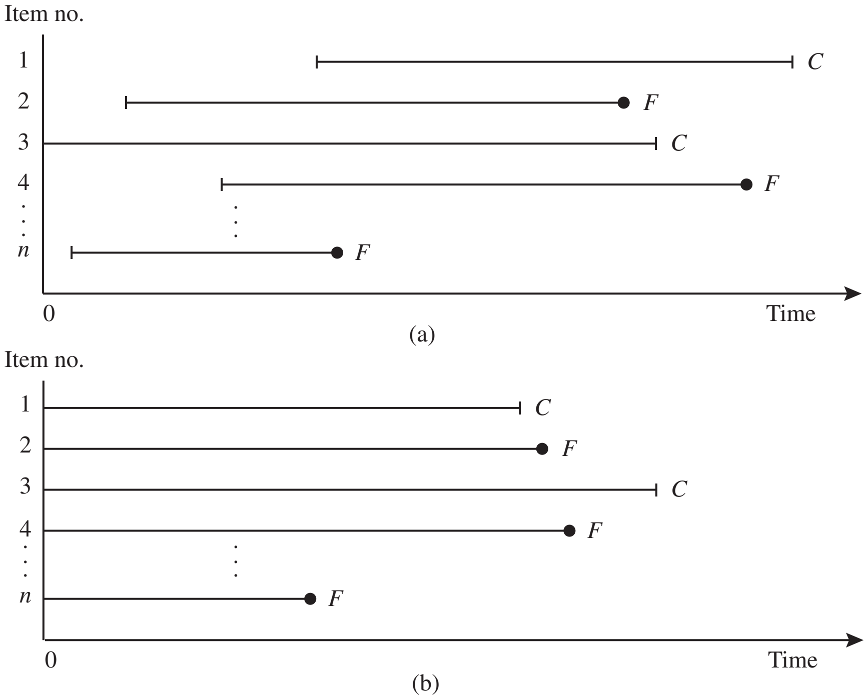


Figure 14.3 An observed dataset (a), and the same dataset shifted to time 0 (b). *F* denotes a failure and *C* denotes censoring.

Entering Survival Times into R

The dataset may be entered into R in several ways. The most common is as follows: (i) a spreadsheet file, (ii) comma-separated values (CSV) file, or (iii) manually as one or more vectors. For the last option, we enter the ordered survival times, for example

t_i	17.88	28.92	33.00	41.52	42.12	45.60
δ_i	1	0	1	1	1	0

We denote the vector of survival times `survtime` and the status vector `status` and enter the following in the R script

```
survtime <- c(17.88,28.92,33.00,41.52,42.12,45.60)
status <- c(1,0,1,1,1,0)
```

For a dataset with many survival times, it may be wise to enter the data into a spreadsheet program and save the file either as a CSV or as an Excel® file.[1]

The Survival R Package

The R package `survival` contains many of the survival analysis functions used in this chapter. How the package is loaded into your R session is illustrated by the following R script, which is based on the data in the R script above. Before loading the script, you must have installed the package `survival`, by the command `install.packages('survival')`. The new dataset is called `my.surv` and is prepared for further analysis by the function `Surv`.

```
library(survival)  # Activate the package survival
survtime <- c(17.88,28.92,33.00,41.52,42.12,45.60)
status <- c(1,0,1,1,1,0)
# Arrange and give the dataset a name
my.surv <- Surv(survtime,status)
# Display the dataset my.surv
print(my.surv)
```

Running this script in R gives the output

```
> print(my.surv)
[1] 17.88  28.92+ 33.00  41.52  42.12  45.60+
```

Observe that + is added to the censored survival times. The + indicates that the time-to-failure would have been somewhat longer if the survival time were not censored.

14.2.3 Categories of Censored Datasets

This section describes four main types of censoring and two subtypes.

Censoring of Type I

A life test of n numbered and identical items is carried to gain information about the probability distribution of the time-to-failure T of the items. A specific time interval $[0, \tau]$ has been allocated for the test. After the test, only the times-to-failure of those items that failed before τ are known.

[1] Commands to import data files into R may be found by searching the Internet for “import data into R.”

This type of censoring is called *censoring of type I*, and the information in the dataset consists of s ($\leq n$) observed, ordered survival times

$$t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(s)}.$$

In addition, we know that $(n - s)$ items have survived the time τ, and this information should also be used.

Because the number of items that fail before time τ obviously is random, there is a chance that none or relatively few of the items will fail before τ. This may be a weakness of the test design.

Censoring of Type II

Consider the same life test as for censoring of type I, but assume that it has been decided to continue the test until exactly r ($< n$) failures have occurred. The test is therefore stopped when the rth failure occurs. This censoring is called *censoring of type II*, and the dataset obtained from the test consists of

$$t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(r)},$$

together with the fact that $(n - r)$ items have survived the time $t_{(r)}$.

In this case, the number r of recorded failures is not random. The price for obtaining this is that the time $t_{(r)}$ to complete the test, is random. A weakness of this design is therefore that we cannot know beforehand how long time the test will last.

Censoring of Type III

Type III censoring is a combination of the first two types. The test terminates at the time that occurs first, τ or the rth failure (τ and r must both be fixed before the test starts).

Censoring of Type IV

Consider a life test of n numbered identical items. Each item may either run to failure or be censored at a random time C. The time-to-failure T is, as before, assumed to have distribution function $F_{\text{T}}(t)$ and probability density function $f_{\text{T}}(t)$, whereas the censoring time C has distribution function $F_{\text{C}}(c)$ and probability density function $f_{\text{C}}(c)$. The two random variables T and C are assumed to be independent. The survival time we observe is therefore the minimum of T and C.

This censoring is called *censoring of type IV* and is sometimes also called *random censoring*. Many of the datasets that are relevant for reliability studies have random censoring, especially when the datasets originate from systems in operation.

Right Censoring

Right censoring means that the item is removed from the study before a failure occurs, or that the study of the item ends before the item fails. For all the examples in this chapter, the censoring is right censoring.

Example 14.1 (Censoring caused by other failures)

Consider a plant where two independent items are located close to each other in a location that is difficult to access. When one of the items fails, both items are replaced or totally refurbished. In this situation, failure of one item leads to censoring of the other item. Because failures occur at random, this is an example of random censoring (i.e. of type IV).

The same censoring applies to a data collection where only a particular failure mode A is of interest. If another failure mode occurs, and is repaired, before failure mode A occurs, the time-to-failure of failure mode A is censored. In this case, we often say that we have *competing failure modes*. □

Informative Censoring

All the examples discussed in this chapter assume that the censoring is *noninformative*. This means that the time-to-failure T is independent of the censoring mechanism. The censoring may also be informative, for example when an item is taken out of service because its level of performance is less than adequate, but without failing.

14.2.4 Field Data Collection Exercises

In field data collection exercises, such as for the OREDA project, survival times are collected from a certain time window (t_1, t_2). We may, for example collect data for failure events that occurred between 1 January 2015 and 31 December 2019. At the beginning of the time window, at time t_1, the items may have different ages $t_i^{(0)}$, whereas some items may be installed during the time window, often as replacements for failed items. In many field data collection exercises, it is assumed that repair of a failed item brings it back to an as-good-as-new condition.

The resulting dataset is sometimes complicated and is best entered into a spreadsheet program, with the following columns:

Number	Item number (i)
Age	Age of item at the start of the data collection ($t_i^{(0)}$)
Start	Starting time of observation (t_i^{start})
Stop	Observation terminated (t_i^{stop})
Status	Status at stop (1 = failed, 0 = censored)

An example of such a dataset is shown in Figure 14.4.

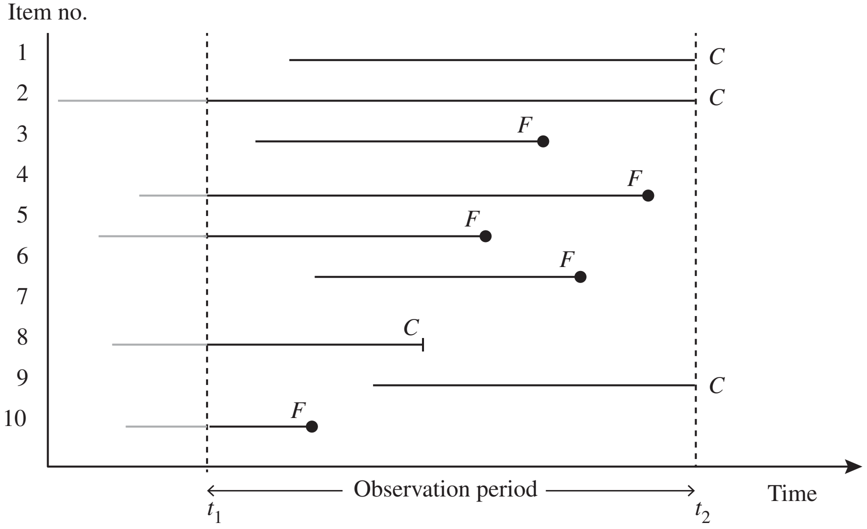


Figure 14.4 Typical dataset for field data.

14.2.5 At-Risk-Set

The at-risk-set at time t is the set of items that have not failed or been censored before time t, that is, the set of items that are *at risk* of failing at time t. When a (single) failure or censoring occurs, one item is removed from the at-risk-set and when a new item enters the study, the at-risk-set is increased by one item. The number of items in the at-risk-set at time t is an important variable in several of the survival analyses methods presented in this chapter.

14.3 Exploratory Data Analysis

An *exploratory data analysis* (EDA) is an essential first step in any data analysis. The EDA gives a “first look at the data” before any modeling effort is done. An EDA has two main parts: (i) calculation of a selection of sample statistics such as the mean, median, and standard deviation, and (ii) data visualization in the form of histograms, empirical distribution functions, Q–Q plots, and so on.

EDA helps the analyst to understand the underlying structure of the data, to identify anomalies and outliers in the dataset, to assess the assumptions about the data, and several more. The examination of the data helps seeing what the data can tell us. EDA got increased importance following the publication of John W. Tukey’s seminal book *Exploratory Data Analysis* (Tukey 1977).

14.3.1 A Complete Dataset

The starting point of an EDA is a specific dataset. This section assumes that we have a *complete* dataset $t_1, t_2, \dots, t_n$ where all survival times are times-to-failure. This means that the status is $\delta_i = 1$ for all items $i = 1, 2, \dots, n$, and that we do not need to enter the status into R. All the n entries in the dataset are assumed to be correct observations of a common variable. Many analytical methods require the dataset to be *sorted* in ascending order. A sorted dataset is also called an *ordered* dataset and is written as $t_{(1)}, t_{(2)}, \dots, t_{(n)}$, such that $t_{(1)} \leq t_{(2)}, \leq \cdots \leq t_{(n)}$.

As an illustration, we use the complete and ordered dataset of 22 observed values in Table 14.1. We call the dataset `survtime` and Table 14.1 shows the most direct way of entering the dataset into R by using the terminal.[2]

How the data is recorded in R is seen by launching the command `print (survtime)` in the terminal. The result is:

```
[1]    17.88  28.92   33.00   41.52   42.12   45.60  48.40  51.84
[9]    51.96  54.12   55.56   67.80   68.64   68.64  68.88  84.12
[17]   93.12  98.64  105.12  105.84  127.92  138.04
```

If the dataset `survtime` were entered into R as an unordered set, it may be ordered by the function `sort(survtime)`.

Remark 14.1 (An advise)
We will illustrate several methods by using the `survtime` dataset. If you want to test our examples or play with R, it may be wise to set up a textfile containing the data. The simplest way is to write one datapoint per line (with a "full stop" as decimal point and save the file as, for example, `dataset.txt` in your R working directory.[3] You may enter the textfile into R and activate the dataset `survtime` by the command `survtime<-read.table("dataset.txt",header=F, dec=".")`

Table 14.1 A complete and ordered dataset of survival times.

```
survtime <-     c(17.88,28.92,33,41.52,42.12,45.6,48.4,51.84,
                51.96,54.12,55.56,67.8,68.64,68.64,68.88,84.12,
                93.12,98.64,105.12,105.84,127.92,138.04)
```

[2] The terminal is called the Console in RStudio.

[3] After having created the working directory, you may check the path by the command `getwd()`.

Instead of a textfile, you may alternatively create an Excel file or a CSV file (but this requires another command to activate the data). □

Ties

Two or more continuously distributed survival times may sometimes be recorded with the same value. This is called a *tie* and may be caused by common-cause failures or by rounding-off. The dataset in Table 14.1 has a tie for 68.64, because two survival times are recorded with the same value. The number of failures that occur at time $t_{(i)}$ is called the *multiplicity* of the tie and is denoted by d_i. The dataset may therefore be recorded in two different ways:

(1) The ordered dataset may be recorded as $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$ with a survival time for each item, thus realizing that some of the survival times may be equal.
(2) The ordered dataset may be recorded as $n_t \leq n$ *distinct* survival times $t_{(1)} < t_{(2)} < \cdots < t_{(n_t)}$ associated with a vector giving the multiplicities of failures $d_1, d_2, \dots, d_{n_t}$.

Most of the following sections use option 1.

14.3.2 Sample Metrics

Valuable information about the dataset can be obtained by applying sample metrics to the dataset. This section defines and shows how to calculate a number of these metrics.

Mean

The mean of a dataset is a measure of the central location of the data values and is calculated as the sum of its data values divided by the number n of data values.

$$\bar{t} = \frac{1}{n} \sum_{i=1}^{n} t_i. \tag{14.2}$$

The R command to obtain the mean of the dataset is `mean(survtime)` and for the data in Table 14.1, we obtain $\bar{t} = 68.08$.

Median

The median t_m of a dataset is the value at the middle of the ordered data. For odd number of values (i.e. $n = 2k + 1$), the median is the $(k + 1)$th smallest in the ordered dataset, that is $t_{(k)}$. For an even number of values (i.e. $n = 2k$), the median is the average of the two values in the middle of the ordered dataset, that is the average of $t_{(k)}$ and $t_{(k+1)}$. If, for example the sorted dataset has the six values 2, 4, 5, 7, 8, 10, the median is $(5 + 7)/2 = 6$. A more formal definition is given in (14.3).

$$t_m = \begin{cases} t_{(k+1)} & \text{for } n = 2k + 1 \\ \dfrac{t_{(k)} + t_{(k+1)}}{2} & \text{for } n = 2k \end{cases}. \tag{14.3}$$

The ordered dataset in Table 14.1 has $n = 22$ values and the median is therefore the average of $t_{(11)}$ and $t_{(12)}$,

$$\text{Median} = \frac{t_{(11)} + t_{(12)}}{2} = 61.68.$$

The same result is obtained by the R function `median(survtime)`. Observe that the mean value is larger than the median for this dataset.

A simple summary, including the mean and the median, of the dataset `survtime` is obtained by the command `summary(survtime)`. If you have created the textfile `dataset.txt` as recommended in Remark 14.1, you may use the script

```
survtime <-read.table("dataset.txt",header=F,dec=".")
summary(survtime)
```

and obtain

```
Min.    : 17.88
1st Qu.: 46.30
Median : 61.68
Mean    : 68.08
3rd Qu.: 90.87
Max.    :138.04
```

Quartiles (Qu.) are introduced below.

Variance and Standard Deviation

The *variance* is a measure of how the data values are dispersed around the *mean* and is calculated as

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (t_i - \bar{t})^2. \tag{14.4}$$

The *standard deviation* is the square root of the variance

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (t_i - \bar{t})^2}. \tag{14.5}$$

The R commands to obtain the variance and the standard deviation of the dataset are `var(survtime)` and `sd(survtime)`, respectively. Observe that the standard deviation is measured with the same unit as the data values, whereas the variance is measured with "squared units." For the dataset in Table 14.1, the (sample) standard deviation obtained by `sd(survtime)` is 32.01.

Quantiles

For $p \in (0, 1)$, the *quantile of order p* of the distribution $F_{\mathrm{T}}(t)$ is the value t_p such that

$$F_{\mathrm{T}}(t_p) = p, \quad \text{which means that} \quad \Pr(T \leq t_p) = p.$$

For realistic life distributions, t_p is unique.

Now, consider an ordered dataset $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$. The *(sample) quantile of order p* may approximately be calculated as $t_{([np]+1)}$, where $[np]$ is the largest integer $< np$. The dataset in Table 14.1 has $n = 22$ values. To determine the (sample) quantile of, say order $p = 0.15$, we first calculate $np = 22 \cdot 0.15 = 3.3$. The largest integer less than np is 3 and the quantile of order 0.15 is therefore $t_{(4)} = 41.52$.

The (sample) quantile of order p is available in R by the function `quantile(survtime,p)`, which gives 41.61. This is not exactly the same result we got by hand calculation because R applies a more elaborate and “correct” formula based on interpolation of the ordered survival times. Interested readers may check the help file in R, `help(quantile)`.

Quartiles

The quantiles of order 0.25 and 0.75 are called the lower and upper *quartiles*, respectively. The *lower quartile* (or 1st quartile), $t_{0.25}$ is the value that cuts off the first 25% of the ordered dataset, and the *upper quartile* (or 3rd quartile) $t_{0.75}$ is the value that cuts off the first 75% of the ordered dataset. Both are provided by the command `summary(survtime)`.

Interquartile Range

The distance between the upper and the lower quartile, $t_{0.75} - t_{0.25}$, called the *interquartile range*, is a common measure for the dispersion of the dataset around its mean or median. The interquartile range for the dataset in Table 14.1 is determined by `quantile(survtime,0.75)-quantile(survtime,0.25)` and the result is 44.57.

Sample Moments and Central Moments

The kth (noncentral) *sample moment* for the dataset $t_1, t_2, \ldots, t_n$ is defined as

$$m_{k,nc} = \frac{1}{n} \sum_{i=1}^{n} t_i^k. \tag{14.6}$$

We observe that the first sample moment is $\bar{t} = \frac{1}{n} \sum_{i=1}^{n} t_i$ is the *average* value (i.e. mean) of the dataset.

The kth ($k \geq 2$) *central sample moment* is centered around its average value of the dataset and is defined as

$$m_{k,c} = \frac{1}{n} \sum_{i=1}^{n} (t_i - \bar{t}\,)^k. \tag{14.7}$$

Moments are available in the R package `moments` that must be installed in R before you can use it.

```
library(moments)
survtime <-read.table("dataset.txt",header=F,dec=".")
k <-3 # Choose the order of the noncentral moment
moment(survtime,order = k,central=F)
moment(survtime,order=k,central=T)
```

The result for the dataset in Table 14.1 is

Order (k)	Noncentral moment	Central moment
2	5612.752	978.3606
3	534022.7	18720.53

Skewness

Skewness is a measure of the asymmetry of the dataset. The skewness value can be positive or negative. When the distribution of the values of the dataset is symmetric, the skewness is zero. When the values are predominantly large (but with some small values), the skewness is negative and when the values are predominantly small (with some large values), the skewness is positive.

The skewness γ_1 is defined by

$$\gamma_1 = \frac{m_{3,c}}{m_{2,c}^{3/2}}, \tag{14.8}$$

where $m_{k,c}$ is the kth central sample moment of the dataset. The skewness γ_1 is available in the R package `moments` by the command `skewness(survtime)`. The result is 0.6117442, which indicates that the dataset is slightly skewed to the left.

Kurtosis

The *kurtosis* describes the shape of the tails of the distribution of the values in the dataset. The normal distribution has zero kurtosis. Negative kurtosis indicate a thin tail of the distribution and positive kurtosis indicate a thicker tail.

The kurtosis γ_2 is defined as

$$\gamma_2 = \frac{m_{4,c}}{m_{3,c}^2} - 3, \tag{14.9}$$

where μ_k is the kth central moment of the dataset. The kurtosis is available in the R package `moments` by the command `kurtosis(survtime)` and the result for the dataset in Table 14.1 is 2.555 003.

14.3.3 Histogram

A *histogram* consists of parallel bars that graphically show the frequency distribution of a variable. As a default, all the bars have the same width. We may choose the number of bars to display. We may also choose whether to (i) show the number of values in the dataset that fall into the interval corresponding to the width of the bar, or (ii) to show the relative number (or percentage) of the values that fall into the interval. With option (ii), the histogram is said to show the *relative frequency distribution* or the distribution density of the values in the dataset.

The information obtained from the histogram depends on the resolution, that is how many intervals we choose. Figure 14.5 shows three different histograms of the data in Table 14.1, with different numbers of columns.

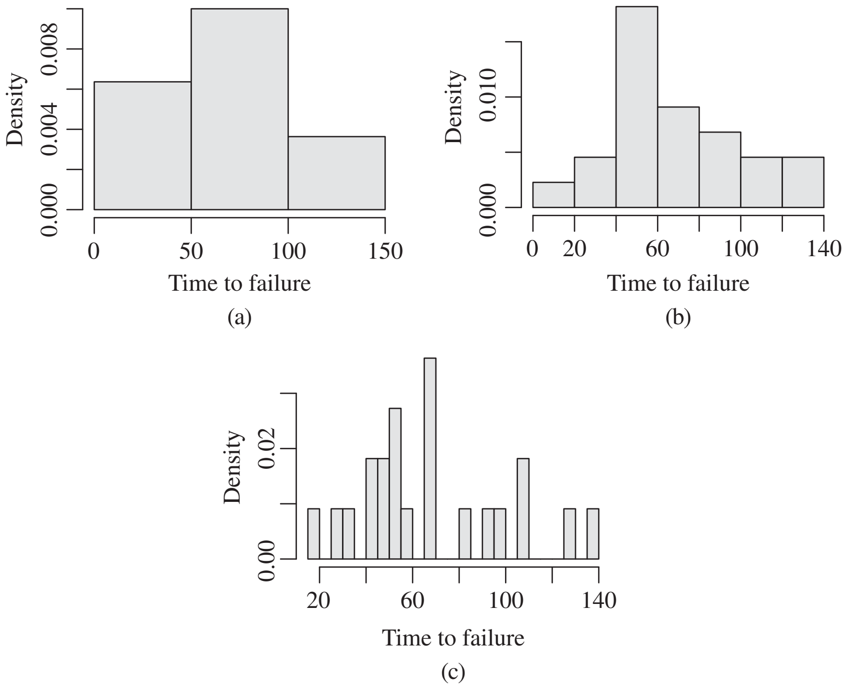


Figure 14.5 Histogram of the dataset in Table 14.1 with different numbers of columns: (a) 3 columns, (b) 7 columns, and (c) 26 columns.

It is not always true that a higher resolution makes it easier to understand the distribution of the data.

Histograms are established by the R script

```
survtime <-read.table("dataset.txt",header=F, dec=".")
hist(survtime$V1,breaks=3,freq=F)   # Plots
the histogram
```

This script provides a histogram according to option (i). A relative frequency histogram [option (ii)] is obtained by replacing `freq=F` with `freq=T`, where F is an abbreviation for false and T is an abbreviation for true.

The reader is encouraged to run the script with different values for `breaks`.

14.3.4 Density Plot

The distribution of the dataset can also be illustrated by a sample density plot, by using the R script:

```
survtime <-read.table("dataset.txt",header=F, dec=".")
d <- density(survtime)   # Returns the density data
plot(d)   # Plots the results
```

The resulting plot is shown in Figure 14.6. The plot is made by an averaging technique and is based on a set of input parameters. The current plot is made with the default parameters of the `density` command. Other parameters and other

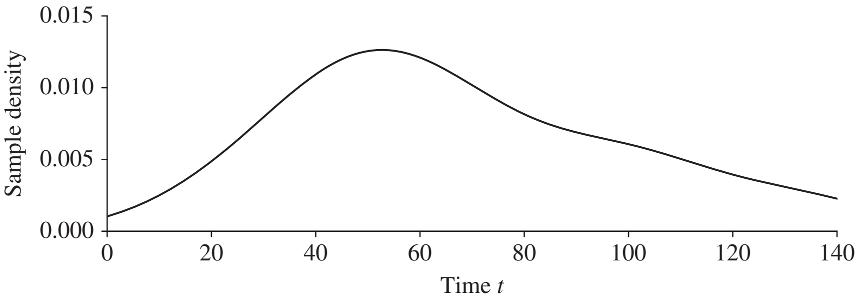


Figure 14.6 A sample density plot of the dataset in Table 14.1.

averaging techniques may be chosen. Interested readers may consult the help file, by the command `help(density)` in the R terminal (console).

14.3.5 Empirical Survivor Function

The survivor function $R(t) = \Pr(T > t)$ is the probability that an item from the population will still be functioning at time t. When a complete dataset is available, the survivor function may be estimated by the empirical survivor function $R_n(t)$

$$R_n(t) = \frac{\text{Number of items with survival time} > t}{n}. \tag{14.10}$$

$R_n(t)$ is from (14.10) seen to be the relative frequency of items that survive time t and is therefore an obvious estimate for $R(t)$.

For a censored dataset, the estimate $R_n(t)$ changes only at the failure times $t_{(i)}$. Between two failure times, such as in an interval $t_{(i)} \le t < t_{(i+1)}$, the number of failures does not change, and $R_n(t)$ remains constant. Observe that $R_n(t)$ is reduced by $1/n$ each time a failure occurs. If more than one failure occurs at the same failure time t and the tie has multiplicity d, $R_n(t)$ is reduced by d/n.

Consider a sample of n items from a population and let $N(t)$ be the number of these items that survive time t. We may consider this as a binomial experiment with n independent trials and probability $R(t)$ of survival, and write the estimator for $R(t)$ as[4]

$$\widehat{R}(t) = \frac{N(t)}{n}. \tag{14.11}$$

The random variable $N(t)$ has a binomial distribution with probability mass function

$$\Pr(N(t) = m) = \binom{n}{m} R(t)^m [1 - R(t)]^{n-m} \quad \text{for } m = 0, 1, \dots, n,$$

with mean and variance

$$E[N(t)] = nR(t).$$
$$\text{var}[N(t)] = nR(t)[1 - R(t)].$$

The mean of the estimator is $E[\widehat{R}(t)] = nR(t)/n = R(t)$ and the estimator is therefore unbiased. The variance of the estimator is

$$\text{var}[\widehat{R}(t)] = \frac{\text{var}[N(t)]}{n^2} = \frac{R(t)[1 - R(t)]}{n} \xrightarrow[n\to\infty]{} 0.$$

The plot of $R_n(t)$, as shown in Figure 14.7, for the dataset in Table 14.1 is also called a *survival curve* and can be made with R using several different packages.

[4] Estimators are discussed in Section 14.4.

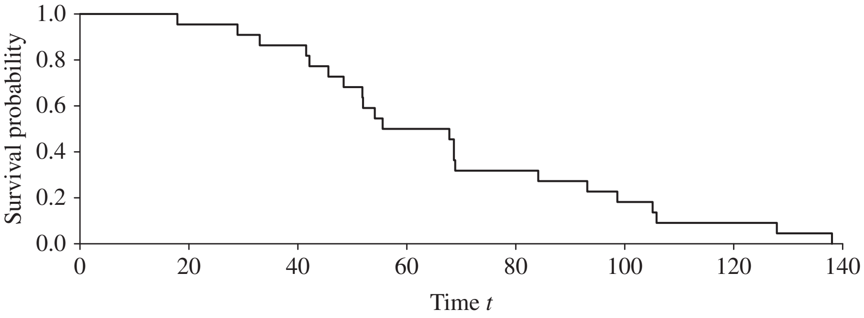


Figure 14.7 Empirical survivor function (survival curve) for the dataset in Table 14.1.

The authors prefer the package `survival` and the survival curve is obtained by the script:

```
library(survival)
survtime <-read.table("dataset.txt",header=F,dec=".")
# Prepare the data and calculate required values
data<- Surv(survtime)
survfunct<- survfit(Surv(survtime)~1,conf.type="none")
plot(survfunct, xlab="Time t", ylab="Survival
probability")
```

A 95% pointwise confidence interval is obtained by replacing `conf.type='none'` with `conf.type='plain'` in the script above. The plot obtained is shown in Figure 14.8.

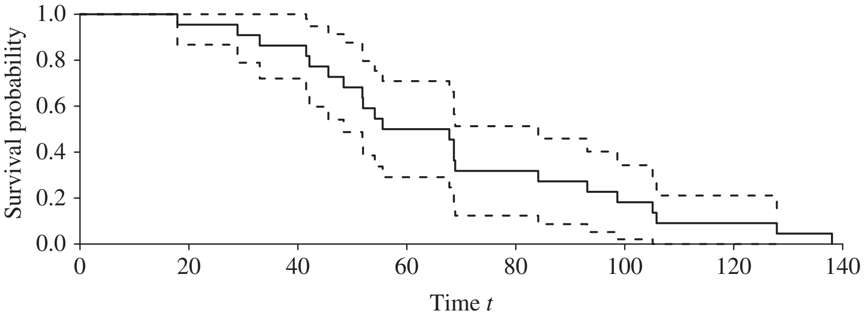


Figure 14.8 Empirical survivor function (survival curve) for the dataset in Table 14.1 with 95% confidence intervals.

14.3.6 Q–Q Plot

A Q–Q plot compares the quantiles of the dataset with the quantiles of a specified probability distribution $F(t)$. The plot is constructed by plotting the kth smallest observation out of n against the expected value of the kth smallest observation out of n from a random sample from $F(t)$. To construct such a plot by hand calculation is time-consuming, and we need a computer program. A Q–Q plot for the normal distribution $\mathcal{N}(0, 1)$ is available in R by the function qqnorm.

If the observations are approximately normally distributed, a normal Q–Q plot of the observations results in an approximately straight line. The R script to produce the Q–Q plot for the dataset survtime in Table 14.1 is

```
survtime<-read.table("dataset.txt",header=F,dec=".")
x<-survtime$V1
qqnorm(x)
qqline(x)
```

The resulting plot for the dataset in Table 14.1 is shown in Figure 14.9. The Q–Q plot shows a fairly good fit to the normal distribution for this particular dataset. If we consider the fit to the normal distribution to be acceptable, the parameters of the normal distribution can be estimated from the slope and intercept of the straight line. This is not pursued any further here.[5]

Q–Q plots for general distributions may be obtained in R by using the function qqplot. To use this function, we need to compare our dataset with a simulated dataset from the distribution we want to compare our data with. Assume that we want to compare the data in Table 14.1 with the exponential distribution with rate $\lambda = 1$. A random sample of size, say, 300 from the exponential distribution is generated in R by the function rexp(300,rate=1). The Q–Q plot comparing the data in Table 14.1 with the exponential distribution is obtained by the script

```
survtime <-read.table("dataset.txt",header=F,dec=".")
y<- survtime$V1
qqplot(rexp(300,rate=1),y)
```

The Q–Q plot produced by this script is shown in Figure 14.10.

Because the data from the exponential distribution are simulated, you do not get exactly the same figure when re-running the script. The exponential Q–Q plot in Figure 14.9 is rather far from a straight line, and we may therefore conclude that the data probably do not come from the exponential distribution.

[5] See, for example, https://en.wikipedia.org/wiki/Q-Q_plot.

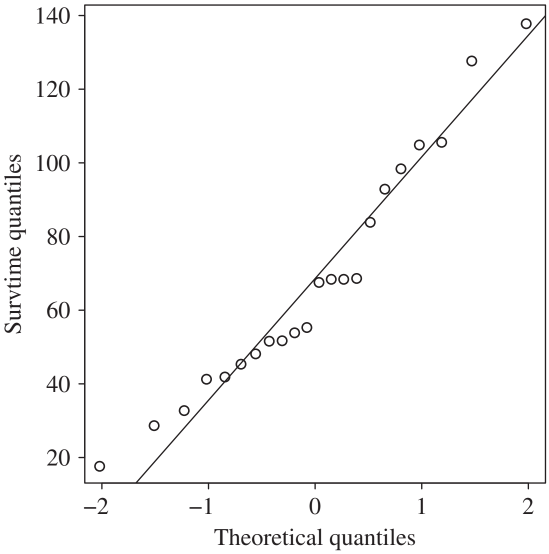


Figure 14.9 Normal Q–Q plot for the dataset in Table 14.1, made with the R function `qqnorm`.

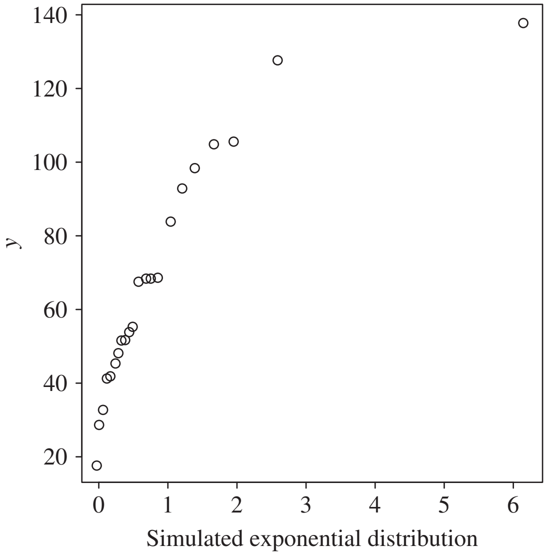


Figure 14.10 Exponential Q–Q plot for the dataset in Table 14.1, made with the R function `qqplot`.

14.4 Parameter Estimation

Probability distributions usually have one or more quantities that we call *parameters*. Examples of parameters include λ in the exponential distribution $\exp(\lambda)$ and the mean μ and the standard deviation σ in the normal distribution $\mathcal{N}(\mu, \sigma^2)$.

Parameters are generally – at least partly – unknown and cannot be measured directly.

This chapter deals with a population of similar items, and we establish a probabilistic model for a typical item in the population. The parameters of this model are therefore called *population parameters*. To get information about population parameters, we take a random sample of a certain number (n) of elements from the population and measure some properties of each element. We then obtain a dataset $\{t_1, t_2, \dots, t_n\}$. Each measurement may be a scalar or a vector of values. This process is shown in Figure 14.1.

Parameter estimation is the process of obtaining information about the parameter(s), based on the dataset. As part of this process, we have to answer questions such as (i) Which measurable properties of the sample elements shall be measured? (ii) How shall we combine these measurements to provide information about the population parameters? (iii) How accurate is this information? This section will shed some light on the parameter estimation process, but first, we need some terminology.

14.4.1 Estimators and Estimates

An *estimator* of a parameter θ is a statistic (i.e. a random variable) that is often denoted $\widehat{\theta}$. An estimator $\widehat{\theta}$ may be considered a metric where observed data can be input to calculate an *estimate* (i.e. a numeric value) for θ. This estimator is sometimes called a *point* estimator and the corresponding estimate is called a point estimate for θ.

We may also talk about *interval* estimators and interval estimates for θ. For interval estimators a probability, often called a *confidence level*, is specified as the probability that the interval contains the "true" value of the parameter. The interval is also called a *confidence interval* for θ.

14.4.2 Properties of Estimators

An estimator $\widehat{\theta}$ may be judged by the following features:

Unbiased

An estimator $\widehat{\theta}$ is said to be an *unbiased* (point) estimator for θ if its expected value is equal to the parameter, that is, if $E(\widehat{\theta}) = \theta$. An unbiased estimator will not systematically overestimate or underestimate the "true" parameter.

An estimator that is not unbiased is said to be *biased*. The bias is calculated as $b_n(\widehat{\theta}) = E(\widehat{\theta}) - \theta$.

An estimator $\widehat{\theta}$ is said to be *asymptotically unbiased* if $\lim_{n\to\infty} b_n(\widehat{\theta}) = 0$.

Small Variance

The estimator $\widehat{\theta}$ should preferably have a small spread or variability, that is, a small variance and standard deviation.

Mean Squared Error

The *mean squared error* (MSE) of the estimator $\widehat{\theta}$ for the parameter θ is defined as

$$\text{MSE}(\widehat{\boldsymbol{\theta}}) = E(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2 = [b_n(\widehat{\boldsymbol{\theta}})]^2 + \text{var}(\widehat{\boldsymbol{\theta}}). \tag{14.12}$$

The estimator $\widehat{\boldsymbol{\theta}}$ is said to be *efficient* if it has the smallest MSE among all competing estimators.

Consistency

An estimator $\widehat{\theta}$ is said to be a *consistent* (point) estimator for θ if $\widehat{\theta} \to \theta$ when the sample size n increases. More formally, we say that the estimator $\widehat{\theta}$ is consistent if we for all $\varepsilon > 0$ have that

$$\Pr(|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}| > \varepsilon) \to 0 \quad \text{when } n \to \infty. \tag{14.13}$$

This means that the distribution of $\widehat{\theta}$ becomes more and more concentrated around the "true" value of θ as the sample size increases.

Chebyshev's Inequality

Chebyshev[6] showed that for all $\varepsilon > 0$

$$\Pr(|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}| \geq \varepsilon) \leq \frac{E(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2}{\varepsilon^2} = \frac{\text{MSE}(\widehat{\boldsymbol{\theta}})}{\varepsilon^2}. \tag{14.14}$$

If we can prove that the MSE of $\widehat{\boldsymbol{\theta}}$ tends to 0 when $n \to \infty$, then $\widehat{\boldsymbol{\theta}}$ is consistent.

Estimator properties are illustrated in the following example.

Example 14.2 (Binomial model)

Consider a sequence of n independent and identically distributed Bernoulli trials with probability p for a specific outcome A. Let X be the number of trials that result in the outcome A. The random variable X is then binomially distributed, binom(n, p)

$$\Pr(X = x \mid p) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, \dots, n.$$

The mean and variance of X is

$$E(X) = np.$$
$$\text{var}(X) = np(1-p).$$

[6] Named after the Russian mathematician Pafnuty Lvovich Chebyshev (1821–1894).

It may be natural to estimate p as the relative frequency of the outcomes that result in A, and a natural estimator is therefore

$$\widehat{p} = \frac{X}{n}. \tag{14.15}$$

This estimator is seen to be unbiased, because

$$E(\widehat{p}) = \frac{E(X)}{n} = p.$$

The estimator $\widehat{p}$ is consistent because it is unbiased and

$$\text{var}(\widehat{p}) = \frac{\text{var}(X)}{n^2} = \frac{np(1-p)}{n^2} \to 0 \quad \text{when } n \to \infty.$$

If we, for example, carry out $n = 50$ independent Bernoulli trials and get $x = 3$ outcomes A, we may put this dataset into the estimator and obtain the point *estimate* $\widehat{p} = 3/50 = 0.06$. Again, observe that the estimator $\widehat{p}$ is a random variable, whereas the estimate is a numerical value. □

Remark 14.2 (Confusing symbols)
Observe that it may be confusing to use the same symbol (here $\widehat{p}$) for both the estimator and the estimate. The same confusion is found in almost all relevant textbooks and papers. □

To find adequate parameter estimators, we may use some general approaches or methods. In this book, we suffice by describing three popular methods for point estimation:

(1) Method of moments estimation (MME)
(2) Maximum likelihood estimation (MLE)
(3) Bayesian estimation, which is treated in Chapter 15.

14.4.3 Method of Moments Estimation

Consider a random variable T. The first *population moment* of T is the same as the mean value $E(T)$, and the kth (noncentral) population moment is $E(T^k)$ (if this mean value exists).

MME is based on the assumption that the sample moments are good estimates of the corresponding population moments. Assume that we have a sample $T_1, T_2, \dots, T_n$ from a distribution $F(t \mid \boldsymbol{\theta})$, where the parameter vector is $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$. The procedure to determine the MME of the parameters has three steps.

(1) Find the k first noncentral population moments $\mu_{1,nc}, \mu_{2,nc}, \dots, \mu_{k,nc}$. Each moment will contain one or more of the parameters $\theta_1, \theta_2, \dots, \theta_k$.

(2) Find the k first noncentral sample moments $m_{1,nc}, m_{2,nc}, \dots, m_{k,nc}$.

(3) From the system of equations $\mu_{i,nc} = m_{i,nc}$, for $i = 1, 2, \dots, k$, solve for the parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$. The solution is the MME $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \widehat{\theta}_2, \dots, \widehat{\theta}_k)$.

Recall that we – from the law of large numbers – know that the first sample moment converges to the first population moment (i.e. the population mean).

$$m_{1,nc} \to \mu_{1,nc} \quad \text{that is} \quad \frac{1}{n}\sum_{i=1}^{n} T_i \to E(T) \quad \text{for } n \to \infty, \tag{14.16}$$

but we do not know much about the higher moments (i.e. for $k \geq 2$).

We illustrate the MME procedure by two examples.

Example 14.3 (Exponential distribution)
We observe the times-to-failure of n similar items. The times-to-failure are denoted by $T_1, T_2, \dots, T_n$, and we assume that they are independent and identically distributed with constant failure rate λ, such that $T_i \sim \exp(\lambda)$, for $i = 1, 2, \dots, n$. In this case, we have only one unknown parameter to estimate, and we can suffice with considering only the first (population) moment $E(T) = 1/\lambda$. The first sample moment is given by the metric $\overline{T} = \frac{1}{n}\sum_{i=1}^{n} T_i$. The method of moment estimator for the parameter λ is therefore determined from $E(T) = \overline{T}$,

$$\frac{1}{\lambda} = \frac{1}{n}\sum_{i=1}^{n} T_i.$$

Solving for λ, we obtain the MME

$$\widehat{\lambda} = \frac{n}{\sum_{i=1}^{n} T_i}.$$

Assume that we have a complete dataset with $n = 8$ items that have run to failure and with a total time in operation $\sum_{i=1}^{8} t_i = 25\ 800$ hours. The MME (estimate) of the failure rate λ with this dataset is then

$$\widehat{\lambda} = \frac{8}{25\ 800}\ \text{h}^{-1} \approx 3.10 \times 10^{-4}\ \text{h}^{-1}.$$

□

Example 14.4 (Gamma distribution)
Let $T_1, T_2, \dots, T_n$ be a random sample of n independent gamma distributed random variables, such that $T_i \sim \text{gamma}(\alpha, \lambda)$, for $i = 1, 2, \dots, n$. The first population moment (i.e. the mean) is from Chapter 5, $\mu_1 = E(T_i) = \alpha/\lambda$. The variance of T_i is

$$\text{var}(T_i) = E(T_i^2) - [E(T_i)]^2 = \frac{\alpha}{\lambda^2}.$$

The second population moment is therefore

$$\mu_2 = E(T_i^2) = \text{var}(T_i) + [E(T_i)]^2 = \frac{\alpha}{\lambda^2} + \left(\frac{\alpha}{\lambda}\right)^2 = \frac{\alpha(\alpha+1)}{\lambda^2}.$$

Setting the first two population moments equal to the first two sample moments yields

$$\frac{\alpha}{\lambda} = \frac{1}{n}\sum_{i=1}^{n} T_i.$$

$$\frac{\alpha(\alpha+1)}{\lambda^2} = \frac{1}{n}\sum_{i=1}^{n} T_i^2.$$

We may now solve the two equations to obtain

$$\widehat{\lambda} = \frac{\frac{1}{n}\sum_{i=1}^{n} T_i}{\frac{1}{n}\sum_{i=1}^{n} T_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} T_i\right)^2}$$

and

$$\widehat{\alpha} = \widehat{\lambda}\,\frac{1}{n}\sum_{i=1}^{n} T_i = \frac{\left(\frac{1}{n}\sum_{i=1}^{n} T_i\right)^2}{\frac{1}{n}\sum_{i=1}^{n} T_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} T_i\right)^2}.$$

By using the dataset in Table 14.1, we may use the following R script to find the estimates of α and λ.

```
survtime <-read.table("dataset.txt",header=F,dec=".")
a<-mean(survtime)
b<-mean(survtime^2)
lambda<- a/(b-a^2)
print(lambda)
alpha<- lambda*a
print(alpha)
```

This gives the estimates $\widehat{\alpha} \approx 4.737$ and $\widehat{\lambda} \approx 0.0696$. □

General Properties of the MME

MMEs have a number of positive and negative properties. We suffice by listing some of these properties, without any proofs:

(1) The MMEs are easy to compute and will always work. The method provides estimators when other methods fail to do so or when estimators are hard to obtain.
(2) The MMEs are consistent.
(3) The MMEs may not be unique.
(4) MMEs are usually not the “best estimators” (i.e. most efficient).

(5) The minimum number of moments we need equals the number of unknown parameters.
(6) Sometimes, the MMEs may be meaningless.

14.4.4 Maximum Likelihood Estimation

The method of maximum likelihood was first introduced in 1922 by the British statistician and geneticist Ronald Aylmer Fischer (1890–1962) and has since been a commonly used method for estimating parameters. With this method, the parameters are estimated by the values that maximize the likelihood function. Before going into further detail, we need to introduce the likelihood function.

Likelihood Function

We start with the likelihood function for a discrete, binomial model. This model is based on a random variable X with probability mass function

$$\Pr(X = x \mid p) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, 2, \dots, n. \tag{14.17}$$

In the classical setup, the parameter p has a deterministic but an unknown value. In (14.17), the unknown parameter p is made visible in the probability mass function $\Pr(X = x \mid p)$ to highlight that the probability is also a function of p. The number n of trials is considered to be a known number and therefore not a parameter.

Assume that the experiment has been carried out and the data has been recorded. The data may, for example be $n = 10$ and $x = 3$. We may now wonder which value of p that produced this particular result. To shed light on this problem, we calculate the probability of obtaining $X = 3$ for different values of p. The probabilities may be calculated by using the function `dbinom(3, size=10, prob=p)` in R.

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7
$\Pr(X = 3 \mid p)$	0.0574	0.201	0.267	0.215	0.117	0.0425	0.009

The probabilities for $p = 0.8$ and $p = 0.9$ are very small and not included in this table. Observe that with these p-values, the probability $\Pr(X = x \mid p)$ is largest for $p = 0.3$, which means that $p = 0.3$ is the most *likely* probability to have produced $X = 3$.

Consider the probability $\Pr(X = 3 \mid p)$ as function of p.

$$L(p \mid 3) = \binom{10}{3} p^3 (1-p)^7 \quad \text{for } 0 \le p \le 1. \tag{14.18}$$

We use the symbol $L(p \mid 3)$, because it seems natural to call this function the *likelihood function* of p for the observed data. It tells how likely it is that a particular value of p has produced the observed result.

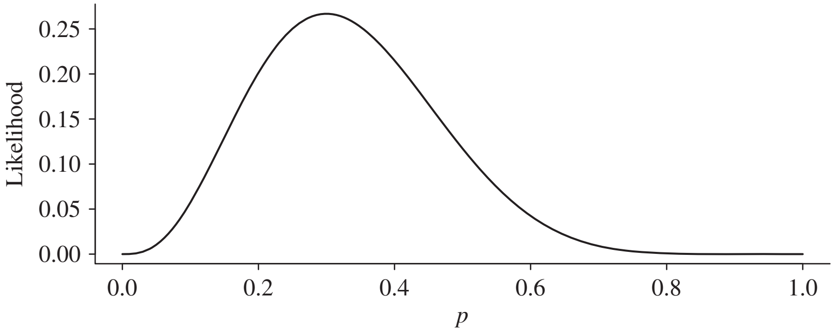


Figure 14.11 Likelihood function for the binomial distribution ($n = 10$ and $x = 3$).

The likelihood function for the observed values $n = 10$ and $x = 3$ is shown in Figure 14.11 as a function of p, and we observe that the most likely p-value to have produced $x = 3$ is $p = 0.3$.

Remark 14.3 (The likelihood function is not a probability distribution) We should observe that $L(p \mid 3)$ is *not* a probability distribution for p, because

$$\int_0^1 L(p \mid 3)\, dp = \binom{10}{3} \int_0^1 p^3(1-p)^7\, dp = \binom{10}{3} B(11,4) = 0.03 \neq 1,$$

where $B(a,b)$ is the beta function that can be written as

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

The beta function $B(a,b)$ is available in R by the function `beta(a,b)`. A factorial, such as 7!, is calculated in R by the function `factorial(7)`. □

Maximum Likelihood Estimate

As indicated above, the parameter value that maximizes the likelihood function for some observed data should be a good estimate for that parameter. This value is called the *maximum likelihood estimate* of the parameter.

To provide a general definition of the maximum likelihood estimate, we have to start with a model for the observed data, $f(\text{data} \mid \theta)$. This model can be a probability density function or a probability mass function depending on whether the model is continuous or discrete. The parameter θ may be one-dimensional or a vector of parameters. On this background, the maximum likelihood estimate is defined as follows.

Definition 14.1 (Maximum likelihood estimate, MLE)
The MLE, $\widehat{\theta}$, is the value of the parameter θ that maximizes the likelihood function with respect to θ. That is,

$$L(\widehat{\theta} \mid \text{data}) = \max_{\theta} L(\theta \mid \text{data}),$$

where the maximum is taken over all possible values of the parameter θ. □

More formally, the MLE $\widehat{\theta}$ may be written as

$$\widehat{\theta} = \arg\max_{\theta} L(\theta \mid \text{data}), \tag{14.19}$$

which means that $\widehat{\theta}$ is the value (the argument) θ that maximizes $L(\theta \mid \text{data})$. The MLE hence is the answer to the question: What value of the parameter θ makes the data most likely to occur?

In many applications, the natural logarithm of the likelihood function, is more convenient to work with. Because the logarithm $\log(\cdot)$ is a monotonically increasing function, the logarithm of $L(\theta \mid \text{data})$ attains its maximum value at the same point as the $L(\theta \mid \text{data})$ and therefore the *log-likelihood function* can be used instead of the likelihood function to obtain the MLE.

The log-likelihood function is written as

$$\ell(\theta \mid \text{data}) = \log L(\theta \mid \text{data}). \tag{14.20}$$

When plotting the log-likelihood function $\ell(\theta \mid \text{data})$ it is most common to plot the *negative log-likelihood function*, $-\ell(\theta \mid \text{data})$, such that the MLE $\widehat{\theta}$ is determined from the minimum value of this function. The negative log-likelihood function for the binomial distribution (14.18) is shown in Figure 14.12.

We now illustrate the maximum likelihood principle by some simple examples.

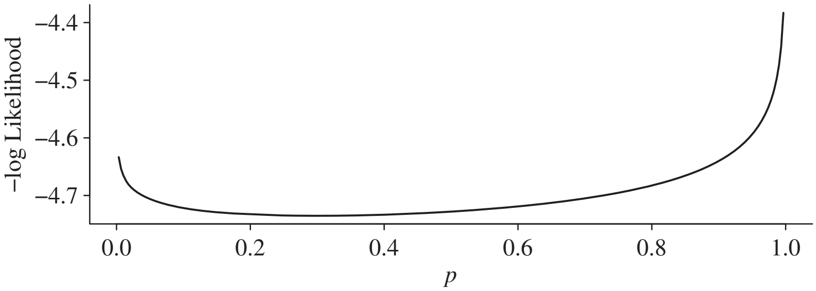


Figure 14.12 The negative log-likelihood function for the binomial distribution.

Example 14.5 (Binomial distribution)

Let $X \sim \text{binom}(n, p)$. The probability mass function is given by (14.18) and the likelihood function is

$$L(p \mid x, n) = \binom{n}{x} p^x (1-p)^{n-x},$$

and the log-likelihood function is

$$\ell(p \mid x, n) = \log \binom{n}{x} + x \log n + (n-x) \log(1-p).$$

The MLE is found by taking the derivative of $\ell(p \mid x, n)$ and setting this derivative equal to zero.

$$\frac{d}{dp}\ell(p \mid x, n) = \frac{x}{p} - \frac{n-x}{1-p} = 0.$$

An extreme point is found for $p = x/n$. We should then check that this extreme point really is a maximum. The ML estimate for the parameter p is therefore

$$\widehat{p} = \frac{x}{n}.$$

The above calculation may be done before any experiment is carried out, and it will apply for any possible values of X and n. We may therefore establish the *metric* for finding the maximum likelihood estimate as

$$\widehat{p} = \frac{X}{n}.$$

Observe the difference between the estimate and the estimator. The estimate is a number that is determined by the observed data and is a numerical *estimate* that is specific for the data. The estimator is a random variable that gives a metric for determining the estimate when the data becomes available. This random variable is called a maximum likelihood *estimator* and, unfortunately, the same symbol and the same abbreviation, MLE, is commonly used for both the estimator and the estimate.

Assume now that the experiment is carried out and that we have observed $x = 5$ in a total of $n = 40$ independent Bernoulli trials. With this data, the ML estimate of p is therefore

$$\widehat{p} = \frac{x}{n} = \frac{5}{40} = 0.125.$$

□

Example 14.6 (Homogeneous Poisson Process)

A homogeneous Poisson process (HPP) with unknown rate λ is observed during a time period $(0, \tau)$. Let $N(\tau)$ be the number of observed events. The probability mass function is

$$\Pr(N(\tau) = n) = \frac{(\lambda\tau)^n}{n!} e^{-\lambda\tau} \quad \text{for } n = 0, 1, 2, \ldots.$$

Assume that we have observed $n = 8$ events during a time period of length $\tau = 10\,560$ hours. The likelihood function is

$$L(\lambda \mid n, \tau) = \frac{(\lambda\tau)^n}{n!} e^{-\lambda\tau} \quad \text{for } \lambda > 0.$$

The log-likelihood function is

$$\ell(\lambda \mid n, \tau) = n \log(\lambda\tau) - \log n! - \lambda\tau.$$

The MLE is found by taking the derivative of $\ell(\lambda \mid n, \tau)$ and setting this derivative equal to zero.

$$\frac{d}{d\lambda}\ell(\lambda \mid n, \tau) = \frac{n\tau}{\lambda\tau} - \tau = 0.$$

The extreme (i.e. maximum) point is found for $\lambda = n/\tau$. As always, we should check that this is really a maximum. With the given data, the MLE (estimate) of λ is

$$\widehat{\lambda} = \frac{n}{\tau} = \frac{8}{10\,560 \text{ h}} \approx 7.58 \times 10^{-4} \text{ h}^{-1}.$$

□

Example 14.7 (Exponential distribution)

Let $T_1, T_2, \dots, T_n$ be n independent and identically distributed random variables with distribution $\exp(\lambda)$. Because the variables are independent and identically distributed, the joint probability density is

$$f(t_1, t_2, \dots, t_n \mid \lambda) = \prod_{i=1}^{n} f(t_i \mid \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda t_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} t_i} \quad \text{for } t \geq 0.$$

Assume that we have observed $n = 5$ variables during an accumulated time period $\tau = \sum_{i=1}^{5} t_i = 15\,600$ hours. The likelihood function is

$$L(\lambda \mid n, \tau) = \lambda^n e^{-\lambda\tau} \quad \text{for } \lambda > 0.$$

The log-likelihood function is

$$\ell(\lambda \mid n, \tau) = n \log \lambda - \lambda\tau.$$

The MLE is found by taking the derivative of $\ell(\lambda \mid n, \tau)$ and setting this derivative equal to zero.

$$\frac{d}{d\lambda}\ell(\lambda \mid n, \tau) = \frac{n}{\lambda} - \tau = 0.$$

The extreme (i.e. maximum) point is found for $\lambda = n/\tau$. Again, we should check that this is really a maximum. With the given data, the MLE (estimate) of λ is

$$\widehat{\lambda} = \frac{n}{\tau} = \frac{5}{15\,600 \text{ h}} \approx 3.2 \times 10^{-4} \text{ h}^{-1}.$$

□

Remark 14.4 (Factors not depending on the parameter can be deleted) As seen from the above examples, the likelihood function can usually be written as a product of two functions, such that, $L(\theta \mid x) = h(x)g(\theta, x)$. The log-likelihood function is then $\ell(\theta \mid x) = \log h(x) + \log g(\theta, x)$. When taking the derivative of $\ell(\theta, x)$ with respect to the parameter θ, we get $d \log h(x)/d\theta = 0$. We may therefore remove additive terms not containing unknown parameters from the log-likelihood function. For the binomial distribution in Example 14.5, the likelihood function is a product of $h(x) = \binom{n}{x}$ and $g(p, x) = p^x(1-p)^{n-x}$. The likelihood function may be simplified to $L(p, x, n) \propto p^x(1-p)^{n-x}$. □

General Properties of the MLE

The MLE has a high number of valuable properties. Here, we suffice by listing some of these properties without proofs.

- Assume that we have found the MLE $\hat{\theta}$ of θ and that $g(\theta)$ is a one-to-one function. The MLE of $g(\theta)$ is then $g(\hat{\theta})$.
- An MLE is asymptotically unbiased. $E(\hat{\theta}_n) \to \theta$ when the sample size n increases.
- Under relatively mild conditions, the MLE is consistent.
- Under certain regularity conditions, the ML estimator has an asymptotically normal distribution.

Interested readers may consult almost any good book on estimation theory to find proofs and further properties.

MLE with R

In most cases, ML estimation results in explicit formulas and R may therefore not be needed to compute the MLEs. If a computer support is deemed to be required, MLE is available by using the R packages: `stats4` , `bbmle`, or `maxLik`. If you want to use one of these packages, please read carefully the package manuals that are found on the Internet (e.g. by searching for "CRAN package bbmle").

To illustrate the analysis, a brief R script using the package `bbmle` to calculate the MLE for p in the binomial distribution is shown. To calculate the maximum likelihood estimate, `bbmle` uses the function `mle2`, which again is based on the negative log-likelihood function.

ML estimation in the binomial model was illustrated in Example 14.5. With R and the dataset `size` $= 40$ and `mydata` $= 5$, we can use the following R script.

```
library(bbmle) # Activate the package bbmle
options(digits=3) # Set the precision of the output
size<-40
mydata<-c(5)
myfunc<-function(size,prob)(-sum(dbinom(mydata,size,prob,
  log=T))#
mle2(myfunc,start=list(prob=0.5),data=list(size=40))
```

As in Example 14.5, the output is `prob=0.125`.

Likelihood Function for Censored Datasets

Consider a sample of n independent and identical items. If item i failed at time t_i, its contribution to the likelihood function is

$$L_i(\theta \mid t_i) = f(t_i \mid \theta) = z(t_i \mid \theta)R(t_i \mid \theta),$$

because to fail at time t_i, the item needs to be functioning just before time t_i [with probability $R(t_i \mid \theta)$] and then it must fail in a very short interval at t_i. Recall the definition of the failure rate function. Here, $f(\cdot)$ and $R(\cdot)$ are regarded as functions of the parameter θ, and t_i is a specific and known time.

If, on the other hand, item i is still functioning at time t_i, all we know is that its time-to-failure exceeds t_i. The contribution to the likelihood function is then

$$L_i(\theta \mid t_i) = R(t_i \mid \theta).$$

Let, as before, δ_i be a failure indicator for item i, such that $\delta_i = 1$ if item i fails and $\delta_i = 0$ if item i is (right) censored, for $i = 1, 2, \ldots, n$. The likelihood function may now be written as

$$L(\theta \mid t_1, t_2, \ldots, t_n) = \prod_{i=1}^{n} L_i(\theta \mid t_i) = \prod_{i=1}^{n} [z(t_i)]^{\delta_i} R(t_i). \tag{14.21}$$

When the failure rate is a constant λ, the likelihood function is

$$L(\lambda \mid t_1, t_2, \ldots, t_n) = \prod_{j=1}^{n} \lambda^{\delta_i} e^{-\lambda t_i}.$$

14.4.5 Exponentially Distributed Lifetimes

The exponential distribution plays an important role in system reliability analysis, and we therefore treat estimation in this distribution separately. Let T be the time-to-failure of an item and assume that T is exponentially distributed with failure rate λ, such that $T \sim \exp(\lambda)$. Further, assume that the survival times of n identical and independent items are observed. The times-to-failure of the n items.

$T_1, T_2, \dots, T_n$ are therefore independent and identically distributed, exp(λ). The dataset of observed survival times is $\boldsymbol{t} = (t_1, t_2, \dots, t_n)$. The dataset may be complete or censored.

Exponentially Distribution: Complete Sample

The joint probability density function of $T_1, T_2, \dots, T_n$ is

$$f(t_1, t_2, \dots, t_n \mid \lambda) = \prod_{i=1}^{n} \lambda \exp(-\lambda t_i) = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} t_i\right).$$

The corresponding likelihood function is

$$L(\lambda \mid \boldsymbol{t}) = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} t_i\right),$$

and the log-likelihood function becomes

$$\ell(\lambda \mid \boldsymbol{t}) = n \log \lambda - \lambda \sum_{i=1}^{n} t_i. \tag{14.22}$$

The MLE is found by setting the derivative of the log-likelihood function equal to zero.

$$\frac{d}{d\lambda}\ell(\lambda \mid \boldsymbol{t}) = \frac{n}{\lambda} - \sum_{i=1}^{n} t_i = 0.$$

Solving for λ gives the ML estimate

$$\widehat{\lambda} = \frac{n}{\sum_{i=1}^{n} t_i}.$$

The corresponding ML *estimator* is

$$\widehat{\lambda} = \frac{n}{\sum_{i=1}^{n} T_i}. \tag{14.23}$$

The ML estimate can hence be expressed by the sample average $\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i$, as

$$\widehat{\lambda} = \frac{1}{\bar{t}}.$$

When a complete dataset $D = \{t_1, t_2, \dots, t_n\}$ is available, the ML estimate can be calculated in R as `1/mean(D)`, and a special R package is not required.

Example 14.8 (Exponential distribution, complete sample)

Assume that we have $n = 10$ observed values and that $\sum_{i=1}^{10} t_i = 68\,450$ hours. With these data, the likelihood function is shown in Figure 14.13 as a function of λ.

The ML estimate in this case is

$$\widehat{\lambda} = \frac{n}{\sum_{i=1}^{10} t_i} = \frac{10}{68\,450\ \text{h}} \approx 1.461 \times 10^{-4}\ \text{h}^{-1},$$

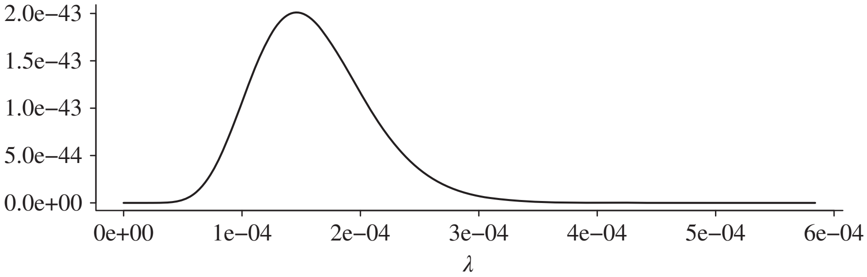


Figure 14.13 Likelihood function for the exponential distribution in Example 14.8.

a value that corresponds with the maximum of the likelihood curve in Figure 14.13. □

We now study the properties of the ML estimator and first find out whether or not it is unbiased.

Because $T_i \sim \exp(\lambda)$, $2\lambda T_i$ is χ^2 distributed with two degrees of freedom for $i = 1, 2, \dots, n$ (e.g. see Ross 2014). Because the T_is are independent, $2\lambda \sum_{i=1}^n T_i$ is χ^2 distributed with $2n$ degrees of freedom.

The ML estimator can be written as

$$\widehat{\lambda} = \frac{n}{\sum_{i=1}^n T_i} = \frac{2n\lambda}{2\lambda \sum_{i=1}^n T_i},$$

and has the same distribution as $2n\lambda/Z$, where Z is χ^2 distributed with $2n$ degrees of freedom. Accordingly,

$$E(\widehat{\lambda}) = 2n\lambda\, E\left(\frac{1}{Z}\right).$$

Here,

$$\begin{aligned} E\left(\frac{1}{Z}\right) &= \int_0^\infty \frac{1}{z}\frac{1}{2^n}\frac{1}{\Gamma(n)} z^{n-1} e^{-z/2}\, dz \\ &= \frac{1}{2(n-1)} \int_0^\infty \frac{1}{2^{n-1}\Gamma(n-1)} z^{n-2} e^{-z/2}\, dz \\ &= \frac{1}{2(n-1)}. \end{aligned}$$

Therefore,

$$E(\widehat{\lambda}) = 2n\lambda\, \frac{1}{2(n-1)} = \frac{n}{n-1}\lambda.$$

The estimator $\widehat{\lambda}$ is accordingly not unbiased, but the estimator λ^*, given by

$$\lambda^* = \frac{n-1}{n}\,\widehat{\lambda} = \frac{n-1}{\sum_{i=1}^n T_i}$$

is seen to be unbiased. Let us determine var(λ^*).

$$\text{var}(\lambda^*) = \left(\frac{n-1}{n}\right)^2 \text{var}\ (\lambda^*) = 4(n-1)^2\lambda^2 \text{var}\ \left(\frac{1}{Z}\right),$$

where Z has the same meaning as above. Now,

$$\text{var}\ \left(\frac{1}{Z}\right) = E\left(\frac{1}{Z^2}\right) - \left[E\left(\frac{1}{Z}\right)\right]^2$$

and

$$E\left(\frac{1}{Z^2}\right) = \int_0^\infty \frac{1}{z^2}\frac{1}{2^n}\frac{1}{\Gamma(n)} z^{n-1} e^{-z/2}\, dz = \frac{1}{4(n-1)(n-2)}.$$

Hence,

$$\begin{aligned}\text{var}(\lambda^*) &= 4(n-1)^2\lambda^2\left(\frac{1}{4(n-1)(n-2)} - \frac{1}{4(n-1)^2}\right)\\ &= (n-1)\lambda^2\left(\frac{1}{n-2} - \frac{1}{n-1}\right) = \frac{\lambda^2}{n-2}.\end{aligned}$$

The estimator

$$\lambda^* = \frac{n-1}{\sum_{i=1}^n T_i} \tag{14.24}$$

is therefore unbiased and has variance

$$\text{var}(\lambda^*) = \frac{\lambda^2}{n-2}. \tag{14.25}$$

To establish a $1-\varepsilon$ confidence interval for λ, we use the fact that $2\lambda\sum_{i=1}^n T_i$ is χ^2 distributed with $2n$ degrees of freedom. Hence,

$$\Pr\left(z_{1-\varepsilon/2,2n} \le 2\lambda \sum_{i=1}^n T_i \le z_{\varepsilon/2,2n}\right) = 1-\varepsilon$$

and

$$\Pr\left(\frac{z_{1-\varepsilon/2,2n}}{2\sum_{i=1}^n T_i} \le \lambda \le \frac{z_{\varepsilon/2,2n}}{2\sum_{j=1}^n T_j}\right) = 1-\varepsilon.$$

Thus, a $1-\varepsilon$ confidence interval for λ is

$$\left(\frac{z_{1-\varepsilon/2,2n}}{2\sum_{i=1}^n T_i},\ \frac{z_{\varepsilon/2,2n}}{2\sum_{j=1}^n T_j}\right). \tag{14.26}$$

Total-Time-on-Test

Let $T_{(1)} \le T_{(2)} \le \cdots \le T_{(n)}$ be the order statistics for the variables $T_1, T_2, \dots, T_n$, and similarly, let $t_{(1)} \le t_{(2)} \le \cdots \le t_{(n)}$ be the ordered dataset that is obtained from the experiment. Assume that all the n items are put into operation at the same time $t = 0$.

We introduce the symbol $\mathcal{T}(t)$ for the accumulated time in operation in the interval $(0, t)$, and call $\mathcal{T}(t)$ the *total-time-on-test* (TTT) at time t. At time $t_{(1)}$, the n items have accumulated a time in operation $\mathcal{T}(t_{(1)}) = nt_{(1)}$. Just after time $t_{(1)}$, there are $n-1$ items left in operation. The accumulated time in operation at time $t_{(2)}$ is therefore $\mathcal{T}(t_{(2)}) = nt_{(1)} + (n-1)(t_{(2)} - t_{(1)})$.

Let $d_i = t_{(i)} - t_{(i-1)}$ be the time interval between the termination of the operation of the $(i-1)$th entry and the termination of the ith entry, such that

$$\begin{aligned} t_{(1)} &= d_1 \\ t_{(2)} &= d_1 + d_2 \\ \vdots \quad & \quad \vdots \\ t_{(r)} &= d_1 + d_2 + \cdots + d_r. \end{aligned}$$

The TTT at time $t_{(r)}$ has two parts

(1) The time on test of the items that have failed in the interval $(0, t_{(r)}]$, which is $\sum_{i=1}^{r} t_{(i)} = rd_1 + (r-1)d_2 + \cdots + d_r$.

(2) The time on test of the $n-r$ items that are still in operation at time $t_{(r)}$, which is $(n-r)t_{(r)} = (n-r)\sum_{i=1}^{r} d_i$.

The TTT at time $t_{(r)}$ is therefore,

$$\begin{aligned} \mathcal{T}(t_{(r)}) &= \sum_{i=1}^{r} t_{(i)} + (n-r)t_{(r)} \\ &= rd_1 + (r-1)d_2 + \cdots + d_r + (n-r)\sum_{i=1}^{r} d_i. \end{aligned}$$

Tidying up this expression yields

$$\mathcal{T}(t_{(r)}) = \sum_{i=1}^{r} [n - (i-1)]d_i. \tag{14.27}$$

By introducing the corresponding random variables, we obtain

$$\mathcal{T}(T_{(r)}) = \sum_{i=1}^{r} [n - (i-1)]D_i. \tag{14.28}$$

Exponentially Distribution: Censored Data

Assume that n independent and identical items with constant failure rate λ have been observed until either failure or censoring. We assume that there are no ties in the dataset $\{t_1, t_2, \dots, t_n\}$. As before, let $\delta_i = 1$ when survival time t_j is a failure time, and $\delta_j = 0$ when t_j is a censored time, for $j = 1, 2, \dots, n$. From (14.21), the likelihood function may then be written as

$$L(\lambda \mid t_1, t_2, \dots, t_n) = \prod_{j=1}^{n} \lambda^{\delta_i} e^{-\lambda t_i}. \tag{14.29}$$

Censoring of Type II

For censoring of type II, the life test is terminated as soon as r failures have been observed. The ordered dataset may be written as $t_{(1)} < t_{(2)} < \cdots < t_{(r)} < t_{(r+1)} < \cdots < t_{(n)}$, for $r < n$. The dataset contains r times-to-failure and $n - r$ censored times. This means that t_r is the longest time-to-failure. The likelihood function for this situation is (see Remark 14.4)

$$L(\lambda \mid t_{(1)}, \dots, t_{(r)}) \propto \lambda^r \exp\left(-\lambda \left[\sum_{j=1}^{r} t_{(j)} + (n-r)t_{(r)} \right] \right)$$
$$= \lambda^r \exp[-\lambda \mathcal{T}(t_{(r)})] \quad \text{for } 0 < t_{(1)} < \cdots < t_{(r)}.$$

The log-likelihood function is

$$\ell(\lambda \mid \boldsymbol{t}) \propto r \log \lambda - \lambda \mathcal{T}(t_{(r)}),$$

where $\boldsymbol{t} = (t_{(1)}, t_{(2)}, \dots, t_{(r)})$. The MLE is found by setting the derivative of the log-likelihood function equal to zero.

$$\frac{d}{d\lambda}\ell(\lambda \mid \boldsymbol{t}) = \frac{r}{\lambda} - \mathcal{T}(t_{(r)}) = 0.$$

The ML estimate λ^*_{II} of λ is, therefore,

$$\lambda^*_{\text{II}} = \frac{r}{\mathcal{T}(t_{(r)})}.$$

The corresponding ML *estimator* is

$$\lambda^*_{\text{II}} = \frac{r}{\mathcal{T}(T_{(r)})}. \tag{14.30}$$

The TTT at time $T_{(r)}$ is

$$\mathcal{T}(T_{(r)}) = nD_1 + (n-1)D_2 + \cdots + [n-(r-1)]D_r$$
$$= \sum_{j=1}^{r} [n-(j-1)]D_j.$$

Introducing

$$D^*_j = [n-(j-1)]D_j \quad \text{for } j = 1, 2, \dots, r.$$

we know that $2\lambda D^*_1, 2\lambda D^*_2, \dots, 2\lambda D^*_r$ are independent and χ^2 distributed, each with 2 degrees of freedom. Hence, $2\lambda\mathcal{T}(T_{(r)})$ is χ^2 distributed with $2r$ degrees of freedom, and we can utilize this to find $E(\lambda^*_{\text{II}})$.

$$E(\lambda^*_{\text{II}}) = E\left(\frac{r}{\mathcal{T}(T_{(r)})}\right) = 2\lambda r E\left(\frac{1}{2\lambda\mathcal{T}(T_{(r)})}\right) = 2\lambda r E\left(\frac{1}{Z}\right),$$

where Z is χ^2 distributed with $2r$ degrees of freedom. This implies that

$$E\left(\frac{1}{Z}\right) = \frac{1}{2(r-1)}.$$

Hence,

$$E(\lambda_{\mathrm{II}}^*) = 2\lambda r \frac{1}{2r-1} = \lambda \frac{r}{(r-1)}.$$

The estimator λ_{II}^* is accordingly not unbiased, but

$$\lambda_{\mathrm{II}}^* = \frac{(r-1)}{\mathcal{T}(T_{(r)})} \tag{14.31}$$

is seen to be unbiased. By the method used for a complete dataset, we find that

$$\mathrm{var}(\widehat{\lambda}_{\mathrm{II}}) = \frac{\lambda^2}{(r-2)}.$$

Confidence intervals, as well as tests for standard hypotheses about λ, may now be derived from the fact that $2\lambda\mathcal{T}(T_{(r)})$ is χ^2 distributed with $2r$ degrees of freedom.

Censoring of Type I

The fact that the number (S) of items failing before time t_0 is random, makes this situation more difficult to deal with from a probabilistic point of view. We therefore confine ourselves to suggesting an intuitive estimator for λ.

First, observe that the estimators for λ, derived in the case of complete datasets and of type II censored data, both could be written as a fraction with numerator equal to "number of recorded failures −1" and denominator equal to "total-time-on-test at the termination of the test." It seems intuitively reasonable to use the same fraction when we have type I censoring.

In this case, the number of failures is S and the TTT is

$$\mathcal{T}(t_0) = \sum_{j=1}^{S} T_{(j)} + (n-S)t_0. \tag{14.32}$$

Hence,

$$\widehat{\lambda}_{\mathrm{I}} = \frac{S-1}{\mathcal{T}(t_0)}$$

seems to be a reasonable estimator for λ.

It can be shown that this estimator is biased for small samples, but asymptotically, it has the same properties as $\widehat{\lambda}_{\mathrm{II}}$ (see Mann et al. 1974, p. 173).

14.4.6 Weibull Distributed Lifetimes

Another important distribution in system reliability analyses is the Weibull distribution. To find the MLEs for the parameters of the Weibull distribution is more complicated than for the exponential distribution. We suffice by treating complete datasets only.

Complete Sample

Let $T_1, T_2, \dots, T_n$ be a complete sample of lifetimes that are independent and identical to Weibull distribution with probability density

$$f_T(t) = \frac{\alpha}{\theta}\left(\frac{t}{\theta}\right)^{\alpha-1} \exp\left[-\left(\frac{t}{\theta}\right)^{\alpha}\right] \quad \text{for } t > 0, \alpha > 0, \theta > 0.$$

The likelihood function is

$$L(\alpha, \theta \mid t_1, t_2, \dots, t_n) = \prod_{j=1}^{n} \frac{\alpha}{\theta}\left(\frac{t_j}{\theta}\right)^{\alpha-1} \exp\left[-\left(\frac{t_j}{\theta}\right)^{\alpha}\right], \tag{14.33}$$

and the log-likelihood is

$$\begin{aligned}\ell(\alpha, \theta \mid t_1, t_2, \dots, t_n) &= \sum_{j=1}^{n}\left[\log\alpha - \alpha\log\theta + (\alpha-1)\log t_j - \left(\frac{t_j}{\theta}\right)^{\alpha}\right] \\ &= n\log\alpha - n\alpha\log\theta + \sum_{j=1}^{n}(\alpha-1)\log t_j - \sum_{j=1}^{n}\left(\frac{t_j}{\theta}\right)^{\alpha}.\end{aligned}$$

The likelihood equations become

$$\frac{\partial\ell}{\partial\theta} = -\frac{n\alpha}{\theta} + \frac{\alpha}{\theta^{\alpha+1}}\sum_{j=1}^{n} t_j^{\alpha} = \frac{\alpha n}{\theta^{\alpha}}\left(\frac{1}{n}\sum_{j=1}^{n} t_j^{\alpha} - \theta^{\alpha}\right) = 0.$$

Solving this equation yields

$$\theta = \left(\frac{1}{n}\sum_{j=1}^{n} t_j^{\alpha}\right)^{1/\alpha}. \tag{14.34}$$

The derivative with respect to α is

$$\begin{aligned}\frac{\partial\ell}{\partial\alpha} &= \frac{n}{\alpha} - n\log\theta + \sum_{j=1}^{n}\log t_j + \sum_{j=1}^{n}\left(\frac{t_j}{\theta}\right)^{\alpha}\log\left(\frac{t_j}{\theta}\right) \\ &= \frac{n}{\alpha} - n\log\theta + \sum_{j=1}^{n}\log t_j + \frac{1}{\theta^{\alpha}}\sum_{j=1}^{n} t_j^{\alpha}(\log t_j - \log\theta).\end{aligned}$$

Inserting (14.46) gives the MLE equation

$$\frac{1}{n}\sum_{j=1}^{n}\log t_j + \frac{1}{\alpha} - \frac{\sum_{j=1}^{n} t_j^{\alpha}\log t_j}{\sum_{j=1}^{n} t_j^{\alpha}} = 0.$$

This is an equation with a single unknown parameter α. We may therefore solve for α to obtain the MLE $\widehat{\alpha}$. It can be proved that there is a unique solution for α.

Weibull Analysis with R

Several R packages can be used to determine ML estimates for the Weibull distribution. Among these are `bbmle`, `stat4`, and `survival`. If you want to use one of these, you should read the package documentation carefully and also search the Internet for example scripts.

A dedicated R package for Weibull analysis, called `WeibullR`, is further available, but is still under development. The package can be used to find the ML estimates for both two-parameter and three-parameter Weibull distributions. Examples of R scripts may be found in the package documentation. The package can be used for both complete and censored datasets. `WeibullR` provides several approaches to estimating the parameters α and θ for a two-parameter Weibull distribution. Here, we illustrate the most simple approach. In the basic setup, we enter the times-to-failure and the censoring times as separate vectors, as shown in the following R script

```
library(WeibullR)
failtime<-c(31.7,39.2,57.5,65.8,70.0,101.7,109.2,130.0)
censored<-c(65.0,75.0,75.2,87.5,88.3,94.2,105.8,110.0)
# Prepare the data for analysis
data<-wblr.conf(wblr.fit(wblr(failtime,censored)),lwd=1)
plot(data)
```

The function `wblr` is used to prepare the dataset for usage in `WeibullR`. The resulting plot is shown in Figure 14.14.

Observe that the plot in Figure 14.14 is obtained by a simplified procedure in `WeibullR` using only default settings. The default names of parameters are “beta” for α and “eta” for θ. The estimates obtained are $\alpha = 2.35$ and $\theta = 115.2$. Confidence bounds are supplied. To choose a more advanced estimation procedure and to adjust the settings, the reader should read the `WeibullR` documentation carefully.

Censoring of Type II

With censoring of type II, the dataset contains r times-to-failure, and $n - r$ censored times, and the censoring takes place at time $t_{(r)}$. Analogous with (14.33) the likelihood function is proportional with

$$\begin{aligned} L(\alpha, \theta \mid \boldsymbol{t}) &\propto \prod_{j=1}^{r} \frac{\alpha}{\theta} \left(\frac{t_{(j)}}{\theta} \right)^{\alpha - 1} \exp\left(-\frac{t_{(j)}}{\theta} \right)^{\alpha} \exp\left(-\left(\frac{t_{(r)}}{\theta} \right)^{\alpha} \right)^{n-r} \\ &= \alpha^{r} \theta^{-\alpha r} \prod_{j=1}^{r} t_{(j)}^{\alpha-1} \exp\left(-(n-r) \left(\frac{t_{(r)}}{\theta} \right)^{\alpha} \right), \end{aligned}$$

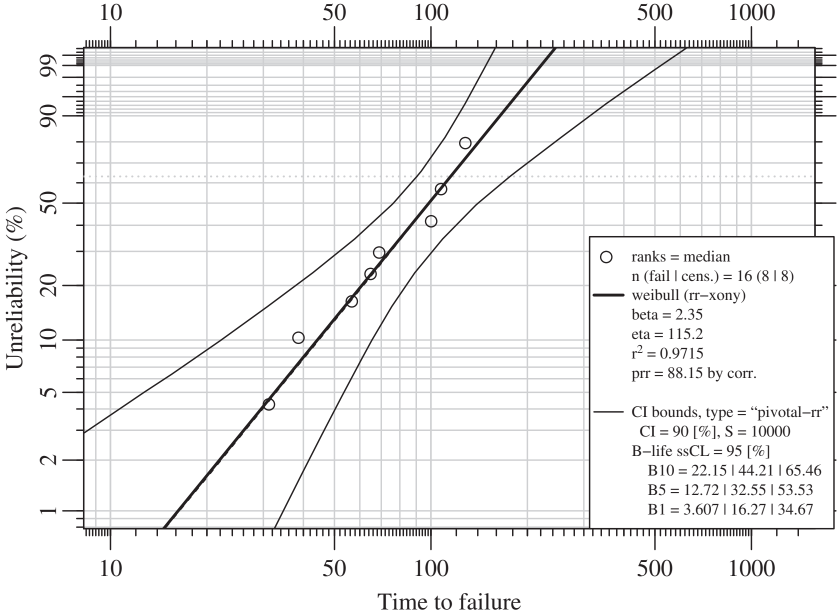


Figure 14.14 Output from a simple script using WeibullR.

where $\boldsymbol{t}$ is the ordered dataset, that is, the r times-to-failure and the $n - r$ censoring times that are all equal to $t_{(r)}$. The log-likelihood is

$$\ell(\alpha, \lambda \mid \boldsymbol{t}) = r \log \alpha - r\alpha \log \theta + (\alpha - 1) \sum_{j=1}^{r} \log t_{(j)} \\ - \sum_{j=1}^{r} \left(\frac{t_{(j)}}{\theta} \right)^{\alpha} - (n - r) \left(\frac{t_{(r)}}{\theta} \right)^{\alpha}.$$

Analogous with the complete data situation, we can determine the MLE estimates α^* and λ^* from

$$\lambda^* = \left(\frac{r}{\sum_{j=1}^{r} t_{(j)}^{\alpha^*} + (n - r) t_{(r)}^{\alpha^*}} \right)^{1/\alpha^*} \tag{14.35}$$

and

$$\frac{r}{\alpha^*} + \sum_{j=1}^{r} \log t_{(j)} - \frac{r \ \sum_{j=1}^{r} t_{(j)}^{\alpha^*} \log t_{(j)} + (n - r) t_{(r)}^{\alpha^*} \log t_{(r)}}{\sum_{j=1}^{r} t_{(j)}^{\alpha^*} + (n - r) t_{(r)}^{\alpha^*}} = 0. \tag{14.36}$$

For further details on ML estimation in the Weibull distribution, e.g. see Meeker and Escobar (1998) and McCool (2012).

14.5 The Kaplan–Meier Estimate

A nonparametric estimate for the survivor function $R(t) = \Pr(T > t)$ was introduced by Kaplan and Meier (1958) and is called the *Kaplan–Meier estimate*.[7] A valued feature of the Kaplan–Meier estimate is that it provides an intuitive graphical representation. We first introduce the estimate for a complete dataset.

14.5.1 Motivation for the Kaplan–Meier Estimate Based a Complete Dataset

Consider a *complete* dataset without ties. For this dataset, the obvious estimate for $R(t)$ is the *empirical survivor function*, which is presented in Section 14.3.5. The empirical survivor function is based on binomial reasoning for each failure time t. As a motivation for the Kaplan–Meier estimate, we now develop the empirical survivor function by a different approach based on the ordered (complete) dataset $0 = t_{(0)} < t_{(1)} < t_{(2)} < \cdots < t_{(n)}$.

Consider a particular survival time, say $t_{(i)}$, in this dataset. To survive $t_{(i)}$, the item has to survive the first interval $(0, t_{(1)})$. Given that this interval is survived, the item has to survive the next interval $(t_{(1)}, t_{(2)})$, and so on, until it must survive the interval $(t_{(i-1)}, t_{(i)})$. Let $t_{(0)} = 0$. The probability of surviving the first interval is

$$R(t_{(1)}) = \Pr(T > t_{(1)}) = \Pr(T > t_{(1)} \mid T > t_{(0)}) = R(t_{(1)} \mid t_{(0)}).$$

The probability of surviving the next interval (when it is known that it has survived the first interval) is

$$R(t_{(2)} \mid t_{(1)}) = \Pr(T > t_{(2)} \mid T > t_{(1)}),$$

and so on. This means that the survivor function at time $t_{(i)}$ can be expressed by using the multiplication rule for conditional probabilities as

$$R(t_{(i)}) = \prod_{j=1}^{i} R(t_{(j)} \mid t_{(j-1)}), \tag{14.37}$$

where $R(t_{(0)}) = R(0) = 1$.

Each factor in (14.37) can be estimated with the same binomial approach we used to obtain the empirical distribution function. Just before time $t_{(1)}$, $n_1 = n$ items are in the at-risk-set and may fail, just before time $t_{(2)}$, $n_2 = n - 1$ items are in the at-risk-set and may fail, and so on. Because we have a complete dataset without censoring and ties, the number of items that failed at time $t_{(j)}$, is $d_j = 1$. The number of items that survives $t_{(j)}$ is therefore $n_j - d_j = (n - j - 1)$.

[7] Named after the authors: Edward Lynn Kaplan (1920–2006) and Paul Meier (1924–2011).

Based on the binomial model, we may then estimate the factors of (14.37) as

$$\hat{R}(t_{(1)} \mid t_{(0)}) = \frac{n_1 - d_1}{n_1} = 1 - \frac{d_1}{n_1} = 1 - \frac{1}{n}$$
$$\hat{R}(t_{(2)} \mid t_{(1)}) = \frac{n_2 - d_2}{n_2} = 1 - \frac{d_2}{n_2} = 1 - \frac{1}{n-1},$$

and so on.

If we use this result in (14.37), we obtain a reformulated estimate for the empirical survivor function

$$\begin{aligned}\hat{R}(t) &= \prod_{j;t_{(j)}<t} \hat{R}(t_{(j)} \mid t_{(j-1)}) = \prod_{j;t_{(j)}<t} \left(1 - \frac{d_j}{n_j}\right) \\ &= \prod_{j;t_{(j)}<t} \left(1 - \frac{1}{n-j+1}\right). \end{aligned} \tag{14.38}$$

For $t > t_{(n)}$, all the n items are failed and $\hat{R}(t) = 0$. The reason why we have written the empirical survivor function in such a complicated way is to pave the way for the introduction of the Kaplan–Meier estimate.

14.5.2 The Kaplan–Meier Estimator for a Censored Dataset

Kaplan and Meier (1958) extend the empirical survivor function to a randomly censored dataset that may also include ties. Their approach is very similar to our derivation of the empirical survivor function and their estimate is given as

$$\hat{R}(t) = \prod_{j;t_{(j)}<t} \left(1 - \frac{d_j}{n_j}\right).$$

The only difference from (14.38) is the values for d_j and n_j. If $t_{(j)}$ is a censoring time, $d_j = 0$, the factor $(1 - d_j/n_j) = 1$ and does not directly influence the estimate $\hat{R}(t)$, but the censoring influences the at-risk-set before the next event (failure or censoring).

We may rewrite the definition of the Kaplan–Meier estimate to include the information of whether a survival time is a failure or a censoring time by including the status δ_j in the formula

$$\hat{R}(t) = \prod_{j;t_{(j)}<t,\delta_j=1} \left(1 - \frac{d_j}{n_j}\right), \tag{14.39}$$

where the product includes all items j that have a *failure* time (i.e. $\delta_j = 1$) such that $t_{(j)} < t$. This formula clearly shows that the factors are only included for survival times that represent failure. Survival times that represent censoring give a factor equal to one and will hence not influence the estimate directly.

With $\widehat{p}_j = 1 - d_j/n_j = (n_j - d_j)/n_j$ we may write (14.54) as

$$\widehat{R}(t) = \prod_{j;t_{(j)}<t,\delta_j=1} \widehat{p}_j. \tag{14.40}$$

The estimate $\widehat{R}(t)$ (14.39) and (14.40) is known as the *Kaplan–Meier estimate* and is also called the *product limit (*PL*) estimate*. The procedure to calculate the Kaplan–Meier estimate is illustrated in Example 14.9.

Example 14.9 (Kaplan–Meier estimate)
Consider the ordered dataset in Table 14.2. The dataset has 16 survival times, of which 9 are censored times (status $\delta = 0$) and 7 are failure times (status $\delta = 1$). The dataset has no ties. With no ties, the number of items at risk just before survival time $t_{(j)}$ is $n_j = n - j + 1$, as listed in the second column of Table 14.2.

Immediately before $t_{(1)}$, $n = 16$ items were at risk. After the failure at $t_{(1)}$, $n - 2 + 1$ items are at risk before $t_{(2)}$, and similar for the other failure times. The Kaplan–Meier estimate $\widehat{R}(t)$ is found from (14.40) by multiplying the $\widehat{p}_j$'s for all survival times $\leq t$.

In Table 14.3, the Kaplan–Meier estimate is presented as a function of time. In the time interval $(0, 31.7)$ until the first failure, it is reasonable to set $\widehat{R}(t) = 1$. The estimate may be displayed graphically as a *Kaplan–Meier plot*. □

Kaplan–Meier Estimate with R

The Kaplan–Meier estimate is available in the R package `survival` and a Kaplan–Meier plot is generated by the script

```
library(survival)
survtime <- c(31.7,39.2,57.5,65.0,65.8,70,0,75.0,75.2,
87.5,88.3,94.2,101.7,105.8,109.2,110.0,130.0)
status <- c(1,1,1,0,1,1,0,0,0,0,0,0,1,0,1,0,1)
data<- Surv(survtime,status==1)
km <- survfit(Surv(survtime, status==1)~1,conf
.type="none")
plot(km,xlab="Time t",ylab="Survival probability")
```

The additional command `print(summary(km))` gives a summary of the results.

```
time n.risk n.event survival std.err
31.7     16       1    0.938  0.0605
39.2     15       1    0.875  0.0827
```

```
  57.5      14        1     0.812  0.0976
  65.8      12        1     0.745  0.1105
  70.0      11        1     0.677  0.1194
 101.7       5        1     0.542  0.1542
 109.2       3        1     0.361  0.1797
 130.0       1        1     0.000     NaN
```

Observe that these results are the same as we found by hand-calculation in Table 14.2, but the estimates are only presented for failure times.

Table 14.2 Computation of the Kaplan–Meier Estimate (censored times are marked with 0 in column “Status”).

Rank j	**Number at risk** $(n - j + 1)$	**Ordered survival times** $t_{(j)}$	**Status** δ_j	$\hat{p}_j$	$\hat{R}(t_{(j)})$
0	—	—	—	1	1.000
1	16	31.7	1	15/16	0.938
2	15	39.2	1	14/15	0.875
3	14	57.5	1	13/14	0.813
4	13	65.0	0	1	0.813
5	12	65.8	1	11/12	0.745
6	11	70.0	1	10/11	0.677
7	10	75.0	0	1	0.677
8	9	75.2	0	1	0.677
9	8	87.5	0	1	0.677
10	7	88.3	0	1	0.677
11	6	94.2	0	1	0.677
12	5	101.7	1	4/5	0.542
13	4	105.8	0	1	0.542
14	3	109.2	1	2/3	0.361
15	2	110.0	0	1	0.361
16	1	130.0	1	0	0.000

Table 14.3 The Kaplan–Meier estimate as a function of time.

t			$\widehat{\boldsymbol{R}}(\boldsymbol{t})$	
0	$\leq t <$	31.7		$= 1.000$
31.7	$\leq t <$	39.2	$\frac{15}{16}$	$= 0.938$
39.2	$\leq t <$	57.5	$\frac{15}{16} \cdot \frac{14}{15}$	$= 0.875$
57.5	$\leq t <$	65.8	$\frac{15}{16} \cdot \frac{14}{15} \cdot \frac{13}{14}$	$= 0.813$
65.8	$\leq t <$	70.0	$\frac{15}{16} \cdot \frac{14}{15} \cdot \frac{13}{14} \cdot \frac{11}{12}$	$= 0.745$
70.0	$\leq t <$	101.7	$\frac{15}{16} \cdot \frac{14}{15} \cdot \frac{13}{14} \cdot \frac{11}{12} \cdot \frac{10}{11}$	$= 0.677$
101.7	$\leq t <$	109.2	$\frac{15}{16} \cdot \frac{14}{15} \cdot \frac{13}{14} \cdot \frac{11}{12} \cdot \frac{10}{11} \cdot \frac{4}{5}$	$= 0.542$
109.2	$\leq t <$	130.0	$\frac{15}{16} \cdot \frac{14}{15} \cdot \frac{13}{14} \cdot \frac{11}{12} \cdot \frac{10}{11} \cdot \frac{4}{5} \cdot \frac{2}{3}$	$= 0.361$
130.0	$\leq t$		$\frac{15}{16} \cdot \frac{14}{15} \cdot \frac{13}{14} \cdot \frac{11}{12} \cdot \frac{10}{11} \cdot \frac{4}{5} \cdot \frac{2}{3} \cdot \frac{0}{1}$	$= 0.000$

We see from (14.39) that $\widehat{R}(t)$ is a step function, continuous from the right, that equals 1 at t = 0. $\widehat{R}(t)$ drops by a factor of $(n_j - 1)/n_j$ at each failure time $t_{(j)}$. The estimate $\widehat{R}(t)$ does not change at the censored times. The censored times influence the values of n_j (i.e. the at-risk-set) and hence, the size of the steps in $\widehat{R}(t)$.

A slightly problematic point is that $\widehat{R}(t)$ never reduces to zero when the longest survival time $t_{(n)}$ recorded is a censored time. For this reason, $\widehat{R}(t)$ is usually taken to be undefined for $t > t_{(n)}$. This issue is further discussed by Kalbfleisch and Prentice (1980).

Some Properties of the Kaplan–Meier Estimator

A thorough discussion of the properties of the Kaplan–Meier estimator $\widehat{R}(t)$ may be found in Kalbfleisch and Prentice (1980), Lawless (1982), Cox and Oakes (1984), and Aalen et al. (2008). Here, we suffice by summarizing a few properties without proofs:

(1) The Kaplan–Meier estimator $\widehat{R}(t)$ can be derived as a nonparametric MLE. This derivation was originally given by Kaplan and Meier (1958).

(2) $\widehat{R}(t)$ is a consistent estimator of $R(t)$ under quite general conditions with estimated asymptotic variance (e.g. see Kalbfleisch and Prentice 1980, p. 14):

$$\widehat{\text{var}}(\widehat{R}(t)) = [\widehat{R}(t)]^2 \sum_{j \in J_t} \frac{d_j}{n_j(n_j - d_j)}. \tag{14.41}$$

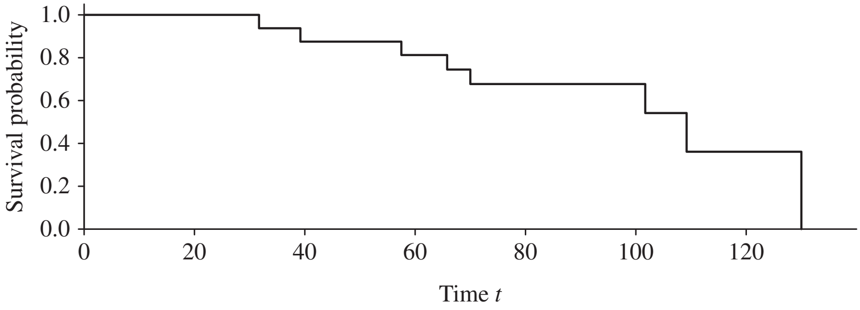


Figure 14.15 Kaplan–Meier plot for the data in Example 14.9. Made with R.

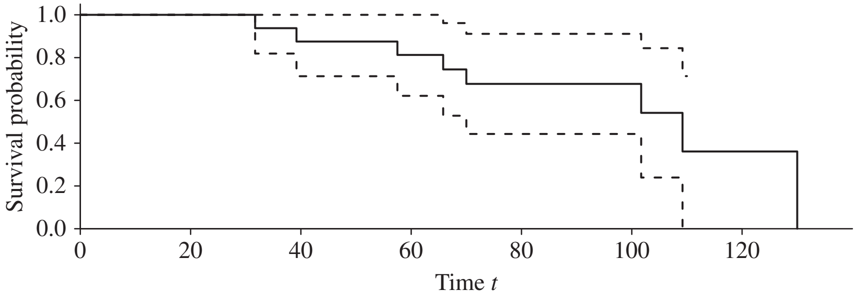


Figure 14.16 Kaplan–Meier plot of the dataset in Example 14.9 with 90% confidence limits, made with R.

Expression (14.41) is known as *Greenwood's formula*.

Confidence limits based on Greenwood's formula are available in R and are obtained by the option `conf.type='plain'` in the R script in Example 14.9. The Kaplan–Meier plot in Figure 14.15 with 90% confidence limits is shown in Figure 14.16. Because the plot is based on only eight failure times, the confidence band is rather wide.

(3) Because it is a maximum likelihood estimator, the Kaplan–Meier estimator has an asymptotic normal distribution. Confidence limits for $R(t)$ can hence be determined using normal approximation. For details see Cox and Oakes (1984).

14.6 Cumulative Failure Rate Plots

Let $R(t)$ be the survivor function for a certain type of items, and assume that the distribution is continuous with probability density $f(t) = R'(t)$, where $f(t) > 0$ for

$t > 0$. No further assumptions are made about the distribution (i.e. a nonparametric model).

The failure rate function was defined in Section 5.3.2 as

$$z(t) = \frac{f(t)}{R(t)} = -\frac{d}{dt} \log R(t).$$

The cumulative failure rate function is

$$Z(t) = \int_0^t z(u)\, du = -\log R(t), \tag{14.42}$$

and the survivor function may therefore be written as

$$R(t) = e^{-Z(t)}.$$

Plotting $Z(t)$ as a function of t gives a *cumulative failure rate plot*. If the plot is convex when plotted on a linear scale, the failure rate function is increasing, and if the plot is concave, the failure rate function is decreasing.

Example 14.10 (Exponential distribution)
The cumulative failure rate function for the exponential distribution, exp(λ), is

$$Z(t) = \lambda t \quad \text{for } t \geq 0,\ \lambda > 0.$$

Plotted as a function of t on a linear scale, the plot of $Z(t)$ is a straight line with slope λ. If we are able to determine an estimate $\widehat{Z}(t)$, the plotted values should follow a reasonably straight line. □

Example 14.11 (Weibull distribution)
The cumulative failure rate function for the Weibull distribution with shape α and scale θ is

$$Z(t) = \left(\frac{t}{\theta}\right)^{\alpha} \quad \text{for } t \geq 0,\ \alpha > 0,\ \theta > 0.$$

Taking logarithm yields

$$\log Z(t) = \alpha \log t - \alpha \log \theta.$$

If $Z(t)$ is plotted versus t on a log–log scale, the plot is a straight line with slope α. If we are able to determine an estimate $\widehat{Z}(t)$, the plotted values should follow a reasonably straight line on a log–log scale. □

The rest of this section is concerned with a particular type of cumulative failure rate plots: the *Nelson–Aalen plot*.

14.6.1 The Nelson–Aalen Estimate of the Cumulative Failure Rate

An obvious estimate of the cumulative failure rate $Z(t)$, based on the Kaplan–Meier estimator $\widehat{R}(t)$, is

$$\widehat{Z}(t) = -\log \widehat{R}(t). \tag{14.43}$$

An alternative estimate of $Z(t)$ is proposed by Nelson (1972) and elaborated by Aalen (1978). This estimate is now known as the *Nelson–Aalen estimate*. Assume that we have a stochastically censored (type IV) dataset. As before, let

$$0 = t_{(0)} < t_{(1)} < t_{(2)} < \cdots < t_{(n)}$$

be the recorded ordered survival times until either failure or censoring, and let δ_j be the status of survival time $t_{(j)}$, for $j = 1, 2, \ldots, n$.

The Nelson–Aalen estimate of the cumulative failure rate is then

$$\widehat{Z}(t) = \sum_{j;t_{(j)}<t,\delta_j=1} \frac{d_j}{n_j}, \tag{14.44}$$

where d_j, as before, is the number of items that fail at time $t_{(j)}$ and n_j is the number of items at risk just before $t_{(j)}$. The Nelson–Aalen estimator of the survivor function at time t is

$$R^*(t) = \exp[-\widehat{Z}(t)]. \tag{14.45}$$

Before we give a justification for these estimators, we illustrate how they are calculated in Example 14.12.

Example 14.12 (Nelson–Aalen estimate for a censored dataset)
Reconsider the censored (type IV) dataset in Table 14.2. The Nelson–Aalen estimate $\widehat{Z}(t)$ may be calculated from (14.44) for the eight failure times $t_{(1)}, t_{(2)}, t_{(3)}, t_{(5)}, t_{(6)}, t_{(12)}, t_{(14)}$, and $t_{(16)}$. Next, $R^*(t)$ is determined from (14.45). The results are shown in Table 14.4. In the last column of Table 14.4, the corresponding Kaplan–Meier estimate $\widehat{R}(t)$ is shown.

As seen, there is good “agreement” between the Kaplan–Meier estimates and the Nelson–Aalen estimates for the survivor function in this dataset, especially for the shortest failure times. For the longest failure times, the discrepancy becomes more significant.

By using the results in Table 14.4, we can now plot the survival times on the x-axis and the corresponding Nelson–Aalen estimates $\widehat{Z}(t)$ on the y-axis and obtain the Nelson–Aalen plot. □

Making the Nelson–Aalen plot manually by the procedure in Example 14.12 may be tedious, but luckily, we may use the R `survival` package.

Table 14.4 Nelson–Aalen estimate for the censored dataset in Example 14.12, compared with the Kaplan–Meier estimate.

j	Survival time	Status δ_j	Nelson–Aalen estimate $\widehat{Z}(t_j)$		Nelson–Aalen $R^*(t_{(j)})$	Kaplan–Meier $\widehat{R}(t_{(j)})$
				$=0.0000$	1.000	1.000
1	31.7	1	$\frac{1}{16}$	$=0.0625$	0.939	0.938
2	39.2	1	$\frac{1}{16}+\frac{1}{15}$	$=0.1292$	0.879	0.875
3	57.5	1	$\frac{1}{16}+\frac{1}{15}+\frac{1}{14}$	$=0.2006$	0.818	0.813
4	65.0	0				
5	65.8	1	$\frac{1}{16}+\frac{1}{15}+\frac{1}{14}+\frac{1}{12}$	$=0.2839$	0.753	0.745
6	70.0	1	$\frac{1}{16}+\frac{1}{15}+\cdots+\frac{1}{11}$	$=0.3748$	0.687	0.677
7	75.0	0				
8	75.2	0				
9	87.5	0				
10	88.3	0				
11	94.2	0				
12	101.7	1	$\frac{1}{16}+\cdots+\frac{1}{11}+\frac{1}{5}$	$=0.5748$	0.563	0.542
13	105.8	0				
14	109.2	1	$\frac{1}{16}+\cdots+\frac{1}{5}+\frac{1}{3}$	$=0.9082$	0.403	0.361
15	110.0	0				
16	130.0	1	$\frac{1}{16}+\cdots+\frac{1}{3}+\frac{1}{1}$	$=1.9082$	0.148	0.000

Nelson–Aalen Plot with R

There is no dedicated package in R for making the Nelson–Aalen plot, but we may use the procedure in the following R script, which illustrates the plot by using the same dataset as in Example 14.12.

```
library(survival)
# Data to be analyzed
survtime<-c(31.7,39.2,57.5,65.0,65.8,70.0,75.0 75.2,
  87.5,88.3,94.2,101.7,105.8,109.2,110.0,130.0)
status<-c(1,1,1,0,1,1,0,0,0,0,0,1,0,1,0,1)
# Prepare hazard data
revrank<-order(survtime,decreasing=T)
haz<- status/revrank
cumhaz<- cumsum(haz)
# Select only failures for plotting.
df<- data.frame(survtime status,cumhaz)
z<- subset(df,status==1)
# Generate cumulative failure rate plot for exp. distr.
plot(z$survtime, z$cumhaz,type="o",pch=19,xlab="Time",
     ylab="Cumulative failure rate")
```

The plot obtained from this script is made with a linear scale on both axes. This means that if the data come from an exponential distribution, the plot should be approximately a straight line (see Example 14.10). The plot is shown in Figure 14.17. The plot is rather far from linear, and we may conclude that the underlying distribution is probably not exponential. To inspect the data used, you may use the commands `print(df)` and `print(z)`.

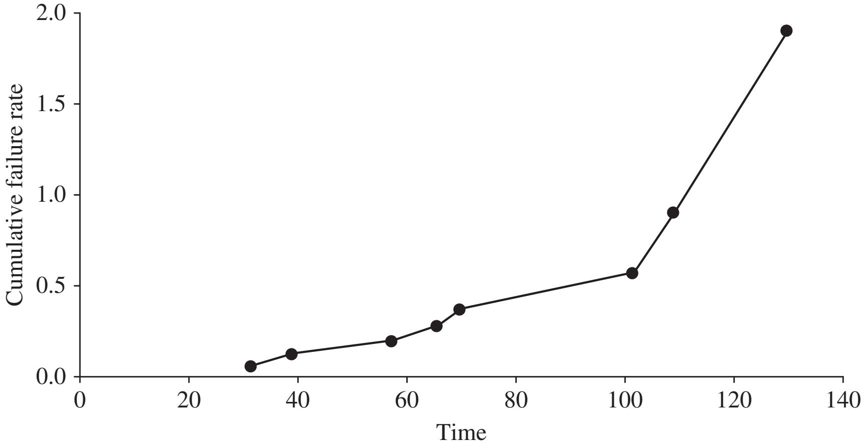


Figure 14.17 Nelson–Aalen plot (linear scale).

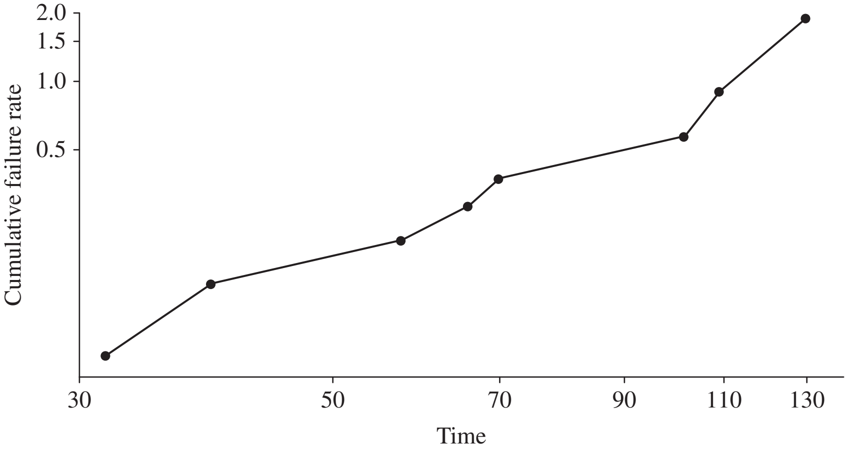


Figure 14.18 Nelson–Aalen plot (log 10 scale).

We may also make the Nelson–Aalen plot with log 10 scale on both axes. As shown in Example 14.11, an approximately straight line would indicate that the underlying distribution may be a Weibull distribution. The plot is obtained by adding the option `log="xy"` to the `plot( )` command in the R script above. The resulting Nelson–Aalen plot is shown in Figure 14.18. The plot is not too far from a straight line, so the Weibull distribution might be an adequate model.

Justification for the Nelson–Aalen Estimate

Some steps in the following justification are approximative and far from rigorous, but we hope the reader may get an understanding of how the Nelson–Aalen estimate is developed. For a more rigorous development of the estimate, see Aalen et al. (2008).

To justify the Nelson–Aalen estimate, we start with arguments similar to those used when introducing the Kaplan–Meier estimate. An ordered dataset $0 = t_{(0)} < t_{(1)} < t_{(2)} < \cdots < t_{(n)}$ is available. The dataset may be censored and include ties. As before, let n_j be the number of items at risk just before survival time $t_{(j)}$ and let d_j be the number of items that fail at time $t_{(j)}$. We again use (14.37)

$$R(t_{(i)}) = \prod_{j=1}^{i} R(t_{(j)} \mid t_{(j-1)}),$$

and assume that the failure rate function in the interval $(t_{(j-1)}, t_{(j)})$ may be approximated by a constant failure rate λ_j, for $j = 1, 2, \ldots$.

For a time t, such that $t_{(m)} < t < t_{(m+1)}$, we get

$$R(t) = \Pr(T > t_{(1)} \mid T > t_{(0)}) \cdots \Pr(T > t \mid T > t_{(m)}). \tag{14.46}$$

As for the Kaplan–Meier estimate, the idea is to estimate each single factor on the right-hand side of (14.46) and use the product of these estimates as an estimate of $R(t)$. What is now a reasonable estimate of $p_j = \Pr(T > t_{(j+1)} \mid T > t_{(j)})$? With the same approach, we used to justify the Kaplan–Meier estimate, the only survival times for which it is natural to estimate p_j with something other than 1, are the survival times $t_{(j)}$ where a *failure* occurs. Because we only consider the times where something happens (failure or censoring), n_j items are at risk in the whole interval $(t_{(j-1)}, t_{(j)})$.

The total functioning time in the $(t_{(j-1)}, t_{(j)})$ is $n_j(t_{(j)} - t_{(j-1)})$. Because we assume a constant failure rate λ_j in $(t_{(j-1)}, t_{(j)})$ a natural estimate of λ_j is

$$\widehat{\lambda}_j = \frac{\text{No. of failures}}{\text{Total functioning time}} = \frac{d_j}{n_j(t_{(j+1)} - t_{(j)})}. \tag{14.47}$$

A natural estimate of p_j, when d_j failures occur at $t_{(j)}$ is therefore

$$\widehat{p}_j = \exp[-\widehat{\lambda}_j(t_{(j)} - t_{(j-1)})] = \exp\left(-\frac{d_j}{n_j}\right). \tag{14.48}$$

Inserting these estimates in (14.46) gives

$$\widehat{R}(t) = \prod_{t_{(j)}<t,\delta_j=1} \exp\left(-\frac{d_j}{n_j}\right) = \exp\left[-\sum_{t_{(j)}<t,\delta_j=1} \frac{d_j}{n_j}\right]. \tag{14.49}$$

Because $R(t) = \exp[-Z(t)]$, a natural estimate for the cumulative failure rate function is

$$\widehat{Z}(t) = \sum_{t_{(j)}<t,\delta_j=1} \frac{d_j}{n_j}, \tag{14.50}$$

which is the Nelson–Aalen estimate.

Uncertainty of the Nelson–Aalen Estimator

The variance of the Nelson–Aalen estimator may be estimated by (e.g. see Aalen et al. 2008)

$$\text{var}[\widehat{Z}(t)] = \widehat{\sigma}^2(t) = \sum_{t_{(j)}<t,\delta_j=1} \frac{(n_j - d_j)\, d_j}{(n_j - 1)\, n_j^2}. \tag{14.51}$$

It may be shown that both the Nelson–Aalen estimator and the variance estimator are close to unbiased. For large samples, it may further be shown that the Nelson–Aalen estimator at time t is approximately normally distributed. We may therefore find a $(1 - \epsilon)$ confidence interval for $\widehat{Z}(t)$ as

$$\widehat{Z}(t) \pm u_{1-\epsilon/2}\widehat{\sigma}(t), \tag{14.52}$$

where $u_{1-\epsilon/2}$ is the $1 - \epsilon/2$ fractile of the standard normal distribution. More properties of the estimator may be found in Aalen et al. (2008).

14.7 Total-Time-on-Test Plotting

A TTT plot is an alternative – but also a supplement – to Kaplan–Meier plot and Nelson–Aalen plots.

14.7.1 Total-Time-on-Test Plot for Complete Datasets

Assume that we have a complete and ordered dataset $t_{(1)} < t_{(2)} < \cdots < t_{(n)}$ of independent lifetimes with continuous distribution function $F(t)$ that is strictly increasing for $F^{-1}(0) = 0 < t < F^{-1}(1)$. Further, it is assumed that the distribution has finite mean μ.

The TTT at time t, $\mathcal{T}(t)$ has earlier been defined as

$$\mathcal{T}(t) = \sum_{j=1}^{i} t_{(j)} + (n-i)t \quad \text{for} \quad i = 0, 1, \ldots, n \qquad \text{and} \qquad t_{(i)} \leq t < t_{(i+1)}, \tag{14.53}$$

and $t_{(0)}$ is defined to be equal to 0 and $t_{(n+1)} = +\infty$.

$\mathcal{T}(t)$ is the total observed lifetime of the n items at time t. We assume that all the n items are put into operation at time $t = 0$ and that the observation is terminated at time t. In the time interval $(0, t]$, a number, i, of the items have failed. The total functioning time of these i items is $\sum_{j=0}^{i} t_{(j)}$. The remaining $n - i$ items survive the time interval $(0, t]$. The total functioning time of these $n - i$ items is thus $(n - i)t$.

The TTT at the ith failure is

$$\mathcal{T}(t_{(i)}) = \sum_{j=1}^{i} t_{(j)} + (n-i)t_{(i)} \quad \text{for} \quad i = 1, 2, \ldots, n. \tag{14.54}$$

In particular,

$$\mathcal{T}(t_{(n)}) = \sum_{j=1}^{n} t_{(j)} = \sum_{j=1}^{n} t_j.$$

The TTT at the ith failure, $\mathcal{T}(t_{(i)})$, may be scaled by dividing by $\mathcal{T}(t_{(n)})$. The *scaled TTT* at time t is defined as $\mathcal{T}(t)/\mathcal{T}(t_{(n)})$.

If we plot the points

$$\left(\frac{i}{n}, \frac{\mathcal{T}(t_{(i)})}{\mathcal{T}(t_{(n)})} \right) \quad \text{for} \quad i = 1, 2, \ldots, n, \tag{14.55}$$

we obtain the *TTT plot* of the dataset.

Example 14.13 Suppose that we have activated 10 identical items and observed their lifetimes (in hours):

6.3	11.0	21.5	48.4	90.1
120.2	163.0	182.5	198.0	219.0

To construct the TTT plot for this (complete) dataset, calculate the necessary quantities and put them in a table as done in Table 14.5. The TTT plot for this (complete) dataset is shown in Figure 14.19. □

To be able to interpret the shape of the TTT plot, we need the following results, which we state without proofs.

(1) Let $U_1, U_2, \dots, U_{n-1}$ be independent random variables with a uniform distribution over $(0, 1]$ (i.e. $U_i \sim \text{unif}(0, 1)$). If the underlying life distribution is *exponential*, the random variables

$$\frac{\mathcal{T}(T_{(1)})}{\mathcal{T}(T_{(n)})}, \frac{\mathcal{T}(T_{(2)})}{\mathcal{T}(T_{(n)})}, \dots, \frac{\mathcal{T}(T_{(n-1)})}{\mathcal{T}(T_{(n)})} \tag{14.56}$$

have the same joint distribution as the $(n-1)$ ordered variables $U_{(1)}, U_{(2)}, \dots, U_{(n-1)}$. For a proof, see Barlow and Campo (1975).

(2) If the underlying life distribution $F(t)$ is exponential, then
   (a) $\text{var}[\mathcal{T}(T_i)/\mathcal{T}(T_n)]$ is finite
   (b) $E[\mathcal{T}(T_i)/\mathcal{T}(T_n)] = 1/n$ for $i = 1, 2, \dots, n$

Table 14.5 TTT Estimates for the dataset in Example 14.15.

i	$t_{(i)}$	$\sum_{j=1}^{i} t_{(j)}$	$\sum_{j=1}^{i} t_{(j)} + (n-i)t_{(i)} = \mathcal{T}(t_{(i)})$	$\frac{i}{n}$	$\frac{\mathcal{T}(t_{(i)})}{\mathcal{T}(t_{(n)})}$
1	6.3	6.3	$6.3 + 9 \cdot 6.3 = 63.0$	0.1	0.06
2	11.0	17.3	$17.3 + 8 \cdot 11.0 = 105.3$	0.2	0.10
3	21.5	38.8	$38.8 + 7 \cdot 21.5 = 189.3$	0.3	0.18
4	48.4	87.2	$87.2 + 6 \cdot 48.4 = 377.6$	0.4	0.36
5	90.1	177.3	$177.3 + 5 \cdot 90.1 = 627.8$	0.5	0.59
6	120.2	297.5	$297.5 + 4 \cdot 120.2 = 778.3$	0.6	0.73
7	163.0	460.5	$460.5 + 3 \cdot 163.0 = 949.5$	0.7	0.90
8	182.5	643.0	$643.0 + 2 \cdot 182.5 = 1008.0$	0.8	0.95
9	198.0	841.0	$841.0 + 1 \cdot 198.0 = 1039.0$	0.9	0.98
10	219.0	1060.0	$1060.0 + 0 = 1060.0$	1.0	1.00

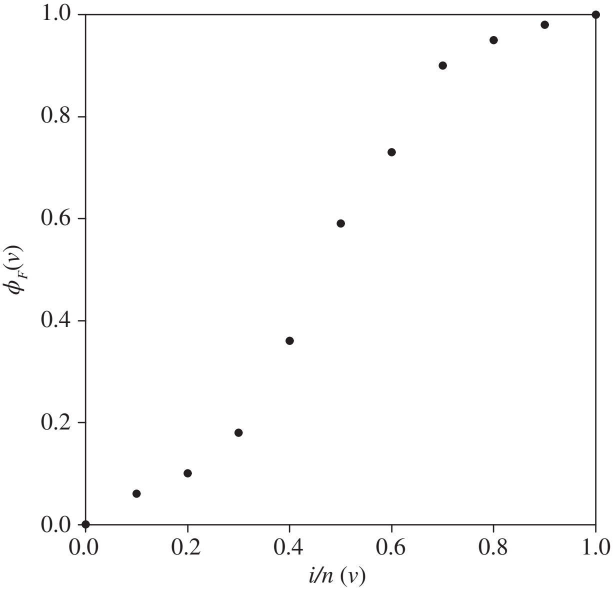


Figure 14.19 TTT plot of the data in Example 14.13.

If the underlying life distribution is exponential, we should, from (14.56), expect that for large n

$$\frac{\mathcal{T}(T_{(i)})}{\mathcal{T}(T_{(n)})} \approx \frac{i}{n} \quad \text{for } i = 1, 2, \dots, (n-1).$$

As this is not the case for the TTT plot in Figure 14.19, we conclude that the underlying life distribution for the data in Example 14.13 is probably not exponential.

To decide from a TTT plot whether or not the corresponding life distribution is increasing failure rate (IFR) or decreasing failure rate (DFR), we need a little more theory. We will be content with a heuristic argument.[8]

We claim that

$$\mathcal{T}(t_{(i)}) = n \int_0^{t_{(i)}} [1 - F_n(u)]\, du, \tag{14.57}$$

where $F_n(t)$ is the empirical distribution function. Assertion (14.57) can be proved in the following way (remember that per definition $t_{(0)} = 0$):

$$n \int_0^{t_{(i)}} [1 - F_n(u)]\, du$$

[8] A rigorous treatment is found, for example, in Barlow and Campo (1975).

$$= n\left[\sum_{j=1}^{i}\int_{t_{(j-1)}}^{t_{(j)}}\left(1-\frac{j-1}{n}\right)\,du\right]$$
$$= \sum_{j=1}^{i}(n-j+1)(t_{(j)}-t_{(j-1)})$$
$$= nt_{(1)}+(n-1)(t_{(2)}-t_{(1)})+\cdots+(n-i+1)(t_{(i)}-t_{(i-1)})$$
$$= \sum_{j=1}^{i}t_{(j)}+(n-i)t_{(i)}=\mathcal{T}(t_{(i)}).$$

We now come to the heuristic part of the argument. First, let n equal $2m+1$, where m is an integer. Then $t_{(m+1)}$ is the median of the dataset. What happens to the integral

$$\int_0^{t_{(m+1)}}[1-F_n(u)]\,du \quad \text{when} \qquad m\to\infty.$$

When $m \to \infty$, we can expect that

$$F_n(u)\to F(u),$$

and that

$$t_{(m+1)}\to\{\text{median of } F\}=F^{-1}(1/2),$$

and therefore that

$$\frac{1}{n}\,\mathcal{T}(t_{(m+1)})\to\int_0^{F^{-1}(1/2)}[1-F(u)]\,du. \tag{14.58}$$

Next, let $n = 4m + 3$. In this case, $t_{(2m+2)}$ is the median of the data, and $t_{(m+1)}$ and $t_{(3m+3)}$ are the lower and upper quartiles, respectively.

When $m \to \infty$, by arguing as we did above, we can expect the following:

$$\begin{aligned}\frac{1}{n}\,\mathcal{T}(t_{(m+1)})&\to\int_0^{F^{-1}(1/4)}[1-F(u)]\,du\\ \frac{1}{n}\,\mathcal{T}(t_{(2m+2)})&\to\int_0^{F^{-1}(1/2)}[1-F(u)]\,du\\ \frac{1}{n}\,\mathcal{T}(t_{(3m+3)})&\to\int_0^{F^{-1}(3/4)}[1-F(u)]\,du.\end{aligned} \tag{14.59}$$

In addition, we have that

$$E(T)=\mu=\int_0^{\infty}[1-F(u)]\,du=\int_0^{F^{-1}(1)}[1-F(u)]\,du. \tag{14.60}$$

When $n \to \infty$, we can therefore expect that

$$\frac{1}{n}\sum_{i=1}^{n}t_i=\frac{1}{n}\,\mathcal{T}(t_{(n)})\to\int_0^{F^{-1}(1)}[1-F(u)]\,du. \tag{14.61}$$

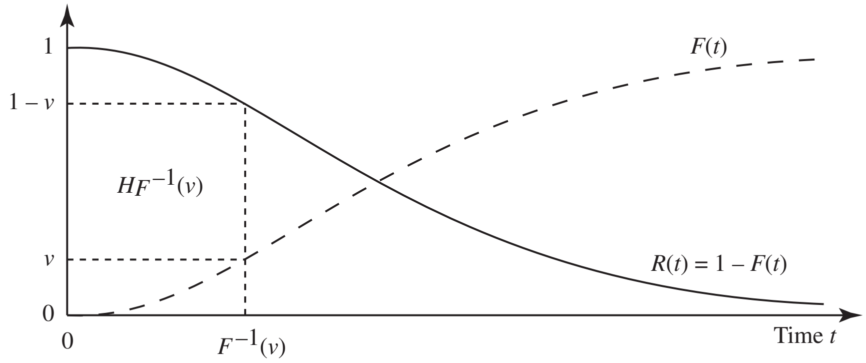


Figure 14.20 The TTT transform of the distribution *F*.

The integrals that we obtain as limits by this approach seem to be of interest and we will look at them more closely. They are all of the type

$$\int_0^{F^{-1}(v)} [1 - F(u)]\, du \quad \text{for} \quad 0 \le v \le 1.$$

The Total-Time-on-Test Transform

We now introduce the *TTT transform* of the distribution $F(t)$ as

$$H_F^{-1}(v) = \int_0^{F^{-1}(v)} [1 - F(u)]\, du \quad \text{for } 0 \le v \le 1. \tag{14.62}$$

The TTT transform of the distribution $F(t)$ is shown in Figure 14.20. Observe that $H_F^{-1}(v)$ is the "area" under the survivor function $R(t)$ between $t = 0$ and $t = F^{-1}(v)$.

It can be shown under assumptions of general nature that there is a one-to-one correspondence between a distribution $F(t)$ and its TTT transform $H_F^{-1}(v)$ (see Barlow and Campo 1975).

Observe from (14.62) that

$$H_F^{-1}(1) = \int_0^{F^{-1}(1)} [1 - F(u)]\, du = \mu. \tag{14.63}$$

The *scaled TTT transform* of $F(t)$ is defined as

$$\varphi_F(v) = \frac{H_F^{-1}(v)}{H_F^{-1}(1)} = \frac{1}{\mu}\, H_F^{-1}(v) \quad \text{for} \quad 0 \le v \le 1. \tag{14.64}$$

Example 14.14 (Exponential distribution)

The distribution function of the exponential distribution is

$$F(t) = 1 - e^{-\lambda t} \quad \text{for} \quad t \ge 0, \qquad \lambda > 0,$$

and hence

$$F^{-1}(v) = -\frac{1}{\lambda}\log(1-v) \quad \text{for} \qquad 0 \le v \le 1.$$

Thus, the TTT transform of the exponential distribution is

$$\begin{aligned} H_F^{-1}(v) &= \int_0^{[-\log(1-v)]/\lambda} e^{-\lambda u}\, du = -\frac{1}{\lambda}\, e^{-\lambda u}\ \Big|_0^{-\frac{1}{\lambda}\log(1-v)} \\ &= \frac{1}{\lambda} - \frac{1}{\lambda}\, e^{\lambda \log(1-v)/\lambda} \\ &= \frac{1}{\lambda} - \frac{1}{\lambda}\,(1-v) = \frac{v}{\lambda} \quad \text{for} \qquad 0 \le v \le 1. \end{aligned}$$

Further

$$H_F^{-1}(1) = \frac{1}{\lambda}.$$

The scaled TTT transform for the exponential distribution is therefore

$$\frac{v/\lambda}{1/\lambda} = v \quad \text{for} \qquad 0 \le v \le 1. \tag{14.65}$$

The scaled TTT transform of the exponential distribution is thus a straight line from $(0,0)$ to $(1,1)$, as shown in Figure 14.21. □

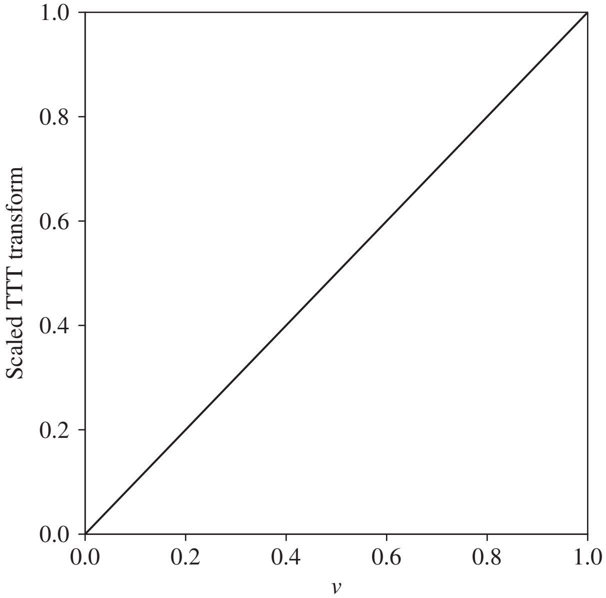


Figure 14.21 Scaled TTT transform of the exponential distribution (Example 14.16).

Example 14.15 (Weibull distribution)

It is usually not straightforward to determine the TTT transform of a life distribution. We illustrate this by trying to determine the TTT transform of the Weibull distribution

$$F(t) = 1 - \exp\left[-\left(\frac{t}{\theta}\right)^{\alpha}\right] \quad \text{for} \quad t \geq 0, \quad \theta > 0, \quad \alpha > 0.$$

The inverse function of F is

$$F^{-1}(v) = \theta\ [-\log(1-v)]^{1/\alpha} \quad \text{for} \quad 0 \leq v \leq 1.$$

The TTT transform of the Weibull distribution is

$$H_F^{-1}(v) = \int_0^{F^{-1}(v)} [1 - F(u)]\ du = \int_0^{\theta\ [-\log(1-v)]^{1/\alpha}} e^{-(u/\theta)^{\alpha}}\ du.$$

By substituting $x = (u/\theta)^{\alpha}$ we obtain

$$H_F^{-1}(v) = \frac{\theta}{\alpha} \int_0^{-\log(1-v)} x^{1/\alpha+1}\ e^{-x}\ dx, \tag{14.66}$$

which shows that the TTT transform of the Weibull distribution may be expressed by the incomplete gamma function. However, several approximation formulas are available.

The mean time-to-failure (MTTF) is obtained by inserting $v = 1$ in $H_F^{-1}(v)$.

$$H_F^{-1}(1) = \frac{\theta}{\alpha} \int_0^{\infty} x^{1/\alpha+1}\ e^{-x}\ dx = \frac{\theta}{\alpha}\Gamma\left(\frac{1}{\alpha}\right) = \theta\ \Gamma\left(\frac{1}{\alpha}+1\right),$$

which coincides with the result we obtained in (5.67). Observe that the scaled TTT transform of the Weibull distribution depends only on the shape parameter α and is independent of the scale parameter θ. Scaled TTT transforms of the Weibull distribution for some selected values of the shape parameter α are shown in Figure 14.22. □

Three Useful Results

We now list three useful results and indicate a proof.

(1) If $F(t)$ is a continuous life distribution that is strictly increasing for $F^{-1}(0) = 0 < t < F^{-1}(1)$, then

$$\frac{d}{dv} H_F^{-1}(v)|_{v=F(t)} = \frac{1}{z(t)}, \tag{14.67}$$

where $z(t)$ is the failure rate of the distribution $F(t)$.

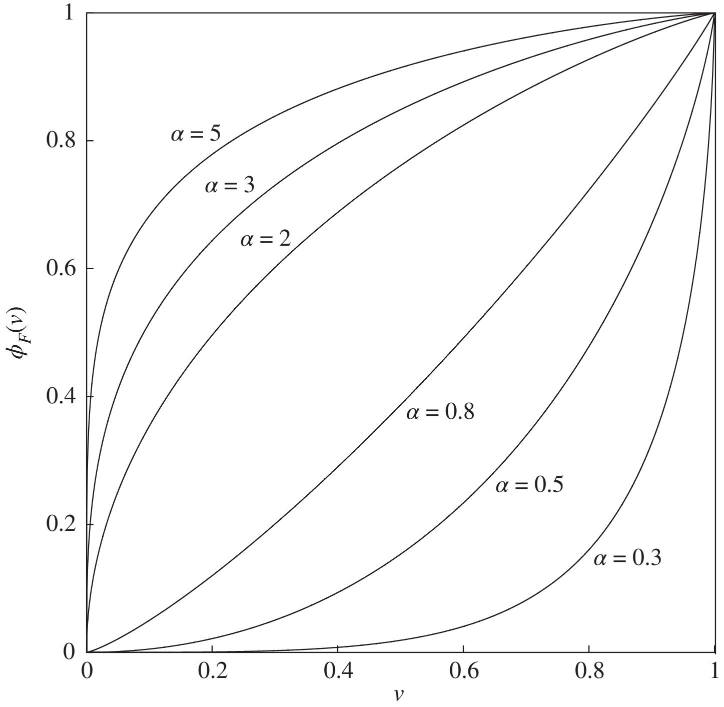


Figure 14.22 Scaled TTT transforms of the Weibull distribution for some selected values of α.

Proof:
Because

$$\frac{d}{dv}H_F^{-1}(v) = \frac{d}{dv}\int_0^{F^{-1}(v)}[1 - F(u)]\,du$$
$$= (1 - F[F^{-1}(v)])\frac{d}{dv}F^{-1}(v) = (1 - v)\frac{1}{f[F^{-1}(v)]},$$

then

$$\frac{d}{dv}H_F^{-1}(v)|_{v=F(t)} = [1 - F(t)]\frac{1}{f(t)} = \frac{1}{z(t)}.$$

From (14.67) we obtain

(2) If $F(t)$ is a continuous life distribution, strictly increasing for $F^{-1}(0) = 0 < t < F^{-1}(1)$, then
   (a) $F \sim \text{IFR} \quad \Longleftrightarrow \quad H_F^{-1}(v)$ concave; $0 \le v \le 1$
   (b) $F \sim \text{DFR} \quad \Longleftrightarrow \quad H_F^{-1}(v)$ convex; $0 \le v \le 1$

The arguments, used to prove properties 1 and 2 are completely analogous. We therefore prove only property 1.

Proof:

$$\begin{aligned}
F \sim \text{IFR} &\iff z(t) \text{ is nondecreasing in } t\\
&\iff \frac{1}{z(t)} \text{ is nonincreasing in } t\\
&\iff \frac{d}{dv} H_F^{-1}(v)|_{v=F(t)} \text{ is nonincreasing in } t\\
&\iff \frac{d}{dv} H_F^{-1}(v) \text{ is nonincreasing in } v\\
&\text{because } F(t) \text{ is strictly increasing}\\
&\iff H_F^{-1}(v) \text{ is concave, } \ 0 \le v \le 1.
\end{aligned}$$

To estimate the scaled TTT transform of $F(t)$ for different v values on the basis of the observed lifetimes, it is natural to use the estimator

$$\frac{\int_0^{F_n^{-1}(v)} [1 - F_n(u)]\, du}{\int_0^{F_n^{-1}(1)} [1 - F_n(u)]\, du} \quad \text{for} \quad v = \frac{i}{n}, \qquad i = 1, 2, \dots, n. \tag{14.68}$$

Introducing the notation

$$H_n^{-1}(v) = \int_0^{F_n^{-1}(v)} [1 - F_n(u)]\, du \quad \text{for} \quad v = \frac{i}{n}, \qquad i = 1, 2, \dots, n, \tag{14.69}$$

this estimator can be written as

$$\frac{H_n^{-1}(v)}{H_n^{-1}(1)} \quad \text{for} \quad v = \frac{i}{n}, \qquad i = 1, 2, \dots, n. \tag{14.70}$$

By comparing (14.70) with (14.64), it seems natural to call $H_n^{-1}(v)/H_n^{-1}(1)$ the empirical, scaled TTT transform of the distribution $F(t)$.

The following result is useful when we wish to exploit the TTT plot to provide information about the life distribution $F(t)$:

(3) If $F(t)$ is a continuous life distribution function, strictly increasing for $F^{-1}(0) = 0 < t < F^{-1}(1)$, then

$$\frac{H_n^{-1}\left(\frac{i}{n}\right)}{H_n^{-1}(1)} = \frac{\mathcal{T}(t_{(i)})}{\mathcal{T}(t_{(n)})} \quad \text{for} \quad i = 1, 2, \dots, n, \tag{14.71}$$

where $\mathcal{T}(t_{(i)})$, as before, is the TTT at time $t_{(i)}$.

Proof:
According to (14.69), for $i = 1, 2, \ldots, n$,

$$
\begin{aligned}
H_n^{-1}\left(\frac{i}{n}\right) &= \int_0^{F_n^{-1}(\frac{i}{n})} [1 - F_n(u)]\,du \\
&= \int_0^{T_{(i)}} [1 - F_n(u)]\,du = \frac{1}{n}\mathcal{T}(T_{(i)}),
\end{aligned}
$$

where as

$$
\begin{aligned}
H_n^{-1}(1) &= \int_0^{F_n^{-1}(1)} [1 - F_n(u)]\,du \\
&= \int_0^{\infty} [1 - F_n(u)]\,du = \frac{1}{n}\mathcal{T}(t_{(n)}) = \frac{1}{n}\sum_{i=1}^{n} t_i.
\end{aligned}
$$

By introducing these results in (14.70), we get (14.71).

Therefore, the scaled TTT at time $t_{(i)}$ seems to be a natural estimate of the scaled TTT transform of $F(t)$ for $v = i/n$, for $i = 1, 2, \ldots, n$. One way of obtaining an estimate for the scaled TTT transform for $(i-1)/n < v < i/n$, is by applying linear interpolation between the estimate for $v = (i-1)/n$ and $v = i/n$. In the following we use this procedure.

Now suppose that we have access to a survival dataset. We first determine $\mathcal{T}(t_{(i)})/\mathcal{T}(t_{(n)})$ for $i = 1, 2, \ldots, n$ as we did in Example 14.13, plot the points $[i/n, \mathcal{T}(t_{(i)})/\mathcal{T}(t_{(n)})]$ and join pairs of neighboring points with straight lines. The curve obtained is an estimate for $H_F^{-1}(v)/H_F^{-1}(1) = \frac{1}{\mu}H_F^{-1}(v)$, for $0 \le v \le 1$.

We may now assess the shape of the curve (the estimate for $H_F^{-1}(v)$) in the light of the result in (14.67) and its proof, and in this way obtain information about the underlying distribution $F(t)$.

A plot, such as the one shown in Figure 14.23a, shows that $H_F^{-1}(v)$ is concave. The plot therefore indicates that the corresponding life distribution $F(t)$ is IFR.

Using the same type of argument, the plot in Figure 14.23b shows that $H_F^{-1}(v)$ is convex, so that the corresponding life distribution $F(t)$ is DFR. Similarly, the plot in Figure 14.23c indicates that $H_F^{-1}(v)$ "is first convex" and "thereafter concave." In other words, the failure rate of the corresponding lifetime distribution has a bathtub shape.

The TTT plot obtained in Example 14.13, therefore indicates that these data originate from a life distribution with bathtub shaped failure rate.

Example 14.16 (Ball bearing failures)
Lieblein and Zelen (1956) provide the numbers of millions of revolutions to failure for each of 23 ball bearings. Below, the original data are put in numerical order for convenience.

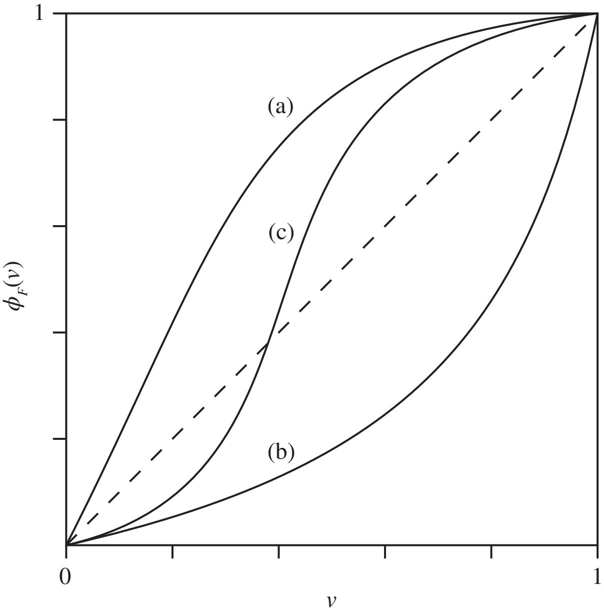


Figure 14.23 TTT plots indicating (a) increasing failure rate (IFR), (b) decreasing failure rate (DFR), and (c) bathtub-shaped failure rate.

17.88	28.92	33.00	41.52	42.12	45.60	48.40
51.84	51.96	54.12	55.56	67.80	68.64	68.64
68.88	84.12	93.12	98.64	105.12	105.84	127.92
128.04	173.40					

The TTT plot of the ball bearing data is presented in Figure 14.24. The TTT plot indicates an IFR. We may try to fit a Weibull distribution to the data. The Weibull parameters α and λ are estimated to be $\widehat{\alpha} = 2.10$ and $\widehat{\lambda} = 1.22 \times 10^{-2}$. The TTT transform of the Weibull distribution with these parameters is plotted as an overlay curve to the TTT plot in Figure 14.24. □

TTT Plotting with R

The scaled TTT-plot is available in the package `AdequacyModel`. We illustrate its use by the data from Example 14.16. A simple script for Figure 14.24 is:

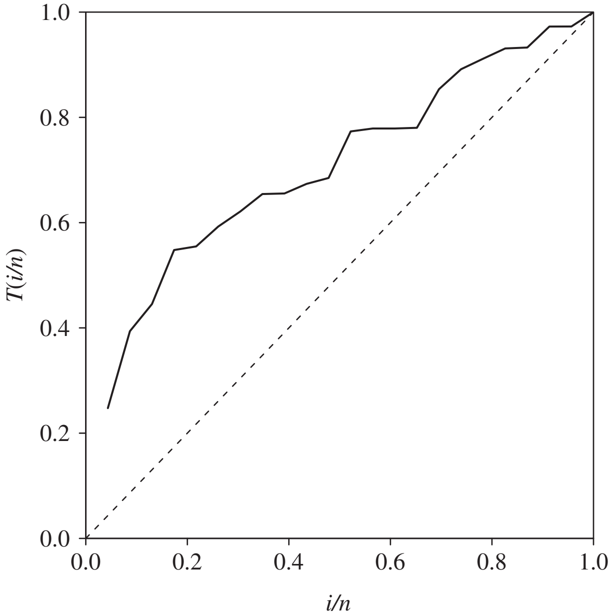


Figure 14.24 TTT plot of the ball bearing data in Example 11.11 together with an overlay curve of the TTT transform of the Weibull distribution with shape parameter $\alpha = 2.10$.

```
library(AdequacyModel)
# Enter the dataset
data <- c(17.88,28.92,33.00,41.52,42.12,45.60,48.40,
       51.84,51.96,54.12,55.56,67.80,68.64,68.64,
       68.88,84.12,93.12,98.64,105.12,105.85,127.92,
       128.04,173.40)
# Make the TTT plot
TTT(data,lwd=1.5,grid=F,lty=3)
```

If you want to establish the scaled TTT-transform for a particular distribution, say a 2-parameter Weibull distribution with shape parameter $\alpha = 3$, this can be obtained by a similar script where the data is a random sample from this distribution. To get a smooth curve, we need a rather high number of simulated values. A script to obtain the TTT-transform is

```
library(AdequacyModel)
# Generate a random sample from a Weibull distribution
data <- rweibull(8000,3,scale=1)
# Make the TTT transform
TTT(data, lwd=1.5,grid=F,lty=3)
```

Example 14.17 (Age replacement)

A well-known application of the TTT transform and the TTT plot is the age replacement problem that is discussed in Section 12.3.1. Here an item is replaced at a cost $c + k$ at *failure* or at a cost c at a *planned replacement* when the item has reached a certain age t_0.

The average replacement cost per time unit of this policy was found to be

$$C(t_0) = \frac{c + kF(t_0)}{\int_0^{t_0} [1 - F(t)]\, dt}. \tag{14.72}$$

The objective is now to determine the value of t_0 that minimizes $C(t_0)$. If the distribution function $F(t)$ and all its parameters are known, it is a straightforward task to determine the optimal value of t_0. One way to solve this problem is to apply the TTT transform.

By introducing the TTT transform (14.62) as

$$C(t_0) = \frac{c + kF(t_0)}{H_F^{-1}[F(t_0)]} = \frac{1}{H_F^{-1}(1)} \frac{c + kF(t_0)}{\varphi_F[F(t_0)]},$$

where $H_F^{-1}(1)$ is the MTTF of the item, and $\varphi_F(v) = H_F^{-1}(v)/H_F^{-1}(1)$ is the scaled TTT transform of the distribution function $F(t)$.

The optimal value of t_0 may be determined by first finding the value $v_0 = F(t_0)$ that minimizes

$$C_1(v_0) = \frac{c + kv_0}{\varphi_F(v_0)},$$

and thereafter determine t_0 such that $v_0 = F(t_0)$. The minimizing value of v_0 may be found by setting the derivative of $C_1(v_0)$ with respect to v_0 equal to zero, and solve the equation for v_0.

$$\frac{d}{dv_0} C_1(v_0) = \frac{\varphi_F(v_0)\, k - \varphi'_F(v_0)(c + kv_0)}{\varphi_F(v_0)^2} = 0.$$

This implies that

$$\varphi'_F(v_0) = \frac{\varphi_F(v_0)}{c/k + v_0}. \tag{14.73}$$

The optimal value of v_0, and hence t_0, may now be determined by the following simple graphical method.

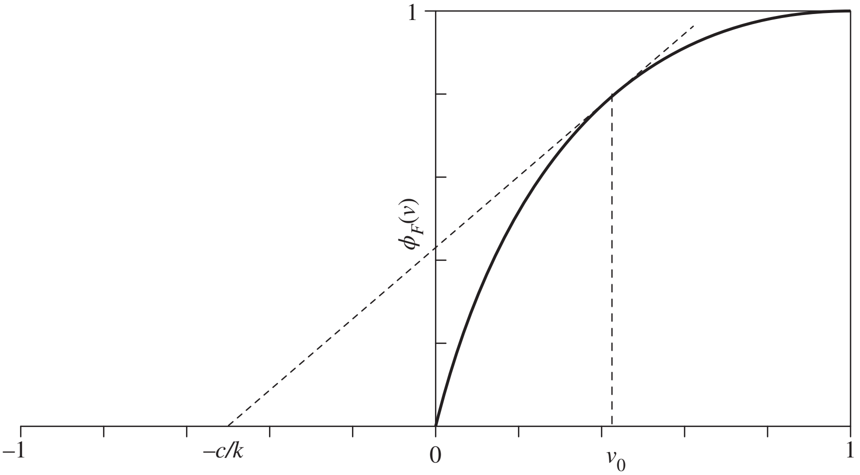


Figure 14.25 Determination of the optimal replacement age from the scaled TTT transform.

(1) Draw the scaled TTT transform in a 1×1 –coordinate system.
(2) Identify the point $(-c/k, 0)$ on the abscissa axis.
(3) Draw a tangent from $(-c/k, 0)$ to the TTT transform.

The optimal value of v_0 can now be read as the abscissa of the point where the tangent touches the TTT transform. If $v_0 = 1$, then $t_0 = \infty$, and no preventive replacements should be performed. The procedure is shown in Figure 14.25.

When a set of times-to-failure of the actual type of item has been recorded, we may use this dataset to obtain the empirical, scaled TTT transform of the underlying distribution function $F(t)$, and draw a TTT plot. The optimal replacement age t_0 may now be determined by the same procedure as described above. This is shown in Figure 14.26. The procedure is further discussed, for example, by Bergman and Klefsjö (1982,1984). □

14.7.2 Total-Time-on-Test Plot for Censored Datasets

When the dataset is incomplete with random censoring (type IV), we may argue as follows to obtain a TTT plot: The TTT transform, as defined in (14.62), is valid for a wide range of distribution functions $F(t)$, also for step functions. Instead of estimating the TTT transform $H_F^{-1}(t)$ by introducing the empirical distribution function $F_n(t)$ as we did in (14.69), we could estimate $F(t)$ by $[1 - \widehat{R}(t)]$, where $\widehat{R}(t)$ is the Kaplan–Meier estimator of $R(t)$.

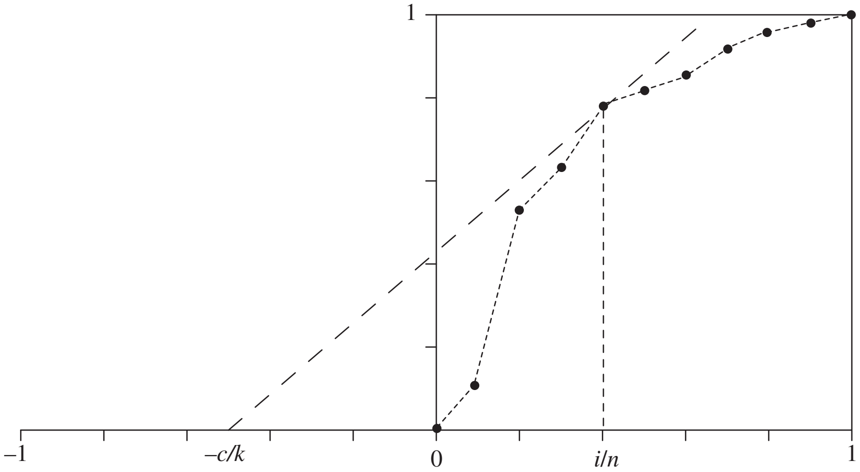


Figure 14.26 Determination of the optimal replacement age from a TTT plot.

Technically, the plot is obtained as follows: Let $t_{(1)}, t_{(2)}, \dots, t_{(k)}$ denote the k ordered failure times among $t_1, t_2, \dots, t_n$ and let

$$v_{(i)} = 1 - \widehat{R}(t_{(i)}) \quad \text{for} \qquad i = 1, 2, \dots, k.$$

Define

$$\widehat{H}^{-1}(v_{(i)}) = \int_0^{t_{(i)}} \widehat{R}(u)\, du = \sum_{j=1}^{i-1} (t_{(j+1)} - t_{(j)}) \widehat{R}(t_{(j)}),$$

where $t(0) = 0$.

The TTT plot is now obtained by plotting the points

$$\left(\frac{v_{(i)}}{v_{(k)}}, \frac{\widehat{H}^{-1}(v_{(i)})}{\widehat{H}^{-1}(v_{(k)})} \right) \quad \text{for} \qquad 1 = 1, 2, \dots, k.$$

Observe that when $k = n$, that is, when the dataset is complete, then

$$v_{(i)} = \frac{i}{n},$$
$$\widehat{H}^{-1}(v_{(i)}) = \mathcal{T}(t_{(i)}),$$

and we get the same TTT plot as we got for complete datasets.

14.7.3 A Brief Comparison

Sections 14.5–14.7 present three nonparametric estimation and plotting techniques that may be applied to both complete and censored data. (The empirical survivor function is equal to the Kaplan–Meier estimate when the dataset is complete, and is therefore considered as a special case of the Kaplan–Meier approach.) The estimates obtained by using the Kaplan–Meier, and the Nelson–Aalen approaches are rather similar, so it is not important which of these is chosen. The nature of the estimate based on TTT transform is different from the other two estimates and may provide supplementary information.

The plots may also be used as a basis for selection of an adequate parametric distribution $F(t)$. In this respect, the three plots provide somewhat different information. The Kaplan–Meier plot is very sensitive to variations in the early and middle phases of an item's lifetime, but is not very sensitive in the right tail of the distribution. The Nelson–Aalen plot is not at all sensitive in the early part of the life distribution, because the plot is "forced" to start in $(0, 0)$. The TTT plot is very sensitive in the middle phase of the life distribution, but less sensitive in the early phase and in the right tail, because the plot is "forced" to start in $(0, 0)$ and end up in $(1, 1)$. To get adequate information about the whole distribution, all the three plots should be studied.

14.8 Survival Analysis with Covariates

The reliability of items is often found to be influenced by one or more *covariates*. Covariates and various models applying covariates were introduced in Section 5.5. This section sheds some light on how to analyze data with different covariate levels. This is a huge and complicated area, so we only scratch the surface of this topic.

We assume that all covariates are measurable, either on a continuous scale, a discrete scale, or simply as "yes" or "no." We further assume that all covariates remain constant during the data collection exercise.

14.8.1 Proportional Hazards Model

By a proportional hazards (PH) model the failure rate function $z(t)$ is modified by a factor $g(\boldsymbol{s})$, where $\boldsymbol{s}$ is the covariate vector. The term "hazard" is here used with the same meaning as failure rate. We could therefore talk about proportional failure rates instead of PH, but PH is the standard term used in most other application areas, such as biostatistics and medical research.

The PH model assumes that the failure rate function related to a specific covariate vector $\boldsymbol{s}$ may be written as

$$z(t \mid \boldsymbol{s}) = z_0(t)\, g(\boldsymbol{s}). \tag{14.74}$$

The failure rate function $z(t \mid \boldsymbol{s})$ is seen to be the product of two factors:

(1) A time-dependent factor $z_0(t)$, which is called the *baseline failure rate* and does not depend on $\boldsymbol{s}$. The baseline failure rate is usually not specified in the PH model.

(2) A *proportionality factor* $g(\boldsymbol{s})$, which is a function of the covariate vector $\boldsymbol{s}$, and not of time t.

Hazard Ratio

We may compare the effects of two covariate vectors $\boldsymbol{s}_1$ and $\boldsymbol{s}_0$ by the ratio:

$$\text{HR}(\boldsymbol{s}_1, \boldsymbol{s}_0) = \frac{z(t \mid \boldsymbol{s}_1)}{z(t \mid \boldsymbol{s}_0)} = \frac{g(\boldsymbol{s}_1)}{g(\boldsymbol{s}_0)}. \tag{14.75}$$

This expression is called the *hazard ratio* (HR) for the covariate vectors $\boldsymbol{s}_1$ and $\boldsymbol{s}_0$. The covariate vector $\boldsymbol{s}_0$ often refers to a basic and known application of the item, called the *baseline application*, whereas the covariate vector $\boldsymbol{s}_1$ refers to the use of a similar item in a new environment. The hazard ratio shows that the two failure rate functions are proportional for any value of t. This proportionality is the reason for calling the model a PH model. The factor of interest is how large $g(\boldsymbol{s}_1)$ is compared to $g(\boldsymbol{s}_0)$ and not the value of each of them. Therefore, we often set $g(\boldsymbol{s}_0) = 1$, such that $g(\boldsymbol{s}_1) = \text{HR}(\boldsymbol{s}_1, \boldsymbol{s}_0)$.

In cases where $g(\boldsymbol{s}_0) = 1$, and we study a single alternative covariate vector, this vector is usually denoted $\boldsymbol{s}$ (i.e. without an index) and the hazard ratio is $\text{HR}(\boldsymbol{s}) = g(\boldsymbol{s})$.

The effect of the covariate vector $\boldsymbol{s}$ is, therefore, determined by $g(\boldsymbol{s})$, which scales the baseline failure rate function $z_0(t)$. Figure 14.27 shows a baseline failure rate function $z_0(t)$ for a Weibull distribution with shape parameter $\alpha = 1.65$ (fully drawn line) together with the failure rate function (dotted line) for an item with covariate vector $\boldsymbol{s}$ and hazard ratio $g(\boldsymbol{s}) = 2$ based on a PH model. The failure rate function for $\boldsymbol{s}$ is obtained by multiplying $z_0(t)$ with HR = 2 for each point of time t.

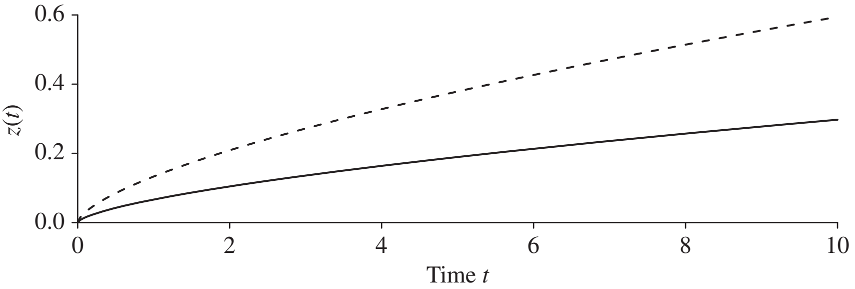


Figure 14.27 Failure rate function for the PH model. The baseline failure rate function (fully drawn line) and for another condition with hazard ratio (HR)= 2 (dotted line). The baseline is a Weibull distribution with shape parameter $\alpha = 1.65$.

Cumulative Failure Rate

In the PH model, the cumulative failure rate $Z(t) = \int_0^t z(u)\,du$ is

$$Z(t \mid \boldsymbol{s}) = Z_0(t)g(\boldsymbol{s}). \tag{14.76}$$

Survivor Function

Let $R_0(t)$ be the survivor function for the baseline application. The survivor function for a new application with covariate vector $\boldsymbol{s}$ is from (14.74)

$$R(t \mid \boldsymbol{s}) = \exp[-Z(t \mid \boldsymbol{s})] = \exp[-Z_0(t)g(\boldsymbol{s})] = [R_0(t)]^{g(\boldsymbol{s})}. \tag{14.77}$$

This result implies that if we know the survivor function in the baseline application and if we are able to determine the hazard ratio, $g(\boldsymbol{s})$, it is easy to find the survivor function – and all the related reliability measures – in the new environment $\boldsymbol{s}$.

Example 14.18 (Exponential distribution)
Consider at item with constant failure rate. During normal operation in a baseline environment, the failure rate is λ_0. The assumption of constant failure rate is considered realistic also for an alternative environment with covariate vector $\boldsymbol{s}$. A simple model for describing the failure rate in this environment is

$$\lambda(\boldsymbol{s}) = \left(\sum_{i=1}^{m} k_1 s_i \right) \lambda_0,$$

where k_i is a constant that determines the effect of s_i on the failure rate, for $i = 1, 2, \dots, m$. To comply with our knowledge about the effect of the various influences, the covariates used may be transformed values of the physical variables. The square of the voltage may, for example, be used as a covariate. □

Example 14.19 (The MIL-HDBK-217 prediction method)
The MIL-HDBK-217F (1995) has for a long time been the state-of-the-art approach for predicting the constant failure rate of an electronic item that is used under non-baseline conditions. Let λ_0 be the constant failure rate when the item is used under baseline conditions. For these conditions, λ_0 can be estimated from data obtained from laboratory testing or from field data. The MIL-HDBK-217F suggests that the failure rate λ in the nonreference conditions is determined as follows:

$$\lambda = \lambda_0 \cdot \pi_S \cdot \pi_T \cdot \pi_E \cdot \pi_Q \cdot \pi_A, \tag{14.78}$$

where

π_S is the stress factor.
π_T is the temperature factor.
π_E is the environment factor.

π_Q is the quality factor.
π_A is the adjustment factor.

These factors may be found in the handbook when we know the conditions the item is used in. The MIL-HDBK-217F therefore applies a simple PH approach. The MIL-HDBK-217F is discussed further in Chapter 16. □

14.8.2 Cox Models

The Cox model was introduced by the British statistician Sir David Roxbee Cox in his famous paper "Regression models and life tables" Cox (1972). The Cox model is a special case of a PH model, where the failure rate function is written as

$$z(t \mid \boldsymbol{s}) = z_0(t)\, e^{\boldsymbol{\beta s}}. \tag{14.79}$$

The hazard ratio $g(\boldsymbol{s})$ of this model is

$$g(\boldsymbol{s}) = e^{\boldsymbol{\beta s}} = \exp\left(\sum_{i=1}^{k} \beta_i s_i\right),$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_k)$ is a vector of unknown *parameters* that need to be estimated from observed data.

Consider two different stress levels; a baseline application with $\boldsymbol{s}_0$ and a new application with $\boldsymbol{s}$. It is common practice to set $\boldsymbol{s}_0 = \boldsymbol{0}$ for the baseline application and to measure the covariates $\boldsymbol{s}$ as the difference from the baseline application. It is further common to scale the function $g(\cdot)$ such that $g(\boldsymbol{s}_0) = 1$. The hazard ratio of this Cox model is

$$\frac{z(t \mid \boldsymbol{s})}{z(t \mid \boldsymbol{0})} = \exp(\boldsymbol{\beta s}) = \exp\left(\sum_{j=1}^{k} \beta_j s_j\right).$$

For the Cox model, the log-failure rate function is a linear function

$$\log z(t \mid \boldsymbol{s}) = \log z_0(t) + \beta_1 s_1 + \beta_2 s_2 + \cdots + \log \beta_k s_k. \tag{14.80}$$

This indicates that (14.80) may be a suitable basis for some sort of regression analysis.

The Cox model is said to be a *semiparametric* model. It is not a parametric model because the baseline failure rate $z_0(t)$ is unspecified, and it is not nonparametric because it is assumed how the failure rate function varies with the value of the covariates. If we make special assumptions about the baseline failure rate function $z_0(t)$, the Cox model becomes a parametric model. The baseline distribution may, for example, be assumed to be an exponential or a Weibull distribution. The advantage of the Cox model is that such assumptions can be avoided. Even though $z_0(t)$ is unspecified, our objective is to estimate the parameters $\boldsymbol{\beta}$. One of the biggest advantages of the Cox model is that we can estimate the parameters $\boldsymbol{\beta}$ that reflect the effects of the covariates without having to make any assumptions about the form of $z_0(t)$.

14.8.3 Estimating the Parameters of the Cox Model

A thorough introduction of the theory required to estimate the parameters ($\boldsymbol{\beta}$) of the Cox model would involve several new concepts and is considered to be outside the scope of this book. The theory described by Cox (1972) is elaborated in several books (e.g. see Cox and Oakes 1984, Ansell and Phillips 1994, Crowder et al. 1991, Kalbfleisch and Prentice 1980, Lawless 1982). Theoretical introductions and surveys may also be found in a high number of presentations and lecture notes that are available on the Internet.

Here, we suffice with a simple introduction, where we highlight some of the main concepts. We start with a right-censored dataset of survival times $\boldsymbol{t} = (t_1, t_2, \dots, t_n)$ from n independent and identical items used in different environments. All the survival times are measured from time $t = 0$. As before, we use the indicator δ_i to tell whether the survival time ended with a failure ($\delta_i = 1$) or with a censoring ($\delta_i = 0$), for $i = 1, 2, \dots, n$.

From (14.21), the likelihood function may be written as

$$L(\boldsymbol{\beta} \mid \text{data}) = \prod_{i=1}^{n} [z(t_i \mid \boldsymbol{\beta}, \boldsymbol{s}_i)]^{\delta_i} R(t_i \mid \boldsymbol{\beta}, \boldsymbol{s}_i),$$

where $z(\cdot)$ and $R(\cdot)$ are seen as functions of $\boldsymbol{\beta}$ and "data" includes all the data available in the dataset, including t_i, δ_i, and $\boldsymbol{s}_i$ for all the items. For the Cox model, the likelihood function may be written as

$$L(\boldsymbol{\beta} \mid \text{data}) = \prod_{i=1}^{n} z_0(t_i)[\exp(\boldsymbol{\beta}\boldsymbol{s}_i)]^{\delta_i}[R_0(t_i)]^{\exp(\boldsymbol{\beta}\boldsymbol{s}_i)}.$$

The corresponding log-likelihood function is

$$\ell(\boldsymbol{\beta} \mid \text{data}) = \sum_{i=1}^{n} \log z_0(t_i) + \delta_i \boldsymbol{\beta}\boldsymbol{s}_i + \exp(\boldsymbol{\beta}\boldsymbol{s}_i) \log R_0(t_i).$$

It is not possible to find the $\boldsymbol{\beta}$ that maximizes this log-likelihood function unless we specify the baseline failure rate function $z_0(t)$. A detailed discussion of this problem is given in Cox and Oakes (1984, chapter 7). Instead, Cox (1972) introduced a *partial likelihood function* that does not depend on $z_0(t)$. This function uses the at-risk-set RS(t) at time t, that is, the set of all items that are functioning and exposed to failure just before time t. Items that have failed or have been censored before time t are not members of RS(t). In this simplified introduction, we assume that there are no ties in the dataset.

Consider a dataset with n distinct survival times $t_1, t_2, \dots, t_n$. To each survival time t_i is connected an indicator δ_i and a covariate vector $\boldsymbol{s}_i$, for $i = 1, 2, \dots, n$. Each covariate vector may be regarded as an observation of a general covariate vector $\boldsymbol{s}$. This means that the same covariates are measured for each and every survival time.

Next, the survival times are ordered, such that $t_1 < t_2 < \cdots < t_n$. To establish the partial likelihood function, Cox (1972) starts by considering the conditional probability that a specific item, say i [$\in \text{RS}(t_i)$] fails at time t_i given that one individual item from the at-risk-set $\text{RS}(t_i)$ fails at time t_i.[9] If the dataset were *complete*, this probability would be

$$L^p(\boldsymbol{\beta} \mid t_i, \boldsymbol{s}_i) = \frac{z(t_i \mid \boldsymbol{\beta}\boldsymbol{s}_i)}{\sum_{j \in \text{RS}(t_i)} z(t_i \mid \boldsymbol{\beta}\boldsymbol{s}_j)},$$

and is the contribution to the partial likelihood [$L^p(\cdot)$] from survival time t_i. The arguments used to arrive at the above result may be summarized as follows:

$$\begin{aligned}
&\Pr(\text{Item } i \text{ fails at time } t_i \mid \text{One item from RS}(t_i) \text{ fails at time } t_i)\\
&\quad= \frac{\Pr(\text{Item } i \text{ fails at } t_i)}{\Pr(\text{One failure in RS}(t_i) \text{ at } t_i)}\\
&\quad= \frac{\Pr(\text{Item } i \text{ fails at } t_i)}{\Pr\left(\sum_{j \in \text{RS}(t_i)} \text{Item } j \in \text{RS}(t_i) \text{ fails at } t_i, t_i + \Delta t\right)}\\
&\quad\approx \frac{\Pr(\text{Item } i \text{ fails in } (t_i, t_i + \Delta t))/\Delta t}{\Pr\left(\sum_{j \in \text{RS}(t_i)} \text{Item } j \in \text{RS}(t_i) \text{ fails in } t_i, t_i + \Delta t\right)/\Delta t}\\
&\quad\approx \frac{\lim_{\Delta t \to 0} \Pr(\text{Item } i \text{ fails in } (t_i, t_i + \Delta t))/\Delta t}{\lim_{\Delta t \to 0} \Pr\left(\sum_{j \in \text{RS}(t_i)} \text{Item } j \in \text{RS}(t_i) \text{ fails in } t_i, t_i + \Delta t\right)/\Delta t}\\
&\quad= \frac{z(t_i \mid \boldsymbol{\beta}\boldsymbol{s}_i)}{\sum_{j \in \text{RS}(t_i)} z(t_i \mid \boldsymbol{\beta}\boldsymbol{s}_j)}.
\end{aligned}$$

To simplify the notation, we introduce

$$\psi_i = \exp(\boldsymbol{\beta}\boldsymbol{s}_i) \quad \text{for} \qquad i = 1, 2, \dots, n,$$

which is the factor we must multiply the baseline failure rate $z_0(t)$ with to obtain the failure rate for the covariate vector $\boldsymbol{s}_i$, that is, $z(t_i \mid \boldsymbol{\beta}, \boldsymbol{s}_i) = \psi_i z_0(t)$. The contribution to the total partial likelihood function from failure time t_i can hence be

[9] Any item in $\text{RS}(t_i)$ would do. We assume that item i corresponds to the ordered survival time t_i, so it is obviously a member of $\text{RS}(t_i)$. We focus on item i to simplify the notation.

written as

$$L^p(\boldsymbol{\beta} \mid t_i, \boldsymbol{s}_i) = \frac{\psi_i}{\sum_{j \in \mathrm{RS}(t_i)} \psi_j}. \tag{14.81}$$

The total partial likelihood for the *complete* dataset is then

$$L^p(\boldsymbol{\beta} \mid \text{data}) = \prod_{i=1}^{n} \frac{\psi_i}{\sum_{j \in \mathrm{RS}(t_i)} \psi_j}.$$

For a right *censored* dataset, the partial likelihood can be shown to be

$$L^p(\boldsymbol{\beta} \mid \text{data}) = \prod_{i=1}^{n} \left[\frac{\psi_i}{\sum_{j \in \mathrm{RS}(t_i)} \psi_j} \right]^{\delta_i}, \tag{14.82}$$

where censored times are excluded by the indicator $\delta_i = 0$ (remember $x^0 = 1$). The partial likelihood function is obtained by multiplying the contributions (14.80) from the actual failure times, but the censoring times are still important because they enter into the at-risk-sets RS(t).

There are at least two reasons why $L^p(\boldsymbol{\beta} \mid \text{data})$ is called a *partial* likelihood:

- It is not a complete likelihood function for all parameters of the density function (because the baseline failure rate function is not covered).
- All the data in the dataset is not used because the actual survival times play no part in (14.81), but only their ranking, i.e. when they enter into the at-risk set.

More thorough treatments may be found in Cox and Oakes (1984), Lawless (1982), and Kalbfleisch and Prentice (1980).

When there are many ties in the dataset, computation of maximum partial-likelihood estimates is still possible but may become time-consuming. Of this reason, the partial likelihood function is often approximated. Two commonly employed approximations are due to Norman E. Breslow and to Bradley Efron. Both approximations are available in the R `survival` package.

The procedures to find estimates for the various parameters are rather technical and are not presented in the current book. Readers who plan to use the Cox model on a practical dataset are advised to consult a more specialized book and to carefully read the documentation of the relevant R packages.

Cox Model Analysis with R

The Cox model is available in R by the function `coxph` in the `survival` package. Related aspects are also treated in several other R packages. Among these are `simPH, coxme, Coxnet, coxphw`, and several more.

We suggest that you start by learning the function `coxph` in the `survival` package. You have many options when using this package, and it is therefore important that you read carefully the package documentation.

Additional theory and several worked examples with R may be found in Moore (2016) and Fox and Weisberg (2019).

14.9 Problems

14.1 Assume that you have determined the lifetimes for a total of 12 identical items and obtained the following results (given in hours): 10.2, 89.6, 54.0, 96.0, 23.3, 30.4, 41.2, 0.8, 73.2, 3.6, 28.0, 31.6

The dataset can be downloaded from the `book companion site`.

(a) Find the sample mean and the sample standard deviation for the dataset. Can you draw any conclusions about the underlying distribution $F(t)$ by comparing the sample mean and the sample standard deviation?

(b) Construct the empirical survivor function for the dataset.

(c) Plot the data on a Weibull paper.[10] What conclusions can you draw from the plot?

(d) Construct the TTT plot for the dataset. What conclusion can you draw from the TTT plot about the corresponding life distribution?

14.2 Failure time data from a compressor were discussed in Example 10.2. All compressor failures at a certain process plant in the time period from 1968 until 1989 have been recorded. In this period, a total of 90 critical failures occurred. In this context, a critical failure is defined to be a failure causing compressor downtime. The compressor is very important for the operation of the process plant, and every effort is taken to restart a failed compressor as soon as possible. The 90 repair times (in hours) are presented chronologically in Table 14.6. The repair time associated with the first failure was 1.25 hours, the second repair time was 135.00 hours, and so on. The dataset can be downloaded from the `book companion site`.

(a) Plot the repair times in chronological order to check whether or not there is a trend in the repair times. Is there any reason to claim that the repair times increase with the age of the compressor?

(b) Assume now that the repair times are independent and identically distributed. Construct the empirical distribution function for the repair times

(c) Plot the repair times on a lognormal plotting paper. Is it reason to believe that the repair times are lognormally distributed?

14.3 Consider the set of material strength data presented by Crowder et al. (1991, p. 46) and given in Table 14.7. An experiment has been carried

[10] Weibull paper may be downloaded from https://www.weibull.com/GPaper/ or you can use the R package `WeibullR`.

Table 14.6 Dataset for Problem 14.2.

1.25	135.00	0.08	5.33	154.00	0.50	1.25	2.50	15.00
6.00	4.50	32.50	9.50	0.25	81.00	12.00	0.25	1.66
5.00	7.00	39.00	106.00	6.00	5.00	17.00	5.00	2.00
2.00	0.33	0.17	0.50	18.00	2.50	0.33	0.50	2.00
0.33	4.00	20.00	6.00	6.30	15.00	23.00	4.00	5.00
28.00	16.00	11.50	0.42	38.33	10.50	9.50	8.50	17.00
34.00	0.17	0.83	0.75	1.00	0.25	0.25	2.25	13.50
0.50	0.25	0.17	1.75	0.50	1.00	2.00	2.00	38.00
0.33	2.00	40.50	4.28	1.62	1.33	3.00	5.00	120.00
0.50	3.00	3.00	11.58	8.50	13.50	29.50	29.50	112.00

Table 14.7 Dataset for Problem 14.3.

26.8^{*}	29.6^{*}	33.4^{*}	35.0^{*}	36.3	40.0^{*}	41.7	41.9^{*}	42.5^{*}
43.9	49.9	50.1	50.8	51.9	52.1	52.3	52.3	52.4
52.6	52.7	53.1	53.6	53.6	53.9	53.9	54.1	54.6
54.8	54.8	55.1	55.4	55.9	56.0	56.1	56.5	56.9
57.1	57.1	57.3	57.7	57.8	58.1	58.9	59.0	59.1
59.6	60.4	60.7						

*Censored data points.

out to gain information about the strength of a certain type of braided cord. A total of 48 pieces of cord were investigated. Seven cords were damaged during the experiment, implying right-censored strength values. The dataset can be downloaded from the `book companion site`.

(a) Establish a Kaplan–Meier plot of the material strength data.
(b) Establish a TTT plot of the material strength data.
(c) Discuss the effect of this type of censoring.
(d) Describe the form of the related failure rate function.

14.4 Establish a graph paper such that the Nelson–Aalen plot of Weibull distributed life data is close to a straight line. Describe how the Weibull parameters α and λ can be estimated from the plot.

14.5 The Pareto distribution has cumulative distribution function $F(x) = \Pr(X \leq x) = 1 - x^{-\theta}$ for $x > 1$. Let $x_1, x_2, \dots, x_n$ be n independent observations of X.

(a) Find the method of moments estimation (MME) estimator for θ.

(b) Find the mean and the standard deviation of this estimator.

14.6 Let $X_1, X_2, \dots, X_n$ be independent and identically distributed variables with uniform distribution unif$(0, \theta)$. Assume that $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$ has been observed.

(a) Find the likelihood function $L(\theta \mid \boldsymbol{x})$.

(b) Find the MLE for θ and derive its mean value.

(c) Find an unbiased estimator for θ.

14.7 Let $X_1, X_2, \dots, X_n$ independent and identically distributed Po(λ), where λ is unknown.

(a) Find an MLE for $e^{-\lambda}$.

(b) Find an unbiased estimator for $e^{-\lambda}$.

14.8 Consider a homogeneous Poisson process (HPP) with rate λ. Let $N(t)$ be the number of failures (events) in a time interval of length t. $N(t)$ is hence Poisson distributed with parameter λt. Assume that the process is observed in a time interval of length $t = 2$ years. In this time period, a total of seven failures have been observed.

(a) Find an estimate for λ

(b) Determine a 90% confidence interval for λ

14.9 Let $X \sim$ Po(λ).

(a) Determine an exact 90% confidence interval for λ when X is observed and found equal to 6. For comparison, also determine an approximate 90% confidence interval for λ, using the approximation of the Poisson distribution to $\mathcal{N}(\lambda, \lambda)$

(b) Solve the same problem as stated in (a) when X is observed and found equal to 14.

14.10 Denote the distribution function of the Poisson distribution with parameter λ by $\mathcal{P}_o(x; \lambda)$, and the distribution function of the χ^2 distribution with ν degrees of freedom by $\Gamma_\nu(z)$.

(a) Show that $\mathcal{P}_o(x \mid \lambda) = 1 - \Gamma_{2(x+1)}(2\lambda)$. (*Hint*: First show that $1 - \Gamma_{2(x+1)}(2\lambda) = \int_{2\lambda}^{\infty} \frac{u^x}{x!} e^{-u}\, du$, and next apply repeated partial integrations to the integral).

Table 14.8 Dataset for Problem 14.11.

12 373	107 318	9 739	13 000	12 207	63 589	31 893
98 474	5 784	9 662	61 731	15 269	4 730	11 269
26 947	27 838	90 682	8 086	7 905	48 162	

(b) Let $\lambda_1(X)$ and $\lambda_2(X)$ be defined by

$$\mathcal{P}_o(x \mid \lambda_1(x)) = \frac{\alpha}{2}.$$

$$\mathcal{P}_o(x - 1 \mid \lambda_2(x)) = 1 - \frac{\alpha}{2}.$$

Use the result of (a) to show that

$$\lambda_1(x) = \frac{1}{2}\, z_{\alpha/2,2x} \quad \text{and}$$

$$\lambda_2(x) = \frac{1}{2}\, z_{1-\alpha/2,2(x+1)},$$

where $z_{\varepsilon,\nu}$ is the upper 100 ε% percentile of the χ^2 distribution with ν degrees of freedom.

14.11 Historical data with a record of 20 times-to-failure (in hours) of a pressure transmitter (PT) are available in Table 14.8. The dataset can be downloaded from the `book companion site`.

(a) Explain why it is reasonable to assume a constant failure rate for the PT.

(b) Determine the empirical cumulative distribution corresponding to this dataset and plot it.

(c) Estimate the failure rate of PT.

(d) Find the survivor function obtained with the estimated failure rate and compare to the one obtained with the empirical distribution. Comment and explain how to improve the result.

14.12 Reconsider the situation in Example 14.16, but assume that the times-to-failure are those that are not starred. They are given in Table 14.9. The dataset can be downloaded from the `book companion site`.

(a) Determine the Kaplan–Meier estimate $\widehat{R}(t)$ and display it graphically.

(b) Determine the Nelson–Aalen estimate $R^*(t)$ for the survivor function and display it graphically.

14.13 Table 14.10 shows the intervals in operating hours between successive failures of air-conditioning equipment in a Boeing 720 aircraft. The first

Table 14.9 Dataset for Problem 14.12.

31.7	39.2*	57.5	65.5	65.8*	70.0	75.0*	75.2*
87.5*	88.3*	94.2	101.7*	105.8*	109.2	110.0	130.0*

*Censored data points.

Table 14.10 Dataset for Problem 14.13.

413	14	58	37	100	65	9	169
447	184	36	201	118	34	31	18
18	67	57	62	7	22	34	

Source: Proschan (1963).

interval is 413, the second is 14, and so on. The data are from Proschan (1963). The dataset can be downloaded from the `book companion site`.

(a) Establish the Nelson–Aalen plot ($N(t)$ plot) of the dataset. Describe (with words) the shape of the rate of occurrence of failures (ROCOF).

14.14 Suppose that the dataset in Problem 14.11 was obtained by simultaneously activating 20 identical items, but that the test was terminated at the 12th failure.

(a) What type of censoring is this?
(b) Estimate λ in this situation.
(c) Calculate a 95% confidence interval for λ.
(d) Compare the results with those derived in Problem 14.11.

14.15 Establish a graph paper such that the Nelson–Aalen plot of normally distributed ($\mathcal{N}(\mu, \sigma^2)$) life data is close to a straight line. Describe how the parameters μ and σ may be estimated from the plot.

14.16 Table 14.11 shows the intervals in days between successive failures of a piece of software developed as part of a large data system. The first interval is 9, the second is 12, and so on. The data are from Jelinski and Moranda (1972).The dataset can be downloaded on the `book companion site`.

(a) Establish the Nelson–Aalen plot ($N(t)$ plot) of the dataset. Is the ROCOF increasing or decreasing?
(b) Assume that the ROCOF follows a log-linear model, and find the maximum likelihood estimates (MLE) for the parameters of this model.

Table 14.11 Dataset for Problem 14.16.

9	12	11	4	7	2	5	8	5	7
1	6	1	9	4	1	3	3	6	1
11	33	7	91	2	1	87	47	12	9
135	258	16	35						

Source: Jelinski and Moranda (1972).

Table 14.12 Dataset for Problem 14.17.

31.7	39.2	57.5	65.0	65.8	70.0	75.0	75.2
87.7	88.3	94.2	101.7	105.8	109.2	110.0	130.0

(c) Draw the estimated cumulative ROCOF in the same diagram as the Nelson–Aalen plot. Is the fit acceptable?

(d) Use the Laplace test to test whether the ROCOF is decreasing or not (use a 5% level of significance).

14.17 Independent lifetimes (given in months in Table 14.12) have been observed with no censoring. The dataset can be downloaded from the `book companion site`.

(a) Give the analytical expression of empirical distribution function and explain your method.

(b) Give a script to get this function.

(c) Give a plot of it.

(d) Assuming that the probability density function is an exponential law of parameter λ, give the optimal value of λ to fit to the given dataset.

(e) Is it reasonable to assume that such a unit has an exponential density function? Why?

14.18 A record of the times-to-failure (given in hours) of a sensor give the following historical dataset in Table 14.13. The dataset can be downloaded from the `book companion site`.

Table 14.13 Dataset for Problem 14.18.

1.2×10^4	9.3×10^4	0.5×10^4	0.2×10^4	1.1×10^4
2.6×10^4	9.4×10^4	1.2×10^4	4.9×10^4	9.6×10^4
0.9×10^4	8.6×10^4	6.5×10^4	0.5×10^4	1.0×10^4
0.1×10^4	0.8×10^4	3.6×10^4	3.2×10^4	

(a) Demonstrate that it is reasonable to assume a constant failure rate for the sensor.

(b) Determine the empirical cumulative distribution corresponding to this dataset and plot it.

(c) Propose two methods to estimate the failure rate.

(d) Determine the survivor function obtained with the estimated failure rate and compare with the one you obtain with the empirical distribution. Comment and explain how you could improve your results.

(e) For all the units, calculate their MTTF and the probability that they survive their own MTTF. Give comments.

(f) Determine a plot for the survivor functions and identify the time horizons t_k for which the survivor function of k items ($k = 0, 1, 2, \ldots$) is higher than 0.9.

References

Aalen, O.O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics* 6: 701–726.

Aalen, O.O., Borgan, Ø., and Gjessing, H.K. (2008). *Survival and Event History Analysis; A Process Point of View*. New York: Springer.

Ansell, J.I. and Phillips, M.J. (1994). *Practical Methods for Reliability Data Analysis*. Oxford: Oxford University Press.

Barlow, R.E. and Campo, R. (1975). Total time on test processes and applications to failure analysis. In: *Reliability and Fault Tree Analysis* (ed. R.E. Barlow, J.B. Fussell, and N.D. Singpurwalla). Philadelphia, PA: SIAM. 451–481.

Bergman, B. and Klefsjö, B. (1982). A graphical method applicable to age-replacement problems. *IEEE Transactions on Reliability* R-31 (5): 478–481.

Bergman, B. and Klefsjö, B. (1984). The total time on test concept and its use in reliability theory. *Operations Research* 32 (3): 596–606.

Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society* B 21: 411–421.

Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman and Hall.

Crowder, M.J., Kimber, A.C., Sweeting, T.J., and Smith, R.L. (1991). *Statistical Analysis of Reliability Data*. Boca Raton, FL: Chapman and Hall.

Fox, J. and Weisberg, S. (2019). *An R Companion to Applied Regression*. Los Angeles, CA: Sage Publications.

Jelinski, Z. and Moranda, P.B. (1972). Software reliability research. In: *Statistical Computer Performance Evaluation* (ed. W. Freiberger), 465–484. New York: Academic Press.

Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. Hoboken, NJ: Wiley.

Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53 (282): 457–481.

Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*. Hoboken, NJ: Wiley.

Lieblein, J. and Zelen, M. (1956). Statistical investigation of the fatigue life of deep-groove ball bearings. *Journal of Research of the National Bureau of Standards* 57: 273–316.

Mann, N.R., Schafer, R.E., and Singpurwalla, N.D. (1974). *Methods for Statistical Analysis of Reliability and Lifetime Data*. Hoboken, NJ: Wiley.

McCool, J.I. (2012). *Using the Weibull Distribution; Reliability, Modeling, and Inference*. Hoboken, NJ: Wiley.

Meeker, W.Q. and Escobar, L.A. (1998). *Statistical Methods for Reliability Data*. Hoboken, NJ: Wiley.

MIL-HDBK-217F (1995). Reliability Prediction of Electronic Equipment. *Military Handbook*. Washington, DC: U.S. Department of Defense.

Moore, D.F. (2016). *Applied Survival Analysis Using R*. Springer.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics* 14: 945–965.

Proschan, F. (1963). Theoretical explanation of observed decreasing failure rate. *Technometrics* 5: 375–383.

Ross, S.M. (2014). *Introduction to Probability Models*, 11e. Academic Press.

Tukey, J.W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

15

Bayesian Reliability Analysis

15.1 Introduction

This chapter gives a brief introduction to Bayesian modeling and Bayesian data analysis. These aspects are presented and explained by simple examples with a single parameter θ and mostly a single random variable X. For these examples, it is straightforward to obtain results by hand calculation. For models with two or more parameters, it is not feasible to solve the equations by hand calculation, and we need to rely on the use of computers. A very brief introduction to computerized Bayesian analysis is given at the end of the chapter. More details may be found in the references cited.

The Bayesian approach has increasing popularity is recent years for two main reasons:

(1) Several user-friendly computer programs have become available that can solve problems that we were not able to solve by hand calculation.
(2) Bayesian methods have become a central element in the development of new technologies, such as artificial intelligence (AI) and machine learning (ML).

In the Bayesian approach to reliability analysis, probability is a measure of the analyst's *degree of belief* about a specific situation or a specific outcome. The Bayesian view is different from the *classical* and the *frequentist* views of probability. The frequentist view is the one we tacitly have used in the first 14 chapters of this book.

15.1.1 Three Interpretations of Probability

The definition and interpretation of the term *probability* has, for a long time, been a controversial issue. What do we really mean when we say that the probability of an event A is 0.90? Is it an objective statement or a subjective statement? Three

System Reliability Theory: Models, Statistical Methods, and Applications, Third Edition.
Marvin Rausand, Anne Barros, and Arnljot Høyland.

Companion Website: www.wiley.com/go/SystemReliabilityTheory3e

dominating views prevail: the classical, the frequentist, and the subjective view. The three views are briefly introduced below.

Classical Probability

With the classical view, experiments with n equally likely single outcomes $e_1, e_2, \dots, e_n$ are considered. The set of all possible single outcomes is called the sample space $S = \{e_1, e_2, \dots, e_n\}$. The probability of a particular event $A \in S$, $\Pr(A)$, is calculated as the number of single outcomes that fulfill A divided by the total number n of possible outcomes. If the experiment consists of throwing a dice, $n = 6$, and the sample space is $\{1, 2, 3, 4, 5, 6\}$. If A = "outcome is an odd number" when throwing the dice, the single outcomes fulfilling A are the three outcomes 1, 3, and 5, and the probability of A is $\Pr(A) = 3/6 = 1/2$. The classical view is applicable only in cases with equally likely single outcomes.

Frequentist Probability

With the frequentist view, it is imagined that a series of n independent and identical experiments can be carried out. Each experiment may, or may not, result in an event A, and the number n_A of experiments that resulted in A is counted. The frequency n_A/n is assumed to approach a limiting value when the number, n, of experiments increases, and this limit is called the probability of A, and is written as $\Pr(A)$. With this view, probability $\Pr(A)$ is an unknown, but existing number, and it is our job to determine this number. To determine this number, we often have to use probabilistic models. Let T be the time-to-failure of an item, and let the event A be $T > t_0$ for a specified t_0. Chapter 5 introduces a range of models that help us to determine $\Pr(A)$. Among these are the exponential, Weibull, and lognormal models. The model to use is often chosen as a compromise between what is realistic and what is feasible with the limited amount of data we have access to. The frequentist view is often claimed to be *objective*, but the choice of a model has several subjective elements that may influence the resulting probabilities.

With the frequentist view, model parameters (e.g. λ in the exponential model) are estimated solely based on the data without using any existing knowledge about the parameters. See Chapter 14.

Both the subjective and frequentist views require that a random variable X is a real-valued, measurable quantity that can be observed and recorded when an experiment is carried out. More formally, a random variable is a function $f : S \to \mathbb{R}$ from the sample space to the real numbers. Quantities that cannot be observed and measured are not random variables.

Subjective Probability

With the subjective view, the probability $\Pr(A)$ of an event A is a measure of the analyst's *degree of belief* about a quantity or an outcome. The degree of belief is formed by the analyst's knowledge and experience with the event A. She may call on experts in the relevant domains and use physical and symmetry arguments to form her degree of belief. For experiments that comply with the classical view, the subjective and classical views give the same probabilities because of symmetry.

With the subjective view, a wide range of quantities and issues may be treated as random variables, many of which have no meaning within the classical and frequentist views. In our context, the most important feature of the subjective view is that we may consider a parameter θ of a probability distribution to be a random variable Θ with (probability) density $\pi(\theta)$. The density $\pi(\theta)$ describes the analyst's degree of belief about the value of the parameter. A *subjective* probability is also called a *Bayesian* probability.

Relevance for Reliability

Reliability analyses give the highest yields in the early design phases of new systems. New systems often comprise new components, based on a new design, new materials, and/or new technologies, with no or limited field experience. To obtain adequate parameter estimates, we need to test the components in a relevant operating context. For high reliability components, this is both time-consuming and costly – and the estimate will usually come after the design decision has been made. In many cases, new components are minor modifications of existing components from which we have some experience. This experience should be used to provide parameter estimates for the new components. For this purpose, the subjective probabilities provide an excellent framework.

15.1.2 Bayes' Formula

Bayesian inference is a method of statistical inference where Bayes' formula[1] is used to update our degree of belief as more evidence or data becomes available. Bayes' formula has been mentioned several times earlier in this book. Here, we explain this formula by first looking at an experiment with discrete outcomes. Let S be the sample space of all possible outcomes of an experiment, and let $E_1, E_2, \ldots, E_m$ be mutually exclusive events such that $\bigcup_{i=1}^m E_i = S$. This means that exactly one of the events E_i will occur when the experiment is performed. Consider another event A in S. This event may be written as

$$A = A \cap S = A \cap \bigcup_{i=1}^{m} E_i = \bigcup_{i=1}^{m} (A \cap E_i).$$

[1] Also called Bayes' theorem.

Because the events $A \cap E_i$ are mutually exclusive for $i = 1, 2, \dots, m$, the probability of A is

$$\Pr(A) = \sum_{i=1}^{m} \Pr(A \cap E_i) = \sum_{i=1}^{m} \Pr(A \mid E_i) \Pr(E_i). \tag{15.1}$$

Equation (15.1) is called the law of total probability (see Section 6.2.4). Suppose that we know that event A has occurred and ask "what is the probability that event E_j also occurs?" This probability is by using the definition of conditional probability

$$\Pr(E_j \mid A) = \frac{\Pr(E_j \cap A)}{\Pr(A)} = \frac{\Pr(A \mid E_j)}{\sum_{i=1}^{m} \Pr(A \mid E_i) \Pr(E_i)}, \tag{15.2}$$

which is Bayes' formula for events. If we consider a discrete random variable Y with sample space $\mathcal{S}_Y = \{y_1, y_2, \dots, y_m\}$, we may let E_i be the event that $Y = y_i$, for $i = 1, 2, \dots, m$. Also assume that we have another random variable X with sample space $\mathcal{S}_X = \{x_1, x_2, \dots, x_n\}$. From Bayes' formula (15.2), the conditional probability mass function for Y given $X = x_\ell$ is

$$\Pr(Y = y_j \mid X = x_\ell) = \frac{\Pr(X = x_\ell \mid Y = y_j) \Pr(Y = y_j)}{\sum_{i=1}^{m} \Pr(X = x_\ell \mid Y = y_i) \Pr(Y = y_i)}, \tag{15.3}$$

which is Bayes' formula for discrete variables.

The similar formula for continuous variables can be developed by analogy to (15.3). Consider the continuously distributed random variable X that can take values in $\mathcal{S}_X$, as our observable variable. Let Θ be a continuous variable that will represent our parameter with density $\pi(\theta)$ with sample space $\mathcal{S}_\Theta$. The (probability) density for X is written $f(x \mid \theta)$. Assume that we have carried out the experiment and have got the result $X = x_\ell$. The density for Θ when we know that $X = x_\ell$ is

$$\pi(\theta \mid x_\ell) = \frac{f(x_\ell \mid \theta)\, \pi(\theta)}{\int_{\theta' \in \mathcal{S}_\Theta} f(x_\ell \mid \theta')\, \pi(\theta')\, d\theta'}, \tag{15.4}$$

which is Bayes' formula for continuous variables. In addition, it is obviously relevant to have one discrete variable and one continuous variable, as will be illustrated later in this chapter.

15.2 Bayesian Data Analysis

To highlight the similarities and differences between the frequentist and the Bayesian approach to data analysis, we start with a brief recap of the main elements of frequentist data analysis.

15.2.1 Frequentist Data Analysis

The approaches to data analyses in Chapter 14 are based on the *frequentist* interpretation of probability and starts with a model of the data. For parametric models, the model has a fixed but an unknown parameter and may be represented by a probability density or a probability mass function. The model is established based on knowledge about the item and the operating context before we start looking at the data.

The data is usually a set of n independent realization, $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$ of a random variable X with density $f(\boldsymbol{x}, \theta)$ or probability mass function $\Pr(\boldsymbol{X} = \boldsymbol{x} \mid \theta)$. The data is combined with the model in a data analysis, and this analysis gives information about the parameter θ, often in the form of an estimate $\hat{\theta}$. The analysis procedure is outlined in Figure 15.1.

No *initial* information about the value of the parameter θ is included in the frequentist data analysis.

15.2.2 Bayesian Data Analysis

The main elements of the Bayesian approach to data analysis are the following:

(1) A *prior distribution* $\pi(\theta)$ that expresses our degree of belief about θ prior to observing any data.
(2) A *likelihood function* $L(\theta \mid d)$ that expresses the "likelihood" that a particular value of the parameter θ has "produced" the obtained data d.
(3) A *posterior distribution* $\pi(\theta \mid d)$ that expresses our degree of belief about θ after the data d has been observed.
(4) A *Bayesian inference* procedure that derives appropriate statements from the posterior distribution, such as point estimates, interval estimates, and probabilities of hypotheses.

The main difference between the frequentist approach and the Bayesian approach is that the Bayesian approach also uses initial knowledge about the value of the parameter. This initial knowledge is called the *prior knowledge* in the

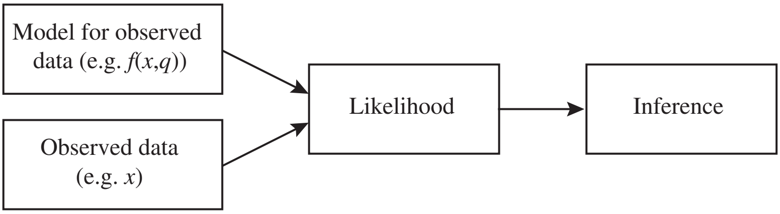


Figure 15.1 The frequentist data analysis process.

Bayesian approach, and the data is used to update this knowledge to a *posterior knowledge*.

The Bayesian approach can be summarized by the following steps:

(1) The data to be studied can be represented by a random variable X. When a data set has been observed, the available data is a set of numbers $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$.
(2) Based on an understanding of the physical situation, the analyst chooses a probability model for the random variable X. For continuous random variables, this involves choosing a density $f(\boldsymbol{x} \mid \theta)$, where θ is the unknown parameter(s) of the model. For discrete random variables, this involves choosing a probability mass function $\Pr(X = x \mid \theta)$.
(3) Prior information about the actual value of the parameter θ (e.g. previous experience with the same, or similar, items, expert judgments) is included in the model as a density $\pi(\theta)$ called the *prior density*. This density represents the analyst's *degree of belief* about θ prior to looking at the data, $\boldsymbol{x}$.
(4) After observing the data $\boldsymbol{x}$, the analyst applies Bayes' formula to update her beliefs and calculates the *posterior distribution* $\pi(\theta \mid \boldsymbol{x})$.
(5) A fifth step is sometimes included. It involves evaluating the fit of the model and the implications of the resulting posterior distribution: how well does the model fit the data, are the substantive conclusions reasonable, and how sensitive are the results to the modeling assumptions? In response, one can alter or expand the model and repeat steps 2–4 (Gelman et al. 2013).

The Bayesian data analysis process is shown in Figure 15.2. The main elements of the Bayesian approach are discussed briefly in the remainder of the section.

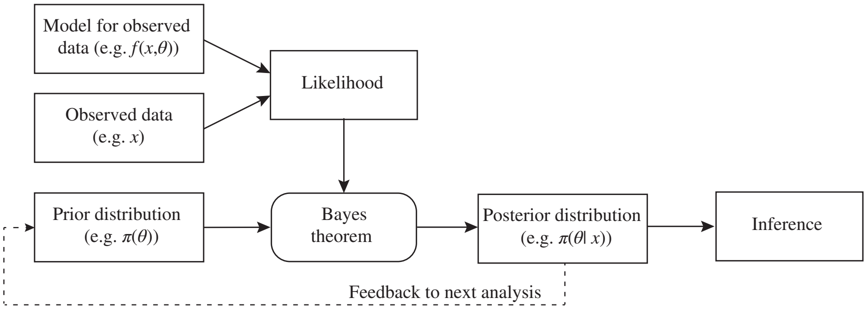

Model for observed data (e.g. $f(x,\theta)$)
Observed data (e.g. x)
Likelihood
Prior distribution (e.g. $\pi(\theta)$)
Bayes theorem
Posterior distribution (e.g. $\pi(\theta| x)$)
Inference
Feedback to next analysis

Figure 15.2 The Bayesian data analysis process.

15.2.3 Model for Observed Data

Earlier in this book, we have studied a number of different models for data to be observed. Among these are the binomial model

$$\Pr(X = x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \qquad \text{for } x = 0, 1, 2, \dots, n, \text{ and } 0 \le \theta \le 1,$$

and the exponential model

$$f_T(t \mid \lambda) = \lambda e^{-\lambda t} \qquad \text{for } t \ge 0, \text{ and } \lambda > 0.$$

The binomial model is a *discrete model* because the sample space for X is discrete $\{0, 1, 2, \dots, n\}$. The exponential model is said to be a *continuous model* because T can take all positive values. Observe that the models are written as conditional models, given the value of the parameter.

15.2.4 Prior Distribution

To simplify the presentation, we assume that the model for the observed data has only one parameter θ. In the Bayesian set-up, this parameter is regarded as a random variable Θ. The data resulting from the experiment come from n independent observations of a random variable X. To simplify the presentation, we assume that X is a continuous random variable with density $f(x \mid \theta)$.

The joint density of the n independent observations $X_1, X_2, \dots, X_n$ is

$$f(x_1, x_2, \dots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

For brevity, we often write $\{x_1, x_2, \dots, x_n\} = \boldsymbol{x}$.

Before analyzing the data, our uncertainty about the value of this parameter is formulated as a *prior distribution*, defined as follows:

Definition 15.1 (Prior distribution)
A probability distribution of an uncertain quantity that expresses the analyst's degree of belief about this quantity before any evidence (i.e. observed data) is taken into account. □

The prior distribution is usually given as a *prior density* $\pi(\theta)$ for the possible values $\theta \in \mathcal{S}_\Omega$, where $\mathcal{S}_\Omega$ is the set (or sample space) of possible values for θ. The prior density expresses our degree of belief and our uncertainty about the value of the parameter before the experiment is carried out. The prior density is determined from our prior knowledge and our beliefs about the value of the parameter and is hence a *subjective* distribution.

Prior distributions may be classified into three categories:

Informative priors provide information that is crucial to the inference made from the model.
Weakly informative priors do not provide any controversial information, but are strong enough to pull the data away from inappropriate inferences that are consistent with the likelihood.
Noninformative priors do not provide any additional information to the inference. A noninformative prior is uniform, or nearly so.

The term *diffuse prior* is sometimes used to denote a noninformative or weakly informative priors. These priors are sometimes chosen based on the argument that we should "allow the data to speak."

Remark 15.1 (Categories of Bayesians)
There are several categories of Bayesian statisticians:

Subjective Bayesians interpret probability strictly as personal degrees of belief.
Objective Bayesians start with a noninformative prior distribution and claim that the resulting posterior is objective.
Empirical Bayesians estimate the prior density from available data. □

15.2.5 Observed Data

The observed data are a realization of the random variable(s) and hence a known single number or sequence of numbers, such as $\{x_1, x_2, \dots, x_n\}$. In the discussions of unspecified models, we refer to the observed data as "data." The observed data may include covariates, but these are tacitly ignored in this simplified presentation.

15.2.6 Likelihood Function

The likelihood function is introduced in Section 14.4.4 and is written as $L(\theta \mid d)$. When the data has been observed in an experiment, for example data $= \boldsymbol{x}$, the likelihood function is given by the same mathematical expression as the joint density $f(\text{data} \mid \theta)$, but the interpretations of the two expressions are quite different. When only one variable is observed and the data is one-dimensional, the two expressions may be interpreted as follows:

- *Probability density*, $f(d \mid \theta)$. The probability that the random variable D gives data d in a small interval $(d, d + \Delta d]$ is

$$\Pr(d < D \leq d + \Delta d) \approx f(d \mid \theta)\Delta d.$$

Knowing the value of θ, the density tells us which outcome d should be expected from the experiment.

- *Likelihood function*, $L(\theta \mid d)$. In this case, the data d has been observed and is known. It is obvious that the value of the parameter θ strongly influences the outcome d of an experiment. The outcome d for a high value of θ will generally be quite different from the outcome for a low value of θ. The likelihood function $L(\theta \mid d)$ tells us the likelihood that a particular value of θ produced the outcome d.

Similar arguments are made for a multidimensional d and for discrete model distributions. Observe that the likelihood function is *not* a probability density function for Θ, given d. Further details about the likelihood function is found in Section 14.4.4.

Example 15.1 (Likelihood function for binomial model)
Consider an experiment where we observe a variable X that is binomially distributed $X \sim \text{binom}(n, \theta)$. The probability of getting the outcome x is

$$\Pr(X = x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x} \qquad \text{for } x = 0, 1, 2, \dots, n, \tag{15.5}$$

where the number of trials, n, is a specified and known number. Assume that the experiment gave the outcome $X = x$, where $x \in \{0, 1, 2, \dots, n\}$ and that this outcome is known. The likelihood function for this outcome is

$$L(\theta \mid x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x} \qquad \text{for } 0 \le \theta \le 1. \tag{15.6}$$

The likelihood function for $n = 10$ and $x = 3$ is shown in Figure 14.11. We observe that $L(\theta \mid x)$ attains its maximum for $\theta = 0.3$, which is the maximum likelihood estimate (MLE) for θ for this data. □

15.2.7 Posterior Distribution

When the prior distribution and the observed data are available, Bayes' formula may be used to update our prior degree of belief into a *posterior distribution*, defined as follows:

Definition 15.2 (Posterior distribution)
The probability distribution of an uncertain quantity that is assigned after relevant evidence or background is taken into account. □

The posterior distribution is also called the *aposteriori distribution* or simply the *posterior*. The posterior distribution summarizes all our current information about the unknown parameter θ, by combining our prior information, specified by $\pi(\theta)$, and the information about θ we obtain from the observed data.

The posterior distribution is usually given as a *posterior density*, $\pi(\theta \mid \text{data})$, for the possible values $\theta \in \Omega$. To make clear that we refer to the posterior and not the prior density, we sometimes write the posterior density as $\pi_{\Theta|\text{data}}(\theta \mid \text{data})$.

The following two examples show how Bayes' formula is used to find the posterior distribution for discrete and continuous models, respectively.

Example 15.2 (Discrete distribution)

Consider a discrete random variable X with probability mass function with a single parameter θ, $\Pr(X = x_i \mid \theta)$, for $i = 1, 2, 3, \dots$ and $\theta \in \Omega$.

Let $\pi(\theta)$ be the prior density that reflects our degree of belief about the value of θ before any data is available. In this case, we have an observable variable X that is discrete and a parameter Θ that has a continuous distribution. This means that we need to use a combination of two Bayes' formulas in (15.3) and (15.4). The new formula can be developed by analogy to the two others versions of Bayes' formula.

Assume that a single experiment is conducted, giving the outcome $X = x_i$ for some i in $\{x_1, x_2, \dots, x_n\}$. Observe that when the experiment is performed, the number x_i is known.

Bayes' formula is used to derive the posterior density

$$\pi(\theta \mid x_i) = \frac{\Pr(X = x_i \mid \theta)\ \pi(\theta)}{\int_\Omega \Pr(X = x_i \mid \theta')\ \pi(\theta')\ d\theta'} \qquad \text{for } \theta \in \Omega.$$

Because the observation x_i is known, the only variable in $\pi(\theta \mid x_i)$ is θ, and we should rather replace the probability $\Pr(X = x_i \mid \theta)$ with the likelihood function $L(\theta \mid \theta)$. The posterior density can then be written as

$$\pi(\theta \mid x_i) = \frac{L(\theta) \mid x_i)\pi(\theta)}{\int_\Omega L(\theta' \mid x_i)\ \pi(\theta')\ d\theta'} \qquad \text{for } \theta \in \Omega. \tag{15.7}$$

The posterior probability $\pi(\theta \mid x_i)$ is a proper probability distribution that fulfills $\int_\Omega \pi(\theta \mid x_i)\ d\theta = 1$. In Section 14.4.4, it is argued that factors of the likelihood function that do not include the parameter may be deleted. The same applies here, and we may therefore write

$$\pi(\theta \mid x_i) \propto L(\theta \mid x_i)\ \pi(\theta), \tag{15.8}$$

where the symbol $\propto$ means proportional to. The posterior density is therefore proportional to the product of the likelihood function and the prior density. Because we know that the posterior density is a proper density, the constant of proportionality may be fitted afterwards (if required). □

Example 15.3 (Continuous distribution)
The random variable X has continuous density $f(x \mid \theta)$, where the actual value of θ is a realization of a random variable Θ with a prior density $\pi(\theta)$ for $\theta \in \Omega$ that describes the analyst's degree of belief about the value of θ.

Assume that an experiment gave the outcome x, where x can be a single number or a vector of numbers. Bayes' formula (15.4) gives the posterior density

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\ \pi(\theta)}{\int_0^\infty f(x \mid \theta')\ \pi(\theta')\ d\theta'} \qquad \text{for } \theta \in \Omega.$$

Because x is a known value, $f(x \mid \theta)$ is the likelihood function $L(\theta \mid d)$, and because the denominator is a constant, the posterior density may, as in Example 15.2, be written as

$$\pi(\theta \mid x) \propto L(\theta \mid x)\ \pi(\theta). \tag{15.9}$$

□

15.3 Selection of Prior Distribution

To be able to determine the posterior distribution by hand calculation, it is important to select a prior distribution that "fits" the model distribution for the observed data, such that (i) the prior distribution is flexible enough to describe our degree of belief about θ and (ii) that it is possible to determine the posterior distribution with hand calculation. We now discuss some typical one-parameter situations.

15.3.1 Binomial Model

Consider the binomially distributed random variable $X \sim \text{binom}(n, \theta)$, where n is a specified and known number of trials.

$$\Pr(X = x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \qquad \text{for } x = 0, 1, \dots, n \text{ and } 0 \le \theta \le 1. \tag{15.10}$$

The possible values of the unknown parameter θ are in the interval $[0, 1]$ so we need a continuous prior distribution that takes values in the same interval. A commonly used distribution for this purpose is the *beta distribution*.

Beta Prior Distribution

For parameters that express probabilities, the beta distribution is often chosen to express our prior information. Θ is beta distributed, $\Theta \sim \text{beta}(r, s)$, over the interval $[0, 1]$ when its density is

$$\pi(\theta) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\ \theta^{r-1}(1-\theta)^{s-1} \qquad \text{for } 0 \le \theta \le 1. \tag{15.11}$$

The beta distribution is introduced in Section 5.7.2. Please observe that we, in this chapter, are using other symbols for the parameters of the beta distribution than in Section 5.7.2 (i.e. $\alpha = r$ and $\beta = s$).

The beta distribution has mean and standard deviation (SD)

$$E(\Theta) = \frac{r}{r+s} \tag{15.12}$$

$$\text{SD}(\Theta) = \sqrt{\frac{rs}{(r+s)^2(r+s+1)}}. \tag{15.13}$$

The beta distribution is available in R where the beta density function, for example, is obtained by `dbeta(x,r,s,log=F)` for given values of $x(=\theta), r$, and s. The beta distribution is rather flexible, and the parameters can be adapted to fit almost any degree of belief we have about θ. Observe that by choosing $r = s = 1$, the density becomes $\pi(\theta) = 1$ for $0 \leq \theta \leq 1$. This is the *uniform distribution* over $[0, 1]$ and means that we consider all probabilities in $[0, 1]$ to be equally likely (see Section 5.7.1). This means that the prior distribution does not provide any information about the value of θ, and is hence *noninformative*.

If we believe that the probability θ is approximately 0.2, we may let $\text{E}(\Theta) = 0.2$, in which case $s = 4r$ The standard deviation can now be expressed by r and becomes

$$\text{SD}(\Theta) = \sqrt{\frac{4}{5(5r+1)}}.$$

Our uncertainty about the value of θ may be expressed by, for example, $\text{SD}(\Theta) = 0.25$, which gives $r = 2.36$ and $s = 9.44$. The corresponding beta distribution is shown in Figure 15.3 and made by the R script

```
# Set the range (i.e. [0,1]) and the number of val-
ues to calculate
x<-seq(0,1,length=100)
# Specify the parameters r and s
r<-2.36
s<-9.44
# Calculate the beta density for each x
y<-dbeta(x,r,s,log=F)
plot(x,y,type="l",xlab=expression(theta),
ylab=expression(pi(theta)))
```

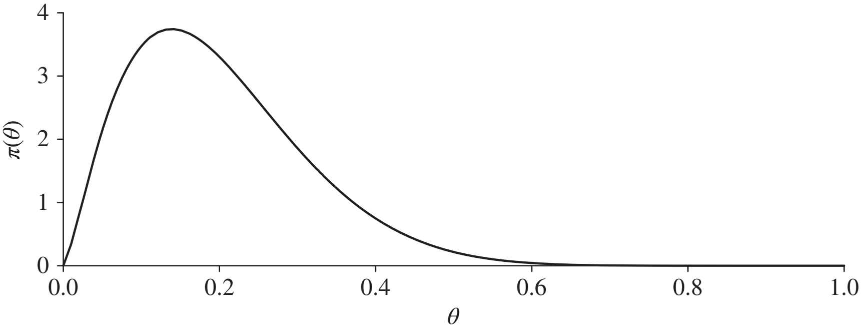


Figure 15.3 Prior beta density with parameters $r = 2.36$ and $s = 4r$.

Posterior Distribution

Assume that $X = x$ has been observed. The posterior density is from (15.8)

$$\begin{aligned}\pi(\theta \mid x) &\propto L(\theta \mid x)\ \pi(\theta)\\ &\propto \theta^x(1-\theta)^{n-x}\theta^{r-1}(1-\theta)^{s-1}\\ &\propto \theta^{x+r-1}(1-\theta)^{n-x+s-1},\end{aligned}$$

which (apart from a constant) is seen to be a beta distribution with parameters $(x + r)$ and $(n - x + s)$. This means that the prior distribution and the posterior distribution come from the same class of distribution. Two distributions (here the binomial and the beta distributions) with this property are said to be *conjugate distributions*.

The prior mean value is $E(\Theta) = r/(r + s)$ and the posterior mean value is

$$E(\Theta \mid x) = \frac{r+x}{x+r+n-x+s} = \frac{x+r}{n+s+r}. \tag{15.14}$$

Remark 15.2 (Conjugate distributions)

To use a prior distribution for the parameter that is a *conjugate* to the model distribution makes the Bayesian analysis simple, and it is therefore important to identify the conjugate distributions when using hand calculation. Current computer programs for Bayesian analysis are based on Monte Carlo simulation and do not use conjugacy in the posterior sampling. When using a computer program for Bayesian analysis, you may choose whichever distribution you like as prior distribution, even a histogram prior. □

15.3.2 Exponential Model – Single Observation

Assume that the time-to-failure T of an item is exponentially distributed with parameter (failure rate) λ.

$$f(t \mid \lambda) = \lambda e^{-\lambda t} \qquad \text{for } t \geq 0 \text{ and } \lambda > 0.$$

Assume that t_1 has been observed and is known. The likelihood function is

$$L(\lambda \mid t_1) = \lambda e^{-\lambda t_1} \qquad \text{for } \lambda > 0. \tag{15.15}$$

The analyst's prior belief about λ may be expressed by the random variable Λ with a prior distribution. A common distribution for this purpose is the *gamma distribution*.

Gamma Prior Distribution

The gamma distribution (see Section 5.4.2) is often the preferred prior distribution for parameters that can take any positive value. A random variable Λ is said to be gamma distributed, $\Lambda \sim \text{gamma}(\alpha, \beta)$ when its density is

$$\pi(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, \lambda^{\alpha-1} e^{-\beta\lambda} \qquad \text{for } \lambda > 0. \tag{15.16}$$

From Section 5.4.2, the mean and the standard deviation of the gamma distribution are

$$E(\Lambda) = \frac{\alpha}{\beta} \tag{15.17}$$

$$\text{SD}(\Lambda) = \frac{\sqrt{\alpha}}{\beta}. \tag{15.18}$$

The gamma distribution is flexible, and we may adapt the parameters α and β to fit our prior belief about the value of the parameter (failure rate) λ. We may, for example, have experience data from items that are similar to the actual item. This may lead us to believe that the failure rate should be around 1.2×10^{-3} h^{-1}, with a standard deviation of about 6×10^{-4}. If we use these values for the mean and the standard deviation, we may solve for α and β and obtain $\alpha = 4.5$ and $\beta = 3700$. The corresponding density is shown in Figure 15.4, which is made with an R script similar to the one used to make Figure 15.3.

Posterior Distribution

When $T = t_1$ has been observed, the posterior density is from (15.4)

$$\begin{aligned}\pi(\lambda \mid t_1) &\propto L(\lambda \mid t_1)\, \pi(\lambda) \\ &\propto \lambda e^{-\lambda t_1} \lambda^{\alpha-1} e^{-\beta\lambda} = \lambda^{\alpha+1-1} e^{-(\beta+t_1)\lambda},\end{aligned} \tag{15.19}$$

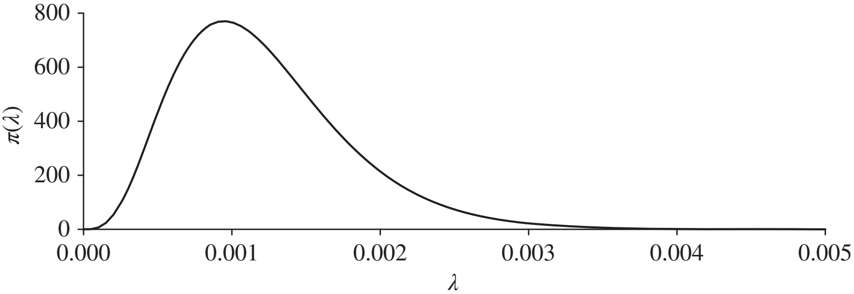


Figure 15.4 The gamma distribution with parameters $\alpha = 4.5$ and $\beta = 3700$.

which (apart from a constant) is seen to be a gamma distribution with parameters $(\alpha + 1)$ and $(\beta + t_1)$. This means that the prior distribution and the posterior distribution come from the same class of distribution. The two distributions (here the exponential and the gamma distributions) are therefore *conjugate distributions*.

The prior mean is $E(\Lambda) = \alpha/\beta$, whereas the posterior mean is

$$E(\Lambda \mid t_1) = \frac{\alpha + 1}{\beta + t_1}. \tag{15.20}$$

15.3.3 Exponential Model – Multiple Observations

Let $\boldsymbol{T} = \{T_1, T_2, \dots, T_n\}$ be n independent and identically exponentially distributed times to failure with failure rate λ. The joint distribution of $\boldsymbol{T}$ is because of independence

$$f_{\boldsymbol{T}}(t_1, t_2, \dots, t_n) = \prod_{i=1}^{n} f_{T_i}(t_i) = \prod_{i=1}^{n} \lambda e^{-\lambda t_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} t_i}.$$

Assume that $\boldsymbol{t} = \{t_1, t_2, \dots, t_n\}$ have been observed and hence are known numbers. The gamma distribution, gamma(α, β), is chosen as prior distribution for the (random) parameter Λ. The posterior density is from (15.4)

$$\begin{aligned}\pi(\lambda \mid \boldsymbol{t}) &\propto L(\lambda \mid \boldsymbol{t})\, \pi(\lambda) \\ &\propto \lambda^n e^{-\lambda \sum_{i=1}^{n} t_i} \lambda^{\alpha-1} e^{-\beta\lambda} = \lambda^{\alpha+n-1} e^{-\lambda\left(\beta + \sum_{i=1}^{n} t_i\right)},\end{aligned} \tag{15.21}$$

which is recognized as a gamma distribution with parameters $(\alpha + n)$ and $\left(\beta + \sum_{i=1}^{n} t_i\right)$. The posterior mean is

$$E(\Lambda \mid \boldsymbol{t}) = \frac{\alpha + n}{\beta + \sum_{i=1}^{n} t_i}, \tag{15.22}$$

where n is the number of failures observed and $\sum_{i=1}^{n} t_i$ is the total time in operation.

Example 15.4 (Sequential updating)

Consider a nonrepairable valve with constant failure rate λ. Experience and various studies lead us to believe that the failure rate is a random variable $\Lambda \sim \text{gamma}(\alpha, \beta)$. The *prior* density of Λ is therefore

$$\pi(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\,\lambda^{\alpha-1}\,e^{-\beta\lambda} \qquad \text{for } \lambda > 0,$$

and the prior mean value of Λ is

$$E(\Lambda) = \frac{\alpha}{\beta}.$$

The density of the time-to-failure T of the valve, when the failure rate λ is known, is

$$f_{T|\Lambda}(t \mid \lambda) = \lambda e^{-\lambda t} \qquad \text{for } t > 0,\ \ \lambda > 0.$$

Assume that we can test n valves of the same type one by one. Before the first test, we assume the prior distribution of the failure rate Λ to be gamma distributed with parameters $\alpha_1 = 2$ and $\beta_1 = 1$,

$$\pi(\lambda) = \lambda\,e^{-\lambda} \qquad \text{for } \lambda > 0.$$

Let T_1 be the time-to-failure of the first valve tested. The joint density of T_1 and Λ becomes

$$\begin{aligned} f_{T_1,\Lambda}(t_1, \lambda) &= f_{T_1|\Lambda}(t_1 \mid \lambda)\ \pi(\lambda) = \lambda e^{-\lambda t_1}\ \lambda e^{-\lambda} \\ &= \lambda^2 e^{-\lambda(t_1+1)} \qquad \text{for } t_1 > 0,\ \ \lambda > 0. \end{aligned}$$

The marginal density of T_1 is

$$f_{T_1}(t_1) = \int_0^\infty \lambda^2 e^{-\lambda(t_1+1)}\,d\lambda = \frac{\Gamma(3)}{(t_1+1)^3} = \frac{2}{(t_1+1)^3} \qquad \text{for } t > 0.$$

The conditional density of Λ, given $T_1 = t_1$, that is the posterior density, is

$$\begin{aligned} \pi(\lambda \mid t_1) &= \frac{\lambda^2 e^{-\lambda(t_1+1)}}{2}(t_1+1)^3 \\ &= \frac{(t_1+1)^3}{\Gamma(3)}\lambda^{3-1}e^{-\lambda(t_1+1)} \qquad \text{for } \lambda > 0, \end{aligned}$$

which is also seen to be a gamma density, now with parameters α_2, and β_2, where

$$\begin{aligned} &\alpha_2 = 3 = \alpha_1 + 1 && \text{because } \alpha_1 = 2 \\ &\beta_2 = (t_1+1) = \beta_1 + t_1 && \text{because } \beta_1 = 1 \end{aligned}.$$

This procedure may now be repeated with $\pi(\lambda \mid t_1)$ as our new prior distribution. Then, we observe the lifetime $T_2 = t_2$ of a similar valve and are lead to a new posterior distribution which is a gamma distribution with parameters:

$$\alpha_3 = \alpha_2 + 1 = \alpha_1 + 2$$
$$\beta_3 = \beta_2 + t_2 = \beta_1 + (t_1 + t_2),$$

and so on.

The posterior density could also have been derived directly because

$$\begin{aligned}\pi(\lambda \mid t_1) &\propto f_{T_1|\Lambda}(t_1 \mid \lambda)\ \pi(\lambda) \\ &\propto \lambda e^{-\lambda t_1}\ \lambda e^{-\lambda} \\ &\propto \lambda^2 e^{-\lambda(t_1+1)}.\end{aligned}$$

Hence,

$$\pi(\lambda \mid t_1) = k(t_1)\ \lambda^2 e^{-\lambda(t_1+1)} \qquad \text{for } t_1 > 0.$$

Because $\pi(\lambda \mid t_1)$ is a density, $k(t_1)$ is easily determined to be $(1 + t_1)^3/2$. This leads to the same posterior density as we derived above.

By repeated arguments, we obtain

$$\begin{aligned} E(\Lambda) &= \frac{2}{1} \\ E(\Lambda \mid T_1 = t_1) &= \frac{2+1}{1+t_1} \quad . \\ E(\Lambda \mid T_1 = t_1, T_2 = t_2) &= \frac{2+1+1}{1+t_1+t_2} \\ &\vdots \end{aligned}$$

Observe how our belief about the mean of Λ is updated, as observations of T become available. □

15.3.4 Homogeneous Poisson Process

Consider an HPP with rate λ, and let $N(t)$ be the number of events in a time interval $(0, t)$. The probability mass function of $N(t)$ is

$$\Pr(N(t) = n \mid \lambda) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \qquad \text{for } n = 0, 1, 2, \ldots .$$

Assume that n_1 failures have been observed during $(0, t)$ such that n_1 and t are known numbers. A gamma(α, β) is again chosen as prior distribution for the (random) parameter Λ. The prior mean is

$$E(\Lambda) = \frac{\alpha}{\beta}.$$

The likelihood function is

$$L(\lambda \mid n_1, t) = \frac{(\lambda t)^{n_1}}{n_1!} e^{-\lambda t} \qquad \text{for } \lambda > 0. \tag{15.23}$$

The posterior density is from (15.4)

$$\begin{aligned} \pi(\lambda \mid n_1, t) &\propto L(\lambda \mid n_1, t)\ \pi(\lambda) \\ &\propto \lambda^{n_1} e^{-\lambda t} \lambda^{\alpha-1} e^{-\beta\lambda} = \lambda^{\alpha+n_1-1} e^{-\lambda(\beta+t)}, \end{aligned} \tag{15.24}$$

which, apart from a constant, is recognized as a gamma distribution with parameters $(\alpha + n_1)$ and $(\beta + t)$. The posterior mean is hence

$$E(\Lambda \mid n_1, t) = \frac{\alpha + n_1}{\beta + t}. \tag{15.25}$$

Example 15.5 (Marginal distribution of $N(t)$)

Consider a plant that has a specified number of identical and independent valves with constant failure rate λ, where λ is a realization of a random variable $\Lambda \sim$ gamma(α, β)

The parameters α and β of the prior distribution are usually "estimated" based on prior experience with the same type of valves, combined with information gained from various reliability data sources (see Chapter 16).

When a valve fails, it is replaced with a valve of the same type. The associated downtime is considered to be negligible. Valve failures are assumed to occur according to an HPP with rate λ. The number of valve failures $N(t) \sim \text{Po}(\lambda t)$.

The *marginal distribution* of $N(t)$ is

$$\begin{aligned} \Pr(N(t) = n) &= \int_0^\infty \Pr(N(t) = n \mid \lambda)\ \pi(\lambda)\ d\lambda \\ &= \int_0^\infty \frac{(\lambda t)^n}{n!}\ e^{-\lambda t} \frac{\beta^\alpha}{\Gamma(\alpha)}\ \lambda^{\alpha-1} e^{-\beta\lambda}\ d\lambda \\ &= \frac{\beta^\alpha t^n}{\Gamma(\alpha) n!} \int_0^\infty \lambda^{\alpha+n-1} e^{-(\beta+t)\lambda}\ d\lambda \\ &= \frac{\beta^\alpha t^n}{\Gamma(\alpha) n!}\ \frac{\Gamma(n+\alpha)}{(\beta+t)^{n+\alpha}} = \frac{\Gamma(n+\alpha)}{\Gamma(\alpha)\Gamma(n+1)} \left(\frac{t}{t+\beta}\right)^n \left(1 - \frac{t}{t+\beta}\right)^\alpha. \end{aligned}$$

When α is an integer and $p = \frac{\beta}{t+\beta}$, the marginal distribution for $N(t)$ can be written

$$\Pr(N(t) = n) = \binom{n+\alpha-1}{n} p^\alpha (1-p)^n, \tag{15.26}$$

which is recognized as the standard *negative binomial distribution* $N(t) \sim$ nbinom(α, p) (see Section 5.8.4). The negative binomial distribution is also defined when α is not an integer. The (prior) marginal mean value of $N(t)$ is

$$E[N(t)] = \frac{\alpha(1-p)}{p} = \frac{\alpha}{\beta}\, t = E(\Lambda)\, t. \tag{15.27}$$

□

15.3.5 Noninformative Prior Distributions

A noninformative prior is a prior distribution that makes all possible values of the parameter θ equally likely. When θ is a probability, the noninformative prior is a uniform distribution over $[0, 1]$ (see Section 15.3.1). When θ can take any positive value, no proper distribution is completely noninformative. In this case, it is common to use a flat prior, that is $\pi(\theta) = k$ for all θ. This prior is not a proper prior because it does not integrate to 1, that is $\int_0^\infty \pi(\theta)\, d\theta \neq 1$. Even so, such a "distribution" may be used because the posterior might still integrate to 1 even if the prior does not.

The posterior distribution is then expressed as

$$\pi(\theta \mid d) \propto L(\theta \mid d)\, \pi(\theta) \propto L(\theta \mid d), \tag{15.28}$$

which means that the posterior distribution is determined solely by the likelihood function.

Example 15.6 (Binomial model)
Let the model distribution be binom(n, θ):

$$\Pr(X = x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}. \tag{15.29}$$

The noninformative prior is a uniform distribution $\pi(\theta) = 1$ for $0 \le \theta \le 1$. The posterior when $X = d$ (where $d \in \{0, 1, 2, \dots, n\}$) is then given by

$$\pi(\theta \mid d) \propto \theta^d (1-\theta)^{n-d}. \tag{15.30}$$

This means that starting with a noninformative prior for θ in the binomial model $X \sim$ binom(n, θ), we get a posterior that is beta distributed with parameters $(d+1)$ and $(n-d+1)$. This is in line with (15.30) because the beta distribution reduces to a (noninformative) uniform distribution when $r = s = 1$. □

Example 15.7 (Exponential model)
The model distribution is

$$f(t \mid \lambda) = \lambda e^{-\lambda t} \qquad \text{for} \;\; t > 0.$$

Let $\pi(\lambda) = 1/k$ be an improper prior distribution for Λ. The posterior for $T = t_1$ is then

$$\pi(\lambda \mid t_1) \propto \lambda e^{-\lambda t_1}, \tag{15.31}$$

which is recognized (apart from a constant) to be a gamma distribution with parameters 2 and t_1. The posterior mean is therefore

$$E(\Lambda \mid t_1) = \frac{2}{t_1}.$$

□

15.4 Bayesian Estimation

When the posterior distribution has been determined, it is often useful to summarize the results with a single point estimate or an interval estimate. A brief introduction to point and interval Bayesian estimation is given in the following.

15.4.1 Bayesian Point Estimation

In Chapter 14, an estimator $\hat{\theta}$ for θ is deemed to be a good estimator when it is *unbiased*, that is $E(\hat{\theta}) = \theta$, and the variance, $\text{var}(\hat{\theta})$, is small. In the Bayesian approach, a *loss function*, $\ell(\hat{\theta}, \theta)$ is used to judge the estimator of the true value θ. The best Bayesian estimator is the estimator that minimizes the expected loss, $E[\ell(\hat{\theta}, \theta)]$.

Many different loss functions may be used. The most common loss functions are

(1) *Squared loss.* $\ell(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$.
(2) *Absolute loss.* $\ell(\hat{\theta}, \theta) =\mid \hat{\theta} - \theta \mid$.

It is also possible to define nonsymmetric loss functions. If, for example, the parameter θ indicates the maximum load to an item and $\hat{\theta}$ is used to determine the required strength of the item, it may be wise to use a loss function as indicated in Figure 15.5. If $\hat{\theta} < \theta$, the item will fail and a certain loss will be incurred. If $\hat{\theta} \ll \theta$, the item may be cheaper to build and the loss may become slightly lower. On the other hand, if $\hat{\theta} > \theta$, a loss related to higher purchase cost will be incurred.

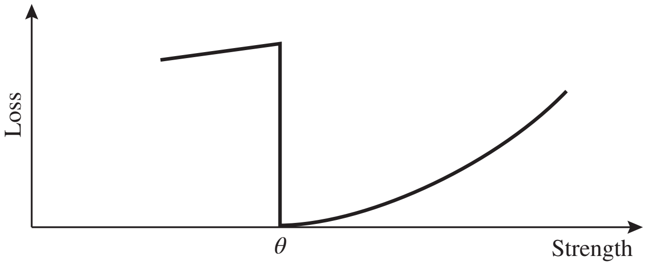


Figure 15.5 Loss function.

In the following, we use the *squared loss function* to exemplify the approach. The Bayesian estimator is determined based on the current knowledge about the parameter θ and can therefore be based on both the prior and the posterior distributions. We illustrate how this is done based on the posterior distribution.

Consider the general setup where θ is a realization of a random variable $\Theta \in \Omega$ with some prior density $\pi(\theta)$, and where X is a random variable with continuous density given $\Theta = \theta, f_{X|\Theta}(x \mid \theta)$. Our task is now to estimate the value θ of Θ that belongs to an observed value x of X. We denote this estimator by $\hat{\theta}(X)$.

As is usual, we prefer an estimator that minimizes the mean squared loss:

$$E[(\hat{\theta}(X) - \Theta)^2].$$

Such an estimator is called a *Bayesian estimator* (of θ) (with minimum expected quadratic loss). Observe that in the Bayesian framework, X and Θ are both random variables. How should $\hat{\theta}(X)$ be chosen?

$$E[(\hat{\theta}(X) - \Theta)^2] = \int_{-\infty}^{+\infty} \int_{\Omega} [\hat{\theta}(X) - \theta]^2 \, f_{X,\Theta}(x, \theta) \; dx \; d\theta.$$

Because $f_{X,\Theta}(x, \theta) = f_{\Theta|X}(\theta \mid x) f_X(x) = \pi(\theta \mid x) f_X(x)$, we get

$$E[(\hat{\theta}(X) - \Theta)^2] = \int_{-\infty}^{+\infty} f_X(x) \left(\int_{\Omega} [\theta - \hat{\theta}(X)]^2 \pi(\theta \mid x) \; d\theta \right) \; dx.$$

Obviously, $E[(\hat{\theta}(X) - \Theta)^2]$ becomes minimized if, for each x, $\hat{\theta}(x)$ is chosen to minimize

$$\int_{\Omega} [\theta - \hat{\theta}(x)]^2 \pi(\theta \mid x) \; d\theta.$$

In probability theory, the following result is well known.

Let Y be a random variable with density $f_Y(y)$ and finite variance τ^2. Then

$$h(\eta) = \int_{-\infty}^{+\infty} (y - \eta)^2 f_Y(y) \; dy \tag{15.32}$$

is minimized when η is chosen as $E(Y)$.

This result, applied to our problem, tells that $E[\hat{\theta}(X) - \Theta]^2$ is minimized for

$$\hat{\theta}(X) = E(\Theta \mid X). \tag{15.33}$$

We can therefore conclude that *the Bayesian estimator of θ is the mean of the posterior distribution of* Θ, when using a squared loss function.

Let us return to our Bayesian model where θ represents a realization of a random variable $\Theta \in \Omega$ with some prior density $\pi_\Theta(\theta)$. We are now considering a situation where our data $(x_1, x_2, \dots, x_n)$ consist of observations of n random variables

$X_1, X_2, \dots, X_n$, assumed to be independent and identically distributed, *conditional on* θ, with density $f_{X|\Theta}(x \mid \theta)$. Then

$$f_{X_1,X_2,\dots,X_n|\Theta}(x_1, x_2, \dots, x_n \mid \theta) = \prod_{j=1}^{n} f_{X|\Theta}(x_j \mid \theta). \tag{15.34}$$

The posterior distribution of Θ, given $X_1, X_2, \dots, X_n$, may now be obtained by the same procedure as we used for a single X, and we get

$$f_{\Theta|X_1,X_2,\dots,X_n}(\theta \mid x_1, x_2, \dots, x_n) \propto \left[\prod_{j=1}^{n} f_{X|\Theta}(x_j \mid \theta)\right] f_{\Theta}(\theta). \tag{15.35}$$

Considering the right-hand side of (15.35) as a function of θ, given $x_1, x_2, \dots, x_n$, this can also be written as

$$f_{\Theta|X_1,X_2,\dots,X_n}(\theta \mid x_1, x_2, \dots, x_n) \propto L(\theta \mid x_1, x_2, \dots, x_n)\pi(\theta), \tag{15.36}$$

where $L(\theta \mid x_1, x_2, \dots, x_n)$ is the likelihood function in the usual meaning.

15.4.2 Credible Intervals

For brevity, let $D = d$ be the data obtained in the experiment (or in the data collection). A *credible interval* is the Bayesian analogue to a confidence interval. A credible interval for Θ, at level $(1 - \varepsilon)$, is an interval $(a(d), b(d))$ such that the conditional probability, given the data d, satisfies

$$\Pr(a(d) \le \Theta \le b(d) \mid d) = \int_{a(d)}^{b(d)} \pi(\theta \mid d)\, d\theta = 1 - \varepsilon. \tag{15.37}$$

The interval $(a(d), b(d))$ is an interval estimate of θ in the sense that the conditional probability of Θ belonging to the interval, given the data, is equal to $1 - \varepsilon$.

As for confidence intervals, the credible intervals are often made symmetrical in the sense that the limits $a(d)$ and $b(d)$ are chosen such that

$$\Pr(\Theta < a(d) \mid d) = \frac{\varepsilon}{2} \quad \text{and} \quad \Pr(\Theta > b(d) \mid d) = \frac{\varepsilon}{2}.$$

Another possibility is to determine the $(1 - \varepsilon)$ credible interval to be the region A of values of Θ that satisfy:

(1) The posterior probability of that region is $(1 - \varepsilon)$, that is, $\Pr(\Theta \in A) = 1 - \varepsilon$.
(2) The minimum posterior density of any point within A is equal to or larger than the posterior density of any point outside A.

The region fulfilling these two requirements is called the *highest posterior density* (HPD) interval. The HPD is an interval in which most of the distribution lies. Some analysts prefer this interval because it is the shortest interval.

15.5 Predictive Distribution

Consider the random variable X with density $f_X(x \mid \theta)$, where θ is considered as a realization of a random variable Θ with prior density $\pi(\theta)$. The marginal density of X is

$$f_X(x) = \int_\Omega f_{X,\Theta}(x, \theta)\, d\theta = \int_\Omega f_X(x \mid \theta)\, \pi(\theta)\, d\theta. \tag{15.38}$$

Some authors (e.g. Gelman et al. 2013) calls this marginal distribution of X the *prior predictive distribution* of X.

Assume that the first experiment has given the result $X = x_0$ and that we are interested in predicting the result of X in the next experiment. This may be done by finding the conditional density of X, given x_0.

$$\begin{aligned} f_X(x \mid x_0) &= \int_\Omega f(x, \theta \mid x_0)\, d\theta = \int_\Omega f(x \mid x_0, \theta)\, \pi(\theta)\, d\theta \\ &= \int_\Omega f(x \mid \theta)\, \pi(\theta \mid x_0)\, d\theta. \end{aligned} \tag{15.39}$$

For this to be correct, we must assume that the experiments are conditionally independent, given θ. This expression is called the *predictive density* of X, given that x_0 has already been observed.

Now, let $x_1, x_2, \dots, x_n$ be n conditionally independent observations of X, given θ. The joint density of $X_1, X_2, \dots, X_n$ and Θ is

$$f_{X,\Theta}(x_1, x_2, \dots, x_n, \theta) = \left[\prod_{i=1}^{n} f_X(x_i \mid \theta)\right] \pi(\theta). \tag{15.40}$$

Let us for brevity denote the data set $x_1, x_2, \dots, x_n$ by d. After having observed the data d, how should we predict the next value of X?

In the same way as for a single data value x_0 above, the *predictive density* becomes

$$f_X(x \mid d) = \int_\Omega f(x \mid \theta)\, \pi(\theta \mid d)\, d\theta. \tag{15.41}$$

Example 15.8 (Exponential distribution)

A total of n identical items are tested to observe the times to failure $T_1, T_2, \dots, T_n$ that are assumed to be exponentially distributed with constant failure rate λ. The times to failure are assumed to be conditionally independent, given $\Lambda = \lambda$. As in Example 15.4, assume that Λ has a gamma prior density with parameters $\alpha = 2$ and $\beta = 1$, such that

$$\pi(\lambda) = \lambda e^{-\lambda} \qquad \text{for } \lambda > 0.$$

Suppose that we have observed the lifetimes $(t_1, t_2, \dots, t_n)$ of n such valves. Based on the arguments in Example 15.4, the posterior density of Λ, given $D = d = \{t_1, t_2, \dots, t_n\}$ is

$$\pi(\lambda \mid d) = \frac{\left(1 + \sum_{i=1}^{n} t_i\right)^{n+2}}{\Gamma(n+2)} \lambda^{n+1} e^{-\lambda\left(1+\sum_{i=1}^{n} t_i\right)} \quad \text{for } \lambda > 0. \tag{15.42}$$

Hence, our guess, based on d, is that the next observation T has (predictive) density

$$\begin{aligned} f_{T|D}(t \mid d) &= \int_0^\infty \lambda e^{-\lambda t} \frac{\left(1 + \sum_{i=1}^{n} t_i\right)^{n+2}}{\Gamma(n+2)} \lambda^{n+1} e^{-\lambda\left(1+\sum_{i=1}^{n} t_i\right)}\, d\lambda \\ &= \frac{\left(1 + \sum_{i=1}^{n} t_i\right)^{n+2}}{\Gamma(n+2)} \int_0^\infty \lambda^{n+2} e^{-\lambda\left[\left(1+\sum_{i=1}^{n} t_i\right)+t\right]}\, d\lambda \\ &= \frac{(n+2)\left(1 + \sum_{i=1}^{n} t_i\right)^{n+2}}{\left(1 + \sum_{i=1}^{n} t_i + t\right)^{n+3}} \quad \text{for } t > 0. \end{aligned} \tag{15.43}$$

Hence, our guess is that the survivor function for a given new valve of the same type is

$$\begin{aligned} \Pr(T > t \mid d) = R(t \mid d) &= \int_t^\infty \frac{(n+2)\left(1 + \sum_{i=1}^{n} t_i\right)^{n+2}}{\left(1 + \sum_{i=1}^{n} t_i + t\right)^{n+3}}\, du \\ &= \left(\frac{1 + \sum_{i=1}^{n} t_i}{1 + \sum_{i=1}^{n} t_i + t}\right)^{n+2} \\ &= \left(1 + \frac{t}{1 + \sum_{i=1}^{n} t_i}\right)^{-(n+2)} \quad \text{for } t > 0. \end{aligned} \tag{15.44}$$

□

15.6 Models with Multiple Parameters

For models with two or more unknown parameters, multidimensional prior distributions must be used and this makes analytic solutions intractable. A wide range of computer programs are available for this type of analysis.

15.7 Bayesian Analysis with R

Within the Bayesian framework, all information about an unknown parameter θ is contained in the posterior distribution. The posterior distribution can be determined in two main ways:

(1) Direct derivation of the posterior, mainly by using conjugate distributions (see above).
(2) Simulation of the posterior
    (a) Sampling from the posterior distribution.
    (b) A Markov chain Monte Carlo (MCMC) approach by using a Gibbs sampler and the Metropolis–Hastings algorithm.

To give a detailed introduction to Bayesian analysis by using MCMC is beyond the scope of this book. Interested readers may find adequate introductions in several other books, such as Albert (2009), Gelman et al. (2013), and Hamada et al. (2008). Here we suffice by briefly mentioning the main approaches.

The programming language BUGS – an acronym for "Bayesian inference using Gibbs sampling" – is the dominating approach for Bayesian analysis by simulation. A main feature of BUGS is to separate the "knowledge base" from the "inference machine" that is used to draw conclusions. BUGS is able to describe rather complex models using very limited syntax. An "expert system" is included in BUGS that can be used to determine an appropriate MCMC scheme for analyzing the specified model. BUGS is thoroughly described in (Lunn et al. 2013). As a programming language, BUGS needs to be implemented into a computer program.

There are three commonly used BUGS implementations:

- WinBUGS (available for MS Windows® computers).
- OpenBUGS (is a native Windows® application, but can run on other platforms by using an emulator (e.g. Wine).
- JAGS – an acronym for "Just Another Gibbs Samples" (is a native application for all major computer platforms).

Among the three, JAGS is often preferred. For any of the three, you have to write the model and the problem to solve in the BUGS language. Each of the engines can be controlled from R. The packages controlling JAGS are, for example called `rjags` and `R2jags` and the R script acts as a frontend to the JAGS/BUGS script. The main programming steps are

(1) Write a BUGS model and save it as a text file.
(2) Open R.
(3) Prepare the inputs for the `R2jags` script and run it.
(4) The model will run in JAGS and you will see the progress and the output in the R terminal/console.

A range of tutorials and examples may be found by searching the Internet. Many WinBUGS examples may, for example, be found in NASA (2009).

The analysis can alternatively be based on Stan, a programming language written in C++. An R interface to Stan is obtained by using the R package `RStan`. Interested readers find numerous tutorials and examples by searching the Internet.

15.8 Problems

15.1 A patient takes a lab test, and the result is A (positive). A positive result (A) indicates that the patient has a special type of cancer (B). It is known that the test returns a correct positive result with probability $\Pr(A \mid B) = 0.97$ and a correct negative result with probability $\Pr(\overline{A} \mid \overline{B}) = 0.95$. Furthermore, evidence indicates that 3% of the population in this age group has this type of cancer, such that our prior belief is $\Pr(B) = 0.03$.

(a) Find the probability that a randomly chosen person is tested with result A (positive).

(b) Find the probability that a patient with a positive test has the special type of cancer?

15.2 Show that the Bayesian estimator of θ, which minimizes the mean absolute error loss $E(\mid \hat{\theta}(X) - \Theta \mid)$ is equal to the *median* of the posterior distribution of Θ (given $X = x$).

15.3 Assume that X has a binomial distribution (n, p), where p represents a realization of a random variable P. The prior distribution of P is $f_P(p) = 1$ for $0 \leq p \leq 1$. Determine the posterior density of P when $X = x$ is observed and determine the Bayesian estimate for p.

15.4 (Based on (Kapur and Lamberson, 1977, p. 402). Seven automobiles are each run over a 36 000 km test schedule. The testing produced a total of 19 failures. Assuming an exponential failure distribution and a gamma prior with parameters $\alpha = 30\ 000$ and $\beta = 3$, answer the following:

(a) What is the Bayesian point estimate for the mean time-to-failure (MTTF)?

(b) What is the 90% lower confidence (credible) limit on the 10 000 km reliability?

15.5 Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed $\mathcal{N}(\theta, \sigma_0^2)$, where σ_0^2 is known, and θ represents a realization of a random variable Θ with normal distribution $\mathcal{N}(\mu_0, \tau_0^2)$, where μ_0 and τ_0^2 are known.

Show that the Bayesian estimate of Θ (minimizing the mean quadratic loss) is a weighed average of the prior mean and the MLE of θ

$$\hat{\theta}(X_1, X_2, \dots, X_n) = \frac{n/\sigma_0^2}{n/\sigma_0^2 + 1/\tau_0^2}\overline{X} + \frac{1/\tau_0^2}{n/\sigma_0^2 + 1/\tau_0^2}\mu_0.$$

Observe that the Bayesian estimate of Θ is a weighed average of the hypothetical estimates of Θ based on the following:

- Data alone (i.e. the standard estimator $\overline{X}$).
- Prior information of Θ but no data, μ_0 (i.e. the Bayesian estimator of μ before any observations are taken).

Again, observe that the influence of the prior mean μ_0 tends to zero as $n \to \infty$.

15.6 Let $X_1, X_2, \dots, X_n$ be independent and identically distributed $\mathcal{N}(0, \sigma^2)$.

(a) Show that the joint density of $X_1, X_2, \dots, X_n$ can be written as

$$C\tau^r e^{-\tau \sum_{i=1}^n x_i^2} \qquad \text{where } r = n/2, \quad \tau = 1/(2\sigma^2).$$

(b) Choose the gamma distribution (k, λ) with density

$$\frac{\lambda}{\Gamma(k)}(\lambda\tau)^{k-1}e^{-\lambda\tau} \qquad \text{for } \tau > 0,$$

as prior density of τ.

Show that the posterior density of τ, given $X_1, X_2, \dots, X_n$ then becomes a gamma distribution $(k + r, \lambda + \sum_{i=1}^n x_i^2)$ with density

$$C(x_1, x_2, \dots, x_n)\tau^{r+k-1}e^{-\tau(\lambda + \sum_{i=1}^n x_i^2)} \qquad \text{for } \tau > 0.$$

(c) Use the result in (b) to show that the Bayesian estimator of σ^2 (with minimum expected quadratic loss) becomes

$$\frac{\lambda + \sum_{i=1}^n X_i^2}{n + 2k - 2}.$$

(*Hint*: Because $2\sigma^2 = 1/\tau$, the Bayesian estimator of $2\sigma^2$ is the posterior expectation of $1/\tau$). This problem is based on an example in Lehmann (1983, p. 246).

15.7 Show that the posterior variance on the average is smaller than the prior variance.

15.8 Explain, as simple as possible, the main differences between a confidence interval and a credible interval.

15.9 Let X have a binomial distribution (n, θ), where θ represents a realization of a random variable Θ with a beta distribution (r, s). Denote the prior mean of Θ by θ_0.

Show that the Bayesian estimate of Θ (minimizing the mean quadratic loss) is a weighed average of the prior mean and the MLE of θ:

$$\hat{\theta}(X) = \frac{n}{r+s+n}\frac{X}{n} + \frac{r+s}{r+s+n}\theta_0.$$

Observe that the Bayesian estimate of Θ is a weighed average of the hypothetical estimates of Θ based on

- Data D alone (i.e. the standard estimator of $\theta, X/n$).
- Prior information of Θ, but no data, θ_0 (i.e. the Bayesian estimator of θ before any observations are taken).

Observe that the influence of the prior mean θ_0 tends to zero as $n \to \infty$.

15.10 Some authors refer to the Bayesian approach to probability as the *Bayesian paradigm*.

(a) Explain what they may mean by referring to this approach as a *paradigm*.

(b) List some advantages obtained by using the Bayesian approach (paradigm).

(c) List some disadvantages related to using the Bayesian approach (paradigm).

(d) List some reasons for the popularity of the Bayesian approach in reliability analyses.

References

Albert, J. (2009). *Bayesian Computation with R*, 2e. Springer.

Gelman, A., Carlin, J.B., Stern, H.S. et al. (2013). *Bayesian Data Analysis*, 3e. Boca Raton, FL: Chapman and Hall.

Hamada, M.S., Wilson, A.G., Reese, C.S., and Martz, H.F. (2008). *Bayesian Reliability*. New York: Springer.

Kapur, K.C. and Lamberson, L.R. (1977). *Reliability in Engineering Design*. Hoboken, NJ: Wiley.

Lunn, D., Jackson, C., Best, N. et al. (2013). *The BUGS Book: A Practical Introduction to Bayesian Analysis*. Boca Raton, FL: Chapman and Hall.

NASA (2009). Bayesian Inference for NASA Probabilistic Risk and Reliability Analysis. *Guide NASA/SP-2009-569*. Washington, DC: U.S. National Aeronautics and Space Administration.

16

Reliability Data: Sources and Quality

16.1 Introduction

By reliability data, we mean estimates of the parameters that enter into our system reliability models, such as failure rates, mean time-to-failures (MTTFs), mean time to repairs (MTTRs), and proof test intervals. To quantify the system reliability, it is necessary to find relevant and realistic estimates for all such parameters. Luckily, there are some databases and prediction methods that can provide some of these estimates. In the following, the term *database* is used to denote any type of data source, from a single table of data to a comprehensive computerized database.

A brief survey of some selected reliability databases is given in this chapter, with focus on databases that are free or commercially available. Quality problems related to reliability databases are briefly discussed.

16.1.1 Categories of Input Data

Quantitative system reliability analyses rely on four main types of input data.

Technical data are needed to understand the functions and the functional requirements and to establish a system model. Technical data are usually supplied by the system vendors.

Operational and environmental data are necessary to define the actual operating context for the system.

Maintenance data, in the form of procedures, resources, quality, and durations, are necessary to establish the system model and to be able to determine the system reliability.

Failure data, that is information about failure modes and failure causes, time-to-failure distributions, and various parameters.

Operational, environmental, and maintenance data are system-specific and can usually not be found in any databases.

System Reliability Theory: Models, Statistical Methods, and Applications, Third Edition.
Marvin Rausand, Anne Barros, and Arnljot Høyland.

Companion Website: www.wiley.com/go/SystemReliabilityTheory3e

Sources of Reliability Data

Reliability data can generally be obtained from the following sources:

(1) Field (i.e. operational) failure event data from the company where the study object is to be used. The failure event data are usually available from the plant's computerized maintenance management system. To provide parameter estimates, the data has to be analyzed by methods as presented in Chapter 14.
(2) Generic reliability databases where the items are classified in broad groups without information about manufacturer, make, and item specifications. OREDA (2015), for example presents estimates for items such as "centrifugal pump; oil processing," "gas turbine; aeroderivative (3000–10,000 kW)," and the like.
(3) Sources providing information about failure modes and failure modes distributions, such as FMD (2016).
(4) Expert judgment is sometimes the only option available to obtain input parameters. The procedure to obtain expert judgments can be more or less structured (e.g. see Meyer and Booker 2001).
(5) Data from manufacturers. These estimates may be based on (i) feedback to the manufacturer from practical use of the items, (ii) engineering analyses of the items, sometimes combined with some test results, (iii) warranty data, and obviously, a combination of all three types.
(6) Reliability prediction models, usually combined with a base case component reliability database, such as MIL-HDBK-217F (1995).
(7) Research reports and papers sometimes present reliability studies of specific items, including the input reliability data.
(8) Data from reliability testing. The testing may be part of the item's qualification process or be available from testing of similar items.

16.1.2 Parameters Estimates

Parameter estimation is dealt with in Chapters 14 and 15. Some brief comments to estimates of the main reliability parameters are given in this section.

Failure Rates

Nearly all the available databases present only constant failure rate estimates. Some databases present failure rates for specific failure modes, whereas other databases present a total failure rate that covers all failure modes. Some few databases provide the number of failures and the operating time on which the estimates are made and also give confidence interval estimates.

CCF Estimates

Common-cause failures (CCFs) are usually modeled by the beta-factor model. The beta-factor is entirely system-specific. Some few databases providing beta-factors are available for nuclear applications. For other applications, the beta-factors usually have to be determined by expert judgment or by using a checklist approach (e.g. see IEC 61508 2010).

Mean Downtime

The estimates of mean downtime (MDT) and MTTR parameters are specific for the particular system and depend on IEEE Std. 352 (2016):

- The physical and mental capabilities of the personnel who operate and maintain the system
- The tools and equipment available for the maintenance action
- The time required to identify and localize the failure
- The time required to isolate the failed part
- The disassembly time
- The availability of spare parts
- The interchange time
- The time to reassemble
- The alignment time
- Checkout time

Most often, these parameters must be provided by expert judgment.

Proof Test Interval and Coverage

The proof test interval for safety items is normally determined from overall safety requirements and should be available as part of the operating procedures. In practice, the proof tests are often adapted to the operational conditions and may therefore vary within an interval covering the stated length of the test interval. The test coverage depends on both the technical properties of the item and the properties of the system where the item is located.

16.2 Generic Reliability Databases

Several generic reliability databases are commercially available as handbooks or computerized databases. This section presents briefly some few of these databases. Most of the databases maintain a website with further descriptions and information about how you can get access to the database.

We start with OREDA that provides reliability data for items used in offshore and onshore oil and gas applications.

16.2.1 OREDA

The OREDA project was initiated by the Norwegian Petroleum Directorate[1] in 1981 to collect and present reliability data for safety equipment used in the Norwegian oil and gas industry. OREDA was later transformed into a joint industry project with international participation. Failure event data are collected from the participating companies and analyzed by a contractor. OREDA has, so far, published six comprehensive data handbooks; in 1984, 1992, 1997, 2002, 2009, and 2015.

Features of OREDA include

- A description of the item and its boundary. The boundary of the item is illustrated by a drawing as shown in Figure 16.1. The lowest level in the system hierarchy at which preventive maintenance is carried out is called *maintainable items*. A list of the maintainable items of the item in question is given, as shown in Table 16.1.
- The number of items from which data have been collected and the number of installations/plants that have supplied data are specified.
- A brief description of the item's operating context is provided.
- Reliability estimates are provided for each failure mode of the item, with 90% confidence intervals. The estimate is denoted "mean," whereas the confidence interval is given by the "lower" and "upper" bounds.

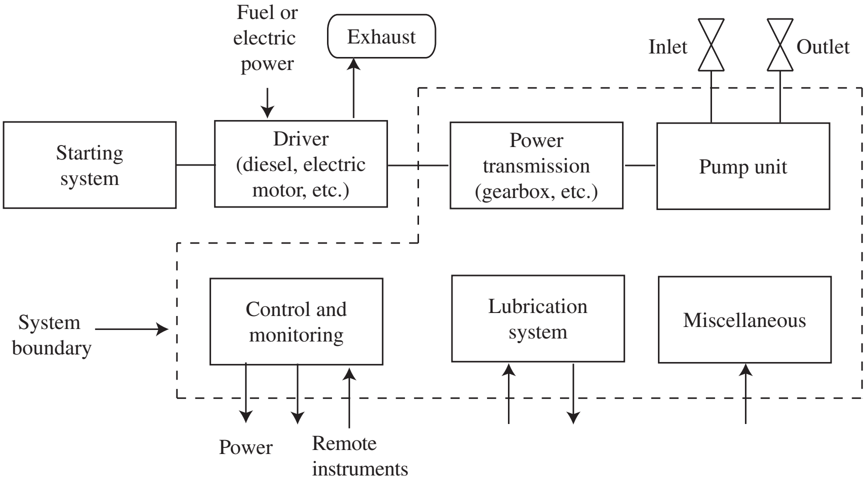


Figure 16.1 Pumps, boundary definition in OREDA.

[1] Now called the Norwegian Petroleum Safety Authority.

Table 16.1 Pumps, subdivision in maintainable items in OREDA.

Pump				
Power transmission	Pump	Control/monitoring	Lubrication	Miscellaneous
– Gearbox/variable drive – Bearing – Seals – Lubrication – Coupling to driver – Coupling to driven unit – Instruments	– Support – Casing – Impeller – Shaft – Radial bearing – Thrust bearing – Seals – Cylinder liner – Piston – Diaphragm – Instruments	– Instruments – Cabling and boxes – Control unit – Actuating device – Monitoring – Internal power supply – Valves	– Instruments – Reservoir with heating system – Pump w/motor – Filter – Cooler – Valves/piping – Oil – Seals	– Purge air – Cooling/heating system – Filter, cyclone – Pulsation damper

- The number of failures for each failure mode and the accumulated time in operation and the accumulated calendar time are specified.
- For on-demand failures, the number of demands is given such that the probability of failure on demand (PFD) can be calculated.
- The active repair time (i.e. the MDT) is estimated for each failure mode together with the number of manhours used for the repair action.

An example of how the data is presented is available on www.oreda.com. The failure data are mainly collected from maintenance records. This means that both item-specific failures and CCFs are included. It also implies that spurious failures such as false alarms may not be included in full detail because such failures not always require a work-order to be corrected. Repair times are recorded whenever possible. For some of the item types, only man-hours were available.

OREDA classifies failure modes in three categories:

(1) *Critical*. A failure that causes immediate and complete loss of a system's capability of providing its output.
(2) *Degraded*. A failure that is not critical, but that prevents the system from providing its output within specifications. Such a failure would usually, but not necessarily, be gradual or partial, and may develop into a critical failure in time.
(3) *Incipient*. A failure that does not immediately cause loss of a system's capability of providing its output, but which, if not attended to, could result in a critical or degraded failure in the near future.

The OREDA handbooks provide data from different time periods, and partly also for different items. This means that data for a particular item may only be found in one of the handbooks.

The OREDA project is still running and is a forum for coordination of reliability data for the oil and gas industry. The detailed data collected during the project is stored in a computerized database that is available to the OREDA Participants. The data in the computerized database is much more detailed that the data presented in the handbooks. The current version of the OREDA handbook is OREDA (2015).

OREDA is often claimed to be the highest quality source of reliability data available and has been a model for other databases. The standard ISO 14224 may be seen as a spin-off of the OREDA project.

Further information about OREDA may be obtained from www.oreda.com.

16.2.2 PDS Data Handbook

The PDS data handbook contains reliability data for items of a safety-instrumented system (SIS) (see Chapter 13). The handbook is based on data from several sources, such as OREDA and vendor data, and subjected to careful expert reviews. The handbook was made to support the PDS method for reliability assessment of SISs, but is a valuable source of reliability data as a stand-alone database. More information is found on https://www.sintef.no/projectweb/pds-main-page/pds-handbooks/pds-data-handbook.

16.2.3 PERD

Process Equipment Reliability Database (PERD) is an ongoing member-based reliability data collection project operated by the Center for Chemical Process Safety (CCPS) of American Institute of Chemical Engineers (AIChE). PERD participants report failures according to a specific taxonomy and in a specified format in line with ISO 14224 (2016).

16.2.4 SERH

Safety Equipment Reliability Handbook (SERH) is an Exida handbook for items in SISs. The handbook has three volumes dedicated to (i) sensors, (ii) logic solvers and interface modules, and (iii) final elements (exida.com 2005).

16.2.5 NPRD, EPRD, and FMD

The data sources Nonelectronic Parts Reliability Data (NPRD), Electronic Parts Reliability Data (EPRD), and Failure Mode Mechanism Distributions (FMD) are supplied by the company Quanterion, through its RMQSI Knowledge Center (www.rmqsi.org). The three sources were earlier developed by the Reliability Information and Analysis Center (RIAC).

NPRD

NPRD provides data for a variety of electrical, mechanical, and electro-mechanical items. The data is a compilation of field experience in military, commercial, and industrial applications. The handbook offers part descriptions, quality level, application environments, point estimates of failure rate, data sources, number of failures, total operating hours, distance, or cycles, and detailed part characteristics. The first edition of NPRD was published in 1978. The most recent version of the handbook is NPRD (2016).

EPRD

EPRD provides reliability estimates for electronic components, such as integrated circuits, discrete semiconductors (diodes, transistors, optoelectronic devices), resistors, capacitors, and inductors/transformers. The estimates are based on failure events in both commercial and military electronic applications. The current version (EPRD 2014) consists of more than 2700 pages and has the same format as NPRD. The handbook provides part descriptions, quality level, application environments, point estimates of failure rate, data sources, number of failures, total operating hours, miles, or cycles, and detailed part characteristics. EPRD is also available in electronic format.

FMD

Failure Mode Mechanism Distributions (FMD) provide field failure mode and mechanism distribution data on a variety of electrical, mechanical, and electromechanical parts and assemblies. The current version (FMD 2016) covers more than 999,000 records. The handbook is also available in electronic format.

Automated Databook

The three data handbooks are also available as an interactive software tool called Quanterion Automated Databook.

16.2.6 GADS

Generating Availability Data System (GADS). This database is operated by the North American Electric Reliability Corporation (NERC). GADS was introduced in 1982, is based on failure and disturbance data from power stations in the United States and Canada, and is a mandatory industry program for conventional generating units over a specified capacity. GADS data consists of three types:

(1) *Design data*. Detailed equipment descriptions.
(2) *Performance data*. Produced power, number of start ups, and so on.
(3) *Event data*. Data related to equipment failures, time, type of outage (forced, maintenance, planned), and so on.

GADS is adapted to IEEE Std 762 (2006) and presents reliability data for total units and major equipment groups. GADS is widely used by industry analysts.

16.2.7 GIDEP

Government Industry Data Exchange Program (GIDEP) is a cooperative information-sharing program between the US government, the Canadian government and industry participants. GIDEP members exchange information about significant problems and nonconforming item data for three main reasons:

(1) To improve safety, reliability, and availability and, at the same time, reduce the development, production, and ownership costs of technical systems.

(2) To ensure that only reliable and conforming parts, material, and software are in use on all government programs.
(3) To avoid the use of counterfeit, known-problem, or discontinued parts and materials.

Further information may be found on www.gidep.org.

16.2.8 FMEDA Approach

An approach combining a reliability database and an analysis to adapt the data is proposed by Exida. Their approach is similar to reliability prediction described in the next section. Exida starts with a failure rate estimate λ_0 from an existing database, such as exida.com (2005) and OREDA (2015). Then a detailed failure modes, effects, and diagnostics analysis (FMEDA) is run to compare the new item to the item covered by the database, to reveal similarities and differences. A proprietary procedure is then used to adjust λ_0 to the new item and to the new operating context. The estimates are then provided to customers. Exida specializes on safety-related equipment, such as sensors and actuating items as discussed in Chapter 13.

16.2.9 Failure Event Databases

Many companies maintain an item failure event database as part of their computerized maintenance management system. Failures and maintenance actions are recorded related to the various items. The data are used in maintenance planning and as a basis for system modifications. In some sectors, the companies are exchanging information recorded in their failure event databases.

Some industries have implemented a failure recording, analysis, and corrective action system (FRACAS) or a defect recording, analysis, and corrective action system (DRACAS). By using FRACAS or DRACAS, failures are formally analyzed, and classified before the reports are stored in the failure report database. Several computer programs supporting FRACAS/DRACAS are available.

16.3 Reliability Prediction

Reliability prediction is the process of forecasting a component's reliability in a given future operating context. The prediction procedures described in this section are mainly used for electronic components, but similar procedures have also been developed for electrical and mechanical components.

For most reliability predictions, a constant failure rate is assumed. Reliability prediction is different from estimation. Estimation (see Chapter 14) deals with

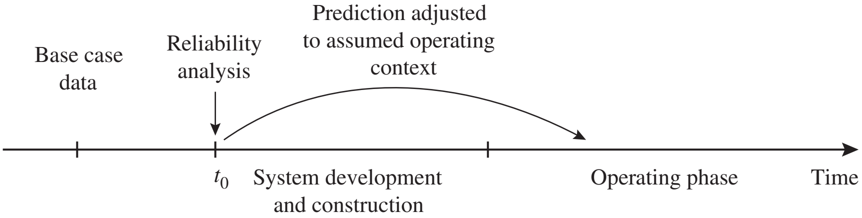

Base case data
Reliability analysis
Prediction adjusted to assumed operating context
t_0
System development and construction
Operating phase
Time

Figure 16.2 Reliability prediction timeline.

quantifying reliability parameters based on an existing dataset, whereas prediction deals with predicting the value of parameters in a future operating context. Very often, we are met with the challenge of predicting the reliability of a new component that has never been used in this operating context. A typical prediction process is illustrated in Figure 16.2, which may be elucidated as follows:

- The component failure rate λ that will be applicable in the (future) operating phase is to be predicted at time t_0.
- Some time before t_0, a base case failure rate λ_0 estimate for a similar component is made available based on data from available sources, expert judgment, or laboratory testing in a controlled environment – or a combination of these.
- The failure rate λ will be used in the system development and construction project, which may be terminated a rather long time after time t_0. The main need of λ is in the early design phase of this project.
- The value of λ to be used must be predicted based on an assumed operating context for the component in the operating phase and also on the technology on which the component may be built.

The analyst applies some procedures to modify λ_0 such that it applies to the stress levels of the given future operating context. Most often, this is accomplished by multiplying λ_0 with a factor $C(\cdot)$, which is a function of the relevant stress levels, such that $\lambda = \lambda_0 C(\text{relevant stress levels})$. The functional form of the factor $C(\cdot)$ varies from approach to approach.

16.3.1 MIL-HDBK-217 Approach

The most common approach for reliability prediction of electronic components is outlined in the military handbook (MIL-HDBK-217F 1995). The handbook provides base case estimates for the constant failure rate λ_0 for various parts used in electronic systems, such as integrated circuits, transistors, diodes, resistors, capacitors, relays, switches, and connectors. The estimates are mainly based on laboratory testing in a controlled base case environment. The failure rates in

MIL-HDBK-217F are hence related only to component-specific (primary) failures. Failures due to external stresses and CCFs are not included. The handbook gives formulas and data to adjust the failure rate of a component to a specified operating context.

Parts Stress

The approach used in MIL-HDBK-217F to predict the failure rate λ in a specified future operating context is called the parts *stress analysis prediction technique* and is based on detailed stress analysis information as well as environment, quality applications, maximum ratings, complexity, temperature, construction, and a number of other application-related factors. The failure rate estimate has the form

$$\lambda_P = \lambda_B \cdot \pi_Q \cdot \pi_E \cdot \pi_A \cdots ,$$

where λ_B is the base case failure rate, that is estimated from reliability tests performed on components under specific and controlled environmental conditions. λ_B is thus given for standardized stresses (e.g. voltage and humidity) and temperature conditions. $\pi_Q, \pi_E, \pi_A, \ldots$ are often called *influence* or *covariate* factors and take into account impact of part quality, equipment environment, application stress, and so on. The values of the basic failure rates and the various factors in the handbook are kept up to date by analysis of failure data on components and systems. The approach does not distinguish between failure modes.

Parts Count

MIL-HDBK-217F describes a special approach for predicting the reliability of a system. The method is called *parts count reliability prediction* and assumes that system success can be achieved only if all the system components are operating, that is, if the system has a series structure. The system failure rate λ_S is obtained by adding the failure rates of the n system components:

$$\lambda_S = \sum_{i=1}^{n} \lambda_i.$$

When the system is not a series system, λ_S gives an upper bound of the failure rate. The parts count method has been heavily criticized (National Research Council 2015, app. D).

The last version of MIL-HDBK-217F was issued in 1995 and has since then not been maintained or updated. It remains a U.S. Department of Defense (DoD) handbook, but Notice 2 of 1995 states that “This handbook for guidance only – Do not cite this document as a requirement.” In spite of this, many producers still adhere to the handbook because it offers a convenient and standard way of estimating reliability.

16.3.2 Similar Methods

Several methods, similar to the MIL-HDBK-217F models have been developed. Among these methods are

Siemens SN 29500 "Electronic Reliability Prediction" is developed by Siemens to be applied to their own products. As for MIL-HDBK-217F, SN 29500 is based on failure rates under specified base case conditions. These failure rates are estimated from application and testing experience and combined with data from external sources, such as MIL-HDBK-217F. Components are categorized into different groups, and each group has a slightly different reliability model. The stress models described in IEC 61709 (2017) are used as a basis for conversion of the failure rate data at reference conditions to the actual operating conditions.

Telcordia SR-322 "Reliability Prediction Procedure for Electronic Equipment" is a reliability prediction method for commercial telecommunication components. Initially, SR-322 was developed by Bellcore because of their dissatisfaction with MIL-HDBK-217F methods applied to commercial products. SR-322 applies three different methods:

- *Method I*. Predictions based on the parts count procedure of MIL-HDBK-217F.
- *Method II*. Predictions based on a combination of parts count and laboratory data.
- *Method III*. Predictions based on a combination of parts count and field data.

FIDES "Reliability Methodology for Electronic Systems" is a French alternative to MIL-HDBK-217F, developed by a consortium of large French companies.

NSWC "Handbook of Reliability Prediction for Mechanical Equipment" (NSWC 2011) is developed for the U.S. DoD by the Naval Surface Warfare Center.

Several computer programs have been developed to support MIL-HDBK 217, Telcordia, and similar databases.

16.4 Common Cause Failure Data

In many reliability studies, the likelihood of CCFs may be more important to estimate than the item failure rates. Very few data sources for CCFs are available and all these are based on the beta-factor model, meaning that it is the value of β that is obtained. There are two main types of data sources:

(1) Data sources based on actual events in the field, such as the *International common cause data exchange* (ICDE) program.

(2) Procedures to predict the value of β based on information about the system and the operating context, such as in IEC 61508 (2010) and IEC 62061 (2005).

16.4.1 ICDE

The ICDE is a project operated by the Nuclear Energy Agency (NEA) on behalf of nuclear industry authorities in several countries. ICDE is focusing mainly on CCF events and how to gain knowledge from these events.

The objectives of the ICDE Project are

- To collect and analyze CCF events to be able to better understand such events, their causes, and their prevention.
- To generate qualitative insights into the causes of CCF events, which can then be used to derive approaches or mechanisms for their prevention or for mitigating their consequences.
- To establish a mechanism for the efficient feedback of experience gained on CCF phenomena, including the development of defenses against their occurrence, such as indicators for risk-based inspections.

The qualitative insights gained from the analysis of CCF events are published in a series of open reports, but the ICDE database is accessible only for the participants of the project.

NRC CCF Insights

The US NRC runs a national project that is similar to ICDE. Data on CCF events are systematically collected and analyzed, stored in a CCF database (NUREG/CR-6268 2007), and insights are published as Insight Summary reports.

16.4.2 IEC 61508 Method

The beta-factor is usually in the range from 1% to 10%. The defenses against CCF events that are actually implemented in the system affect the fraction of CCF events, and estimates of β based on generic data are therefore of limited value.

An approach to estimate β is suggested in IEC 61508-6, Annex D. This method is called the *IEC 61508 method* and is made for SIS hardware failures. The method requires answering 37 predefined questions. The questions are grouped into the following categories:

(1) Physical design (20 questions)
- Separation/segregation (5)
- Diversity/redundancy (9)
- Complexity/design/application/maturity/experience (6)

(2) Analysis (3 questions)
- Assessment/analysis and feedback of data

(3) Human/operator issues (10 questions)
- Procedures/human interface (8)
- Competence/training/safety culture (2)

(4) Environmental issues (4 questions)
- Environmental control (3)
- Environmental testing (1)

The IEC 61508 provides formulas to process the answers and to come up with a estimate of β. The method is explained in detail in IEC 61508 (2010) and also by Rausand (2014).

16.5 Data Analysis and Data Quality

A significant effort has been devoted to the collection and processing of reliability data during the last 40 years. Despite this great effort, the quality of the data available is still not good enough. The quality of the data presented in the databases obviously depends on the way the data are collected and analyzed – and on the competence of the persons who classify and analyze the data.

Several standards and guidelines have been issued to obtain high quality in data collection and analysis. Among these are

- IEC 60300-3-2 (2004) *Dependability management. Part 3-2: Application guide – Collection of dependability data from the field.*
- ISO 14224 *Petroleum and natural gas industries – Collection and exchange of reliability and maintenance data for equipment*. This standard may be considered as a spin-off of the OREDA project.
- *Guidelines for Improving Plant Reliability Through Data Collection and Analysis* (CCPS 1998).
- *Reliability Data Quality Handbook* (ESReDA 1999).

In the following, we discuss briefly some main problems related to data analysis and reliability databases.

16.5.1 Outdated Technology

A typical case for collection and use of reliability data is illustrated in Figure 16.3, which may be interpreted as follows:

- The items considered are often so reliable that they have to be observed for a long period to give enough failures to provide meaningful estimates. The data collection must therefore cover a rather long time interval even if we observe a fair number of identical items – usually several years.

- The data collection is often organized as a project with a specified start and stop. This means that many of the observed items were not new when the observation period started.
- After the data collection project is terminated, there is usually a "waiting time" where the data are classified and checked for quality and consistency. Next, the data are analyzed and reliability estimates are provided.
- When the estimates are to be used as input to a new system development, the estimates are needed in the first design phases of the project. The system development project may sometimes take a long time, up to several years before the items in question are installed and ready for use.

Figure 16.3 indicates that when the system items are purchased or built (based on current technology) and installed, their reliability assessments may have been based on estimates for items based a much older, and quite different technology. The OREDA project has shown that some of the items – from which reliability estimates are made – were installed 20–30 years ago. It is then pertinent to ask: Is the technology used in these old items sufficiently similar to the technology of the new items that are to be installed?

16.5.2 Inventory Data

Field data are typically collected from maintenance records. Failures that require a maintenance task are usually recorded, but false alarms and temporary failures may not be recorded in the maintenance files. Another challenge in field data collection is to cover the total *inventory*. We need to find the answer to questions such as:

- How many items of this particular type do we have in the plant?
- How big percentage of the time is each item loaded and in operation?
- What is the operating context for each item?

16.5.3 Constant Failure Rates

Almost all commercially available reliability databases provide only constant failure rates, even for mechanical equipment that degrade due to mechanisms such

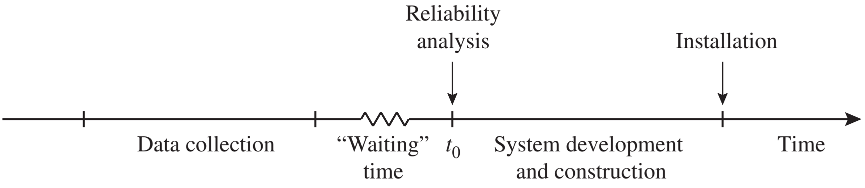


Figure 16.3 Estimates from field data sources.

as erosion, corrosion, and fatigue. Based on knowledge about the deteriorating mechanisms, the failure rate of such equipment should be increasing. The data available for the analysis is usually the number n of failures during a total time t in service. The failure rate estimated by n/t is thus an "average failure rate." The failure data are usually collected from a rather limited time period, that may be called the *observation window* – see Section 16.5.1.

Assume that the failed items are replaced, or restored to an as-good-as-new condition, such that we have a renewal process. A number of items are observed during a specified observation window. The observation window may, for example be from 1 January 2000 till 1 January 2003. In this period, we only record the number (n) of failures and the accumulated time (t) in service. A constant failure rate λ is estimated by $\widehat{\lambda} = n/t$. If the (real) life distribution is a Weibull distribution with an increasing failure rate function, $z(t)$, and we use a constant failure rate estimate, we overestimate the failure rate in the early phase of the item's life, and underestimate the failure rate in the last part of its life. This is illustrated in Figure 16.4. The result will especially be wrong if we extrapolate the estimated constant failure rate beyond the time interval where we have collected data.

People who analyze life data are not always aware of the difference between the concepts failure rate function (force of mortality [FOM]), and rate of occurrence of failures (ROCOF) as discussed in Chapters 5 and 10. Assume that we have a system with an increasing ROCOF, $w(t)$. If we collect failure data in an observation window in an early phase of the system's life, the resulting "average failure rate" is often very different from what we would get in a later observation window. This is illustrated in Figure 16.5.

This effect has been seen in several offshore data collection projects, for example for downhole safety valves. When a valve has failed, it has been replaced with a new valve of the same type, and we have (erroneously) believed that we had a

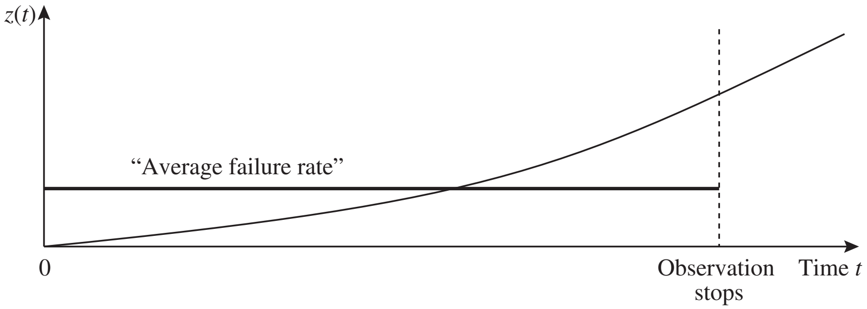


Figure 16.4 The real failure rate and the erroneously estimated constant failure rate.

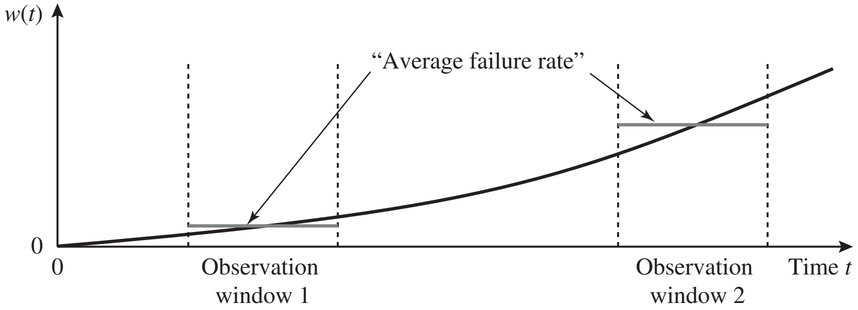


Figure 16.5 Average failure rates estimated in two different observation windows.

renewal process. The environmental conditions in the well had, however, changed with time and produced a more hostile environment.

16.5.4 Multiple Samples

In generic databases, failure rate estimates for generic items are presented. The individual items that are classified within the same generic class of items do not need to be identical and do not need to be exposed to exactly the same operating context. The data collected is therefore not a homogeneous sample.

Assume that we have m samples of failure data and that each sample is homogeneous. It is, however, not certain that all the m samples are homogeneous. Sample i consists of n_i recorded failures during a total time in operation t_i. The items in this sample are assumed to have constant failure rate λ_i, for $i = 1, 2, \dots, m$. The failure rate λ_i can be estimated by

$$\widehat{\lambda}_i = \frac{n_i}{t_i},$$

and a 90% confidence is given by (10.16)

$$\left(\frac{1}{2t_i}\, z_{0.95,2n_i},\ \frac{1}{2t_i}\, z_{0.05,2(n_i+1)} \right).$$

The estimates and the confidence intervals for the m samples are illustrated in Figure 16.6.

If we (erroneously) assume that all samples have the same failure rate λ, the estimate would be

$$\widehat{\lambda} = \frac{\sum_{i=1}^{m} n_i}{\sum_{i=1}^{m} t_i}. \tag{16.1}$$

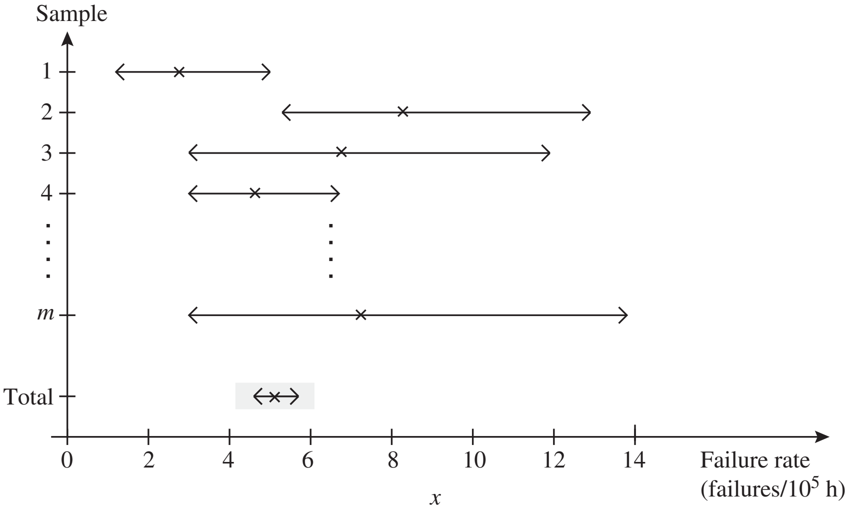


Figure 16.6 Estimates and confidence intervals for inhomogeneous samples.

Because the total number of failures is relatively large, and the total time in operation is relatively long, the confidence interval is rather short, as illustrated by "total" in Figure 16.6. It is seen from Figure 16.6 that the "total" confidence interval does not reflect the uncertainty of the failure rates.

We should therefore carefully check that the samples are homogeneous before we merge them. In many databases, the samples are merged without any checking. In OREDA (2015), an alternative approach is used. The failure rate λ is assumed to be a random variable, that can take different values for the different samples. An estimate of the standard deviation (SD) of the distribution of λ is presented together with the failure rate estimates for each failure mode. A high value of SD indicates that the samples are inhomogeneous. The (average) failure rate is estimated as a weighted average of the failure rate estimates for each sample, following a semi-Bayesian approach. The approach is described in detail in Lydersen and Rausand (1989) and in the OREDA documentation.

Another approach to handle inhomogeneous samples is presented in Molnes et al. (1986),, where failure data from safety valves in oil wells are analyzed. The valves are installed in wells with different characteristics called *stressors*. The stressors are factors such as flowrate, gas/oil ratio, CO_2 content, H_2S content, and sand content. Some main valve characteristics, like diameter, and equalizing principle, are also defined as stressors. The failure rate is modeled as a function of the stressors, as proportional hazards models, and analyzed by Cox regression, as mentioned in Chapter 14. In this case, we obtain estimates based on a physical modeling of the differences between the samples.

16.5.5 Data From Manufacturers

Some manufacturers provide reliability data for their equipment. These stem mainly from reported failures by customers and warranty claims, sometimes supplemented by laboratory tests. Manufacturers seldom get any information from users after the warranty period is over. Some manufacturers therefore offer service schemes to obtain such data. A general problem with this type of data is that we can never be sure that all failures have been reported, neither does we have full information about how long time the items have been in operation. Data for manufacturers can provide some information, but analysts should be careful not to put too much confidence in failure rate estimates.

16.5.6 Questioning the Data Quality

When using data from a reliability database to predict the reliability of a particular item, one should at least consider the following questions:

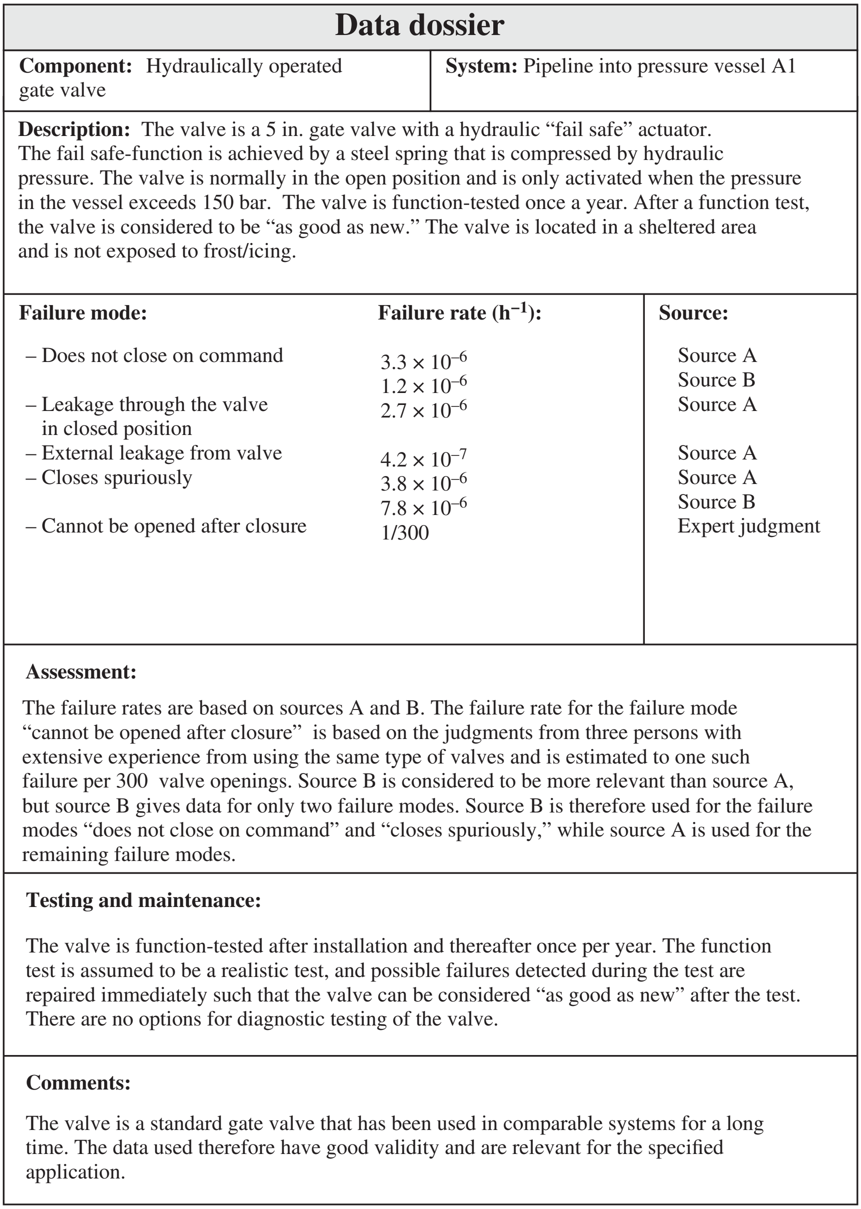
- Does the data originate from the same type of items?
- Has this type of items recently been subject to significant changes of technology or materials?
- Is the operating context the same or similar?
- Is the data source based on a big enough set to give a trustworthy estimate?

16.6 Data Dossier

When performing a reliability analysis, it is important to document all the input parameters that are used in the calculations. It is therefore recommended that a *data dossier* be set up that presents and justifies the choice of data for each element or channel of the system. An example of such a data dossier is shown in Figure 16.7. In many applications, a simpler data dossier may be used.

16.6.1 Final Remarks

Collection and analysis of field data are often difficult tasks, where it is easy to make mistakes. Further information about reliability databases and associated problems may be found in Flamm and Luisi (1992) and Cooke and Bedford (2002).



Data dossier

Component: Hydraulically operated gate valve	**System:** Pipeline into pressure vessel A1

Description: The valve is a 5 in. gate valve with a hydraulic "fail safe" actuator. The fail safe-function is achieved by a steel spring that is compressed by hydraulic pressure. The valve is normally in the open position and is only activated when the pressure in the vessel exceeds 150 bar. The valve is function-tested once a year. After a function test, the valve is considered to be "as good as new." The valve is located in a sheltered area and is not exposed to frost/icing.

Failure mode:	**Failure rate (h^{-1}):**	**Source:**
– Does not close on command	3.3×10^{-6}	Source A
	1.2×10^{-6}	Source B
– Leakage through the valve in closed position	2.7×10^{-6}	Source A
– External leakage from valve	4.2×10^{-7}	Source A
– Closes spuriously	3.8×10^{-6}	Source A
	7.8×10^{-6}	Source B
– Cannot be opened after closure	1/300	Expert judgment

Assessment:

The failure rates are based on sources A and B. The failure rate for the failure mode "cannot be opened after closure" is based on the judgments from three persons with extensive experience from using the same type of valves and is estimated to one such failure per 300 valve openings. Source B is considered to be more relevant than source A, but source B gives data for only two failure modes. Source B is therefore used for the failure modes "does not close on command" and "closes spuriously," while source A is used for the remaining failure modes.

Testing and maintenance:

The valve is function-tested after installation and thereafter once per year. The function test is assumed to be a realistic test, and possible failures detected during the test are repaired immediately such that the valve can be considered "as good as new" after the test. There are no options for diagnostic testing of the valve.

Comments:

The valve is a standard gate valve that has been used in comparable systems for a long time. The data used therefore have good validity and are relevant for the specified application.

Figure 16.7 Example of a reliability data dossier.

References

CCPS (1998). *Guidelines for Improving Plant Reliability through Data Collection and Analysis*. New York: Center for Chemical Process Safety of the American Institute of Chemical Engineers.

Cooke, R.M. and Bedford, T. (2002). Reliability databases in perspective. *IEEE Transactions on Reliability* 51 (3): 294–310.

EPRD (2014). Electronic Parts Reliability Data. *Handbook EPRD 2014*. Utica, NY: Quanterion Solutions Inc.

ESReDA (1999). Handbook on Quality of Reliability Data. Working group report. Statistical Series No. 4. Høvik, Norway: European Reliability Data Association, DNV-GL.

exida.com (2005). *Safety Equipment Reliability Handbook*, 4e. Sellersville, PA: exida.com.

Flamm, J. and Luisi, T. (eds.) (1992). *Reliability Data Collection and Analysis*. Deventer: Kluwer Academic Publishers.

FMD (2016). Failure Mode and Mechanism Distributions. *Handbook FMD 2016*. Utica, NY: Quanterion Solutions Inc.

IEC 60300-3-2 (2004). *Dependability management. Part 3-2: Application guide–collection of dependability data from the field*, *International standard*. Geneva: International Electrotechnical Commission.

IEC 61508 (2010). *Functional safety of electrical/electronic/programmable electronic safety-related systems*. Parts 1-7, *International standard*. Geneva: International Electrotechnical Commission.

IEC 61709 (2017). *Electric components–reliability–reference conditions for failure rates and stress models for conversion*, *International standard*. Geneva: International Electrotechnical Commission.

IEC 62061 (2005). *Safety of machinery–functional safety of safety-related electrical, electronic and programmable electronic control systems*, *International standard*. Geneva: International Electrotechnical Commission.

IEEE Std. 352 (2016). *IEEE guide for general principles of reliability analysis of nuclear power generating station protection station systems and other nuclear facilities*, *Standard*. New York: Institute of Electrical and Electronics Engineers.

IEEE Std 762 (2006). *Standard definitions for use in reporting electric generating unit reliability, availability, and productivity*, *Standard*. New York: Institute of Electrical and Electronics Engineers.

ISO 14224 (2016). *Petroleum, petrochemical, and natural gas industries: collection and exchange of reliability and maintenance data for equipment*, *International standard*. Geneva: International Organization for Standardization.

Lydersen, S. and Rausand, M. (1989). Failure rate estimation based on data from different environments and with varying quality. In: *Reliability Data Collection* *and Use in Risk and Availability Assessment* (ed. V. Colombari). Springer. 243–255.

Meyer, M.A. and Booker, J.M. (2001). *Eliciting and Analyzing Expert Judgment*. Philadelphia, PA: SIAM.

MIL-HDBK-217F (1995). Reliability Prediction of Electronic Equipment. *Military Handbook*. Washington, DC: U.S. Department of Defense.

Molnes, E., Rausand, M., and Lindqvist, B.H. (1986). Reliability of Surface Controlled Subsurface Safety Valves. *Technical Report STF75A86024*. Trondheim, Norway: SINTEF.

National Research Council (2015). *Reliability Growth: Enhancing Defense System Reliability*. Washington, DC: The National Academies Press.

NPRD (2016). Nonelectronic Parts Reliability Data. *Handbook NPRD 2016*. Utica, NY: Quanterion Solutions Inc.

NSWC (2011). Handbook of Reliability Prediction Procedures for Mechanical Equipment. *Handbook NSWC-11*. West Bethesda, ML: Naval Surface Warfare Center, Carderock Division.

NUREG/CR-6268 (2007). Common-Cause Failure Database and Analysis System: Event Data Collection, Classification, and Coding. *Report NUREG/CR-6268*. Washington, DC: U.S. Nuclear Regulatory Commission.

OREDA (2015). *Offshore and Onshore Reliability Data*, 6e. 1322 Høvik, Norway: OREDA Participants, DNV GL.

Rausand, M. (2014). *Reliability of Safety-Critical Systems: Theory and Applications*. Hoboken, NJ: Wiley.

Appendix A

Acronyms

The main abbreviations and acronyms that are used in the book are listed in Appendix A. Abbreviations that are used only once, and where the meaning of the abbreviation is spelled out are not included.

AIChE	American Institute of Chemical Engineers
AFT	accelerated failure time
AMSAA	U.S. Army Material Systems Analysis Activity
ARA	arithmetic reduction of age
ARI	arithmetic reduction of intensity
ARINC	Aeronautical Radio, Incorporated
ARMA	auto-regressive moving average
BDD	binary decision diagram
BN	Bayesian network
BPM	basic parameter model
CCF	common-cause failure
CCPS	Center for Chemical Process Safety (of AIChE)
CDF	core damage frequency
CBM	condition-based maintenance
CM	corrective maintenance
CMMS	computerized maintenance management system
CONOPS	concept of operations
CPT	conditional probability table
CVS	comma-separated values
DAG	directed acyclic graph
DFR	decreasing failure rate
DFRA	decreasing failure rate average
DFT	dynamic fault tree
DIM	differential importance metric

System Reliability Theory: Models, Statistical Methods, and Applications, Third Edition.
Marvin Rausand, Anne Barros, and Arnljot Høyland.

Companion Website: www.wiley.com/go/SystemReliabilityTheory3e

DoD	Department of Defense
DRACAS	defect recording, analysis, and corrective action system
EDA	exploratory data analysis
E/E/PE	electrical/electronic/programmable electronic
EN	European norm
EPRD	electronic parts reliability data
ESD	emergency shutdown
ESDV	emergency shutdown valve
ESReDa	European Safety, Reliability & Data Association
ETA	event tree analysis
EUC	equipment under control
FAR	fatal accident rate
FAST	functional analysis system technique
FFA	functional failure analysis
FMD	failure mode mechanism distributions
FMEA	failure modes and effects analysis
FMECA	failure modes, effects, and criticality analysis
FMEDA	failure modes, effects, and diagnostics analysis
FOM	force of mortality
FRACAS	failure reporting analysis and corrective action system
FSI	functional significant item
FTA	fault tree analysis
FTF	fail to function
GADS	generating availability data system
GIDEP	government industry data exchange program
HAZOP	hazard and operability (study)
HPD	highest posterior density
HPP	homogeneous Poisson process
ICDE	international common cause data exchange
IDEF	integrated definition language
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronic Engineers
IEV	international electrotechnical vocabulary
IFR	increasing failure rate
IFRA	increasing failure rate average
i.i.d.	independent and identically distributed
IP	improvement potential
ISO	International Organization for Standardization
KTT	kinetic tree theory
LCC	life cycle cost
LOPA	layer of protection analysis

MBF	multiple beta-factor (model)
MCPS	minimal cut parallel structure
MCMC	Markov chain Monte Carlo
MDT	mean downtime
MFDT	mean fractional deadtime
MFSC	multiple failures with a shared cause
MGL	multiple Greek letter (model)
MLE	maximum likelihood estimator
MME	method of moments estimator
MRL	mean residual lifetime
MSG	maintenance steering group
MTBF	mean time between failures
MTBR	mean time between replacements/renewals
MTTF	mean time to failure
MTTFF	mean time to first failure
MTTR	mean time to repair
MUT	mean up-time
NBU	new better than used
NBUE	new better than used in expectation
NEA	nuclear energy agency
NERC	North American Electric Reliability Corporation
NHPP	nonhomogeneous Poisson process
NPRD	non-electronic parts reliability data
NRC	Nuclear Regulatory Commission (U.S.)
NTNU	Norwegian University of Science and Technology
NUREG	Title of reports from the U.S. NRC
NWU	new worse than used
OEE	overall equipment efficiency
OREDA	offshore and onshore reliability data
PDMP	piecewise-deterministic Markov process
PERD	process equipment reliability database
PFD	probability of failure on demand
PH	proportional hazards
PHM	prognostics and health management
PM	preventive maintenance
PRA	probabilistic risk assessment
PSA	probabilistic safety assessment
PSV	pressure safety valve
QRA	quantitative risk analysis
RAM	reliability, availability, and maintainability
RAMS	reliability, availability, maintainability, and safety

RAW	risk achievement worth
RRW	risk reduction worth
RBD	reliability block diagram
RCM	reliability centered maintenance
RIAC	reliability information analysis center
ROCOF	rate of occurrence of failures
RUL	remaining useful lifetime
SADT	structured analysis and design technique
SAE	The Engineering Society For Advancing Mobility in Land Sea Air and Space
SERH	safety equipment reliability handbook
SIF	safety instrumented function
SIL	safety integrity level
SIS	safety instrumented system
SRP	superimposed renewal process
SRS	system reliability services
TPM	total productive maintenance
TRL	technology readiness level
TRP	trend renewal process
TQM	total quality management
UKAEA	United Kingdom Atomic Energy Authority

Appendix B

Laplace Transforms

Let $f(t)$ be a function that is defined on the interval $(0, \infty)$. The Laplace[1] transform $f^*(s)$ of the function $f(t)$ is defined by

$$f^*(s) = \int_0^\infty e^{-st} f(t)\, dt, \tag{B.1}$$

where s is a real number and if the integral exists. In more advanced treatments of the Laplace transform, s is permitted to be a complex number. All functions do not have a Laplace transform. For instance, if $f(t) = \exp(t^2)$, the integral diverges for all values of s.

The Laplace transform of $f(t)$ is also written as $\mathcal{L}[f(t)]$:

$$\mathcal{L}[f(t)] = f^*(s) = \int_0^\infty e^{-st} f(t)\, dt, \tag{B.2}$$

to indicate the relationship between the functions f and f^*. When $f(t)$ is the probability density function of a nonnegative random variable T, the Laplace transform of $f(t)$ is seen to be equal to the expected value of the random variable e^{-sT}.

$$E(e^{-sT}) = \int_0^\infty e^{-st} f(t)\, dt = f^*(s).$$

The function $f(t)$ is called the inverse Laplace transform of $f^*(s)$ and is written

$$f(t) = \mathcal{L}^{-1}[f^*(s)]. \tag{B.3}$$

Theorem B.1

Let $f(t)$ be a function that is piecewise continuous on every finite interval in the range $t \geq 0$ and satisfies

$$|f(t)| \leq M\, e^{\alpha t} \qquad \text{for all } t \geq 0,$$

[1] Named after the French mathematician Pierre-Simon Laplace (1749–1827).

System Reliability Theory: Models, Statistical Methods, and Applications, Third Edition.
Marvin Rausand, Anne Barros, and Arnljot Høyland.

Companion Website: www.wiley.com/go/SystemReliabilityTheory3e

and for some constants α and M. Then the Laplace transform of $f(t)$ exists for all $s > \alpha$. □

Example B.1

Consider the function $f(t) = e^{\alpha t}$, where α is a constant. We have

$$\begin{aligned} f^*(s) &= \int_0^\infty e^{-st} e^{\alpha t}\, dt = \int_0^\infty e^{-t(s-\alpha)}\, dt \\ &= \lim_{\tau\to\infty} \left[\frac{-1}{s-\alpha} e^{-t(s-\alpha)} \right]_0^\tau \\ &= \frac{1}{s-\alpha} \qquad \text{for } s > \alpha. \end{aligned}$$

Thus

$$\mathcal{L}[e^{\alpha t}] = \frac{1}{s-\alpha} \qquad \text{when} \quad s > \alpha.$$

□

B.1 Important Properties of Laplace Transforms

Table B.1 lists some important properties of the Laplace transform. Proofs may be found in many standard textbooks on mathematical analysis.

B.2 Laplace Transforms of Some Selected Functions

Table B.2 lists the Laplace transforms of some selected functions.

You will find a lot more about Laplace transforms by searching the Internet.

Table B.1 Some main properties of Laplace transforms.

(1) $\mathcal{L}[f(_1(t) + f_2(t)] = \mathcal{L}[f_1(t)] + \mathcal{L}[f_2(t)]$
(2) $\mathcal{L}[\alpha f(t)] = \alpha\mathcal{L}[f(t)]$
(3) $\mathcal{L}[f(t-\alpha)] = e^{-\alpha s}\mathcal{L}[f(t)]$
(4) $\mathcal{L}[e^{\alpha t} f(t)] = f^*(s-\alpha)$
(5) $\mathcal{L}[f'(t)] = s\mathcal{L}[f(t)] - f(0)$
(6) $\mathcal{L}[\int_0^t f(u)\, du] = \frac{1}{s}\mathcal{L}[f(t)]$
(7) $\mathcal{L}[\int_0^t f_1(t-u) f_2(u)\, du] = \mathcal{L}[f_1(t)] \cdot \mathcal{L}[f_2(t)]$
(8) $\lim_{s\to\infty} sf^*(s) = \lim_{t\to 0} f(t)$
(9) $\lim_{s\to 0} sf^*(s) = \lim_{t\to\infty} f(t)$

Table B.2 Some Laplace transforms.

$f(t),\ t \geq 0$	$f^*(s) = \mathcal{L}[f(t)]$
1	$\dfrac{1}{s}$
t	$\dfrac{1}{s^2}$
t^2	$\dfrac{2!}{s^3}$
t^n	$\dfrac{n!}{s^{n+1}}$ for $\alpha > -1$ for $n = 0, 1, 2, \dots$
t^α	$\dfrac{\Gamma(\alpha+1)}{s^{\alpha+1}}$ for $\alpha > -1$
$e^{\alpha t}$	$\dfrac{1}{s-\alpha}$
$e^{\alpha t} t^n$	$\dfrac{n!}{(s-\alpha)^{n+1}}$
$\cos \omega t$	$\dfrac{s}{s^2+\omega^2}$
$\sin \omega t$	$\dfrac{\omega}{s^2+\omega^2}$
$\cosh \alpha t$	$\dfrac{s}{s^2-\alpha^2}$
$\sinh \alpha t$	$\dfrac{\alpha}{s^2-\alpha^2}$

Author Index

Authors who are explicitly referred to are listed. Additional authors may be "hidden" in technical reports issued by organizations. We have tried to provide full names, but for many authors, full names have not been traceable.

System Reliability Theory: Models, Statistical Methods, and Applications, Third Edition.
Marvin Rausand, Anne Barros, and Arnljot Høyland.

Companion Website: www.wiley.com/go/SystemReliabilityTheory3e

Subject Index

A number of terms are explicitly defined in the book. These terms are indicated by a boldfaced page number. Terms that are treated in longer sections and/or paragraphs are mainly referred to at the first –or the most important –occurrence.

System Reliability Theory: Models, Statistical Methods, and Applications, Third Edition.
Marvin Rausand, Anne Barros, and Arnljot Høyland.

Companion Website: www.wiley.com/go/SystemReliabilityTheory3e

WILEY SERIES IN PROBABILITY AND STATISTICS

ESTABLISHED BY WALTER A. SHEWHART AND SAMUEL S. WILKS

The ***Wiley Series in Probability and Statistics*** is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches.

This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

† ABRAHAM and LEDOLTER · Statistical Methods for Forecasting
AGRESTI · Analysis of Ordinal Categorical Data, *Second Edition*
AGRESTI · An Introduction to Categorical Data Analysis, *Second Edition*
AGRESTI · Categorical Data Analysis, *Second Edition*
ALTMAN, GILL, and McDONALD · Numerical Issues in Statistical Computing for the Social Scientist
AMARATUNGA and CABRERA · Exploration and Analysis of DNA Microarray and Protein Array Data
AND\BR\EL · Mathematics of Chance
ANDERSON · An Introduction to Multivariate Statistical Analysis, *Third Edition*
* ANDERSON · The Statistical Analysis of Time Series
ANDERSON, AUQUIER, HAUCK, OAKES, VANDAELE, and WEISBERG · Statistical Methods for Comparative Studies
ANDERSON and LOYNES · The Teaching of Practical Statistics
ARMITAGE and DAVID (editors) ·Advances in Biometry
ARNOLD, BALAKRISHNAN, and NAGARAJA · Records
* ARTHANARI and DODGE · Mathematical Programming in Statistics
* BAILEY · The Elements of Stochastic Processes with Applications to the Natural Sciences
BAJORSKI · Statistics for Imaging, Optics, and Photonics
BALAKRISHNAN and KOUTRAS · Runs and Scans with Applications

* Now available in a lower priced paperback edition in the Wiley Classics Library
† Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

BALAKRISHNAN and NG · Precedence-Type Tests and Applications
BARNETT · Comparative Statistical Inference, *Third Edition*
BARNETT · Environmental Statistics
BARNETT and LEWIS · Outliers in Statistical Data, *Third Edition*
BARTHOLOMEW, KNOTT, and MOUSTAKI · Latent Variable Models and Factor Analysis: A Unified Approach, *Third Edition*
BARTOSZYNSKI and NIEWIADOMSKA-BUGAJ · Probability and Statistical Inference, *Second Edition*
BASILEVSKY · Statistical Factor Analysis and Related Methods: Theory and Applications
BATES and WATTS · Nonlinear Regression Analysis and Its Applications
BECHHOFER, SANTNER, and GOLDSMAN · Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons
BEIRLANT, GOEGEBEUR, SEGERS, TEUGELS, and DEWAAL · Statistics of Extremes: Theory and Applications
BELSLEY · Conditioning Diagnostics: Collinearity and Weak Data in Regression
† BELSLEY, KUH, and WELSCH · Regression Diagnostics: Identifying Influential Data and Sources of Collinearity
BENDAT and PIERSOL · Random Data: Analysis and Measurement Procedures, *Fourth Edition*
BERNARDO and SMITH · Bayesian Theory
BHAT and MILLER · Elements of Applied Stochastic Processes, *Third Edition*
BHATTACHARYA and WAYMIRE · Stochastic Processes with Applications
BIEMER, GROVES, LYBERG, MATHIOWETZ, and SUDMAN · Measurement Errors in Surveys
BILLINGSLEY · Convergence of Probability Measures, *Second Edition*
BILLINGSLEY · Probability and Measure, *Anniversary Edition*
BIRKES and DODGE · Alternative Methods of Regression
BISGAARD and KULAHCI · Time Series Analysis and Forecasting by Example
BISWAS, DATTA, FINE, and SEGAL · Statistical Advances in the Biomedical Sciences: Clinical Trials, Epidemiology, Survival Analysis, and Bioinformatics
BLISCHKE and MURTHY (editors) · Case Studies in Reliability and Maintenance
BLISCHKE and MURTHY · Reliability: Modeling, Prediction, and Optimization
BLOOMFIELD · Fourier Analysis of Time Series: An Introduction, *Second Edition*
BOLLEN · Structural Equations with Latent Variables
BOLLEN and CURRAN · Latent Curve Models: A Structural Equation Perspective
BOROVKOV · Ergodicity and Stability of Stochastic Processes
BOSQ and BLANKE · Inference and Prediction in Large Dimensions
BOULEAU · Numerical Methods for Stochastic Processes
* BOX· Bayesian Inference in Statistical Analysis
BOX · Improving Almost Anything, *Revised Edition*
* BOX and DRAPER · Evolutionary Operation: A Statistical Method for Process Improvement

* Now available in a lower priced paperback edition in the Wiley Classics Library
† Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

BOX and DRAPER · Response Surfaces, Mixtures, and Ridge Analyses, *Second Edition*

BOX, HUNTER, and HUNTER · Statistics for Experimenters: Design, Innovation, and Discovery, *Second Editon*

BOX, JENKINS, and REINSEL · Time Series Analysis: Forcasting and Control, *Fourth Edition*

BOX, LUCEÑO, and PANIAGUA-QUIÑONES · Statistical Control by Monitoring and Adjustment, *Second Edition*

* BROWN and HOLLANDER · Statistics: A Biomedical Introduction

CAIROLI and DALANG · Sequential Stochastic Optimization

CASTILLO, HADI, BALAKRISHNAN, and SARABIA · Extreme Value and Related Models with Applications in Engineering and Science

CHAN · Time Series: Applications to Finance with R and S-Plus$^{\backslash sr\backslash}$, *Second Edition*

CHARALAMBIDES · Combinatorial Methods in Discrete Distributions

CHATTERJEE and HADI · Regression Analysis by Example, *Fourth Edition*

CHATTERJEE and HADI · Sensitivity Analysis in Linear Regression

CHERNICK · Bootstrap Methods: A Guide for Practitioners and Researchers, *Second Edition*

CHERNICK and FRIIS · Introductory Biostatistics for the Health Sciences

CHILES and DELFINER · Geostatistics: Modeling Spatial Uncertainty, *Second Edition*

CHOW and LIU · Design and Analysis of Clinical Trials: Concepts and Methodologies, *Second Edition*

CLARKE · Linear Models: The Theory and Application of Analysis of Variance

CLARKE and DISNEY · Probability and Random Processes: A First Course with Applications, *Second Edition*

* COCHRAN and COX · Experimental Designs, *Second Edition*

COLLINS and LANZA · Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences

CONGDON · Applied Bayesian Modelling

CONGDON · Bayesian Models for Categorical Data

CONGDON · Bayesian Statistical Modelling, *Second Edition*

CONOVER · Practical Nonparametric Statistics, *Third Edition*

COOK · Regression Graphics

COOK and WEISBERG · An Introduction to Regression Graphics

COOK and WEISBERG · Applied Regression Including Computing and Graphics

CORNELL · A Primer on Experiments with Mixtures

CORNELL · Experiments with Mixtures, Designs, Models, and the Analysis of Mixture Data, *Third Edition*

COX · A Handbook of Introductory Statistical Methods

CRESSIE · Statistics for Spatial Data, *Revised Edition*

CRESSIE and WIKLE · Statistics for Spatio-Temporal Data

CSÖRGÖ and HORVÁTH · Limit Theorems in Change Point Analysis

[*] Now available in a lower priced paperback edition in the Wiley Classics Library

[†] Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

DAGPUNAR · Simulation and Monte Carlo: With Applications in Finance and MCMC

DANIEL · Applications of Statistics to Industrial Experimentation

DANIEL · Biostatistics: A Foundation for Analysis in the Health Sciences, *Eighth Edition*

* DANIEL · Fitting Equations to Data: Computer Analysis of Multifactor Data, *Second Edition*

DASU and JOHNSON · Exploratory Data Mining and Data Cleaning

DAVID and NAGARAJA · Order Statistics, *Third Edition*

* DEGROOT, FIENBERG, and KADANE · Statistics and the Law

DEL CASTILLO · Statistical Process Adjustment for Quality Control

DEMARIS · Regression with Social Data: Modeling Continuous and Limited Response Variables

DEMIDENKO · Mixed Models: Theory and Applications

TXT~~000004

DENISON, HOLMES, MALLICK and SMITH · Bayesian Methods for Nonlinear Classification and Regression

DETTE and STUDDEN · The Theory of Canonical Moments with Applications in Statistics, Probability, and Analysis

DEY and MUKERJEE · Fractional Factorial Plans

DE ROCQUIGNY · Modelling Under Risk and Uncertainty: An Introduction to Statistical, Phenomenological and Computational Models

DILLON and GOLDSTEIN · Multivariate Analysis: Methods and Applications

* DODGE and ROMIG · Sampling Inspection Tables, *Second Edition*

* DOOB · Stochastic Processes

DOWDY, WEARDEN, and CHILKO · Statistics for Research, *Third Edition*

DRAPER and SMITH · Applied Regression Analysis, *Third Edition*

DRYDEN and MARDIA · Statistical Shape Analysis

DUDEWICZ and MISHRA · Modern Mathematical Statistics

DUNN and CLARK · Basic Statistics: A Primer for the Biomedical Sciences, *Fourth Edition*

DUPUIS and ELLIS · A Weak Convergence Approach to the Theory of Large Deviations

EDLER and KITSOS · Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment

* ELANDT-JOHNSON and JOHNSON · Survival Models and Data Analysis

ENDERS · Applied Econometric Time Series, *Third Edition*

† ETHIER and KURTZ · Markov Processes: Characterization and Convergence

EVANS, HASTINGS, and PEACOCK · Statistical Distributions, *Third Edition*

EVERITT, LANDAU, LEESE, and STAHL · Cluster Analysis, *Fifth Edition*

FEDERER and KING · Variations on Split Plot and Split Block Experiment Designs

FELLER · An Introduction to Probability Theory and Its Applications, Volume I, *Third Edition,* Revised; Volume II, *Second Edition*

* Now available in a lower priced paperback edition in the Wiley Classics Library

† Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

FITZMAURICE, LAIRD, and WARE · Applied Longitudinal Analysis, *Second Edition*

* FLEISS · The Design and Analysis of Clinical Experiments

FLEISS · Statistical Methods for Rates and Proportions, *Third Edition*

† FLEMING and HARRINGTON · Counting Processes and Survival Analysis

FUJIKOSHI, ULYANOV, and SHIMIZU · Multivariate Statistics: High-Dimensional and Large-Sample Approximations

FULLER · Introduction to Statistical Time Series, *Second Edition*

† FULLER · Measurement Error Models

GALLANT · Nonlinear Statistical Models

GEISSER · Modes of Parametric Statistical Inference

GELMAN and MENG · Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives

GEWEKE · Contemporary Bayesian Econometrics and Statistics

GHOSH, MUKHOPADHYAY, and SEN · Sequential Estimation

GIESBRECHT and GUMPERTZ · Planning, Construction, and Statistical Analysis of Comparative Experiments

GIFI · Nonlinear Multivariate Analysis

GIVENS and HOETING · Computational Statistics

GLASSERMAN and YAO · Monotone Structure in Discrete-Event Systems

GNANADESIKAN · Methods for Statistical Data Analysis of Multivariate Observations, *Second Edition*

GOLDSTEIN · Multilevel Statistical Models, *Fourth Edition*

GOLDSTEIN and LEWIS · Assessment: Problems, Development, and Statistical Issues

GOLDSTEIN and WOOFF · Bayes Linear Statistics

GREENWOOD and NIKULIN · A Guide to Chi-Squared Testing

GROSS, SHORTLE, THOMPSON, and HARRIS · Fundamentals of Queueing Theory, *Fourth Edition*

GROSS, SHORTLE, THOMPSON, and HARRIS · Solutions Manual to Accompany Fundamentals of Queueing Theory, *Fourth Edition*

* HAHN and SHAPIRO · Statistical Models in Engineering

HAHN and MEEKER · Statistical Intervals: A Guide for Practitioners

HALD · A History of Probability and Statistics and their Applications Before 1750

† HAMPEL · Robust Statistics: The Approach Based on Influence Functions

HARTUNG, KNAPP, and SINHA · Statistical Meta-Analysis with Applications

HEIBERGER · Computation for the Analysis of Designed Experiments

EDAYAT and SINHA · Design and Inference in Finite Population Sampling

HEDEKER and GIBBONS · Longitudinal Data Analysis

HELLER · MACSYMA for Statisticians

HERITIER, CANTONI, COPT, and VICTORIA-FESER · Robust Methods in Biostatistics

HINKELMANN and KEMPTHORNE · Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design, *Second Edition*

* Now available in a lower priced paperback edition in the Wiley Classics Library

† Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

HINKELMANN and KEMPTHORNE · Design and Analysis of Experiments, Volume 2: Advanced Experimental Design

HINKELMANN (editor) ·Design and Analysis of Experiments, Volume 3: Special Designs and Applications

* HOAGLIN, MOSTELLER, and TUKEY · Fundamentals of Exploratory Analysis of Variance

* HOAGLIN, MOSTELLER, and TUKEY · Exploring Data Tables, Trends and Shapes

* HOAGLIN, MOSTELLER, and TUKEY · Understanding Robust and Exploratory Data Analysis

HOCHBERG and TAMHANE · Multiple Comparison Procedures

HOCKING · Methods and Applications of Linear Models: Regression and the Analysis of Variance, *Second Edition*

HOEL · Introduction to Mathematical Statistics, *Fifth Edition*

HOGG and KLUGMAN · Loss Distributions

HOLLANDER and WOLFE · Nonparametric Statistical Methods, *Second Edition*

HOSMER and LEMESHOW · Applied Logistic Regression, *Second Edition*

HOSMER, LEMESHOW, and MAY · Applied Survival Analysis: Regression Modeling of Time-to-Event Data, *Second Edition*

HUBER · Data Analysis: What Can Be Learned From the Past 50 Years

HUBER · Robust Statistics

† HUBER and RONCHETTI · Robust Statistics, *Second Edition*

HUBERTY · Applied Discriminant Analysis, *Second Edition*

HUBERTY and OLEJNIK · Applied MANOVA and Discriminant Analysis, *Second Edition*

HUITEMA · The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, *Second Edition*

HUNT and KENNEDY · Financial Derivatives in Theory and Practice, *Revised Edition*

HURD and MIAMEE · Periodically Correlated Random Sequences: Spectral Theory and Practice

HUSKOVA, BERAN, and DUPAC · Collected Works of Jaroslav Hajek—with Commentary

HUZURBAZAR · Flowgraph Models for Multistate Time-to-Event Data

INSUA, RUGGERI and WIPER · Bayesian Analysis of Stochastic Process Models

JACKMAN · Bayesian Analysis for the Social Sciences

† JACKSON · A User's Guide to Principle Components

JOHN · Statistical Methods in Engineering and Quality Assurance

JOHNSON · Multivariate Statistical Simulation

JOHNSON and BALAKRISHNAN · Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz

JOHNSON, KEMP, and KOTZ · Univariate Discrete Distributions, *Third Edition*

JOHNSON and KOTZ (editors) ·Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present

[*] Now available in a lower priced paperback edition in the Wiley Classics Library

[†] Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 1, *Second Edition*

JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 2, *Second Edition*

JOHNSON, KOTZ, and BALAKRISHNAN · Discrete Multivariate Distributions

JUDGE, GRIFFITHS, HILL, LTKEPOHL, and LEE · The Theory and Practice of Econometrics, *Second Edition*

JUREK and MASON · Operator-Limit Distributions in Probability Theory

KADANE · Bayesian Methods and Ethics in a Clinical Trial Design

KADANE AND SCHUM · A Probabilistic Analysis of the Sacco and Vanzetti Evidence

KALBFLEISCH and PRENTICE · The Statistical Analysis of Failure Time Data, *Second Edition*

KARIYA and KURATA · Generalized Least Squares

KASS and VOS · Geometrical Foundations of Asymptotic Inference

† KAUFMAN and ROUSSEEUW · Finding Groups in Data: An Introduction to Cluster Analysis

KEDEM and FOKIANOS · Regression Models for Time Series Analysis

KENDALL, BARDEN, CARNE, and LE · Shape and Shape Theory

KHURI · Advanced Calculus with Applications in Statistics, *Second Edition*

KHURI, MATHEW, and SINHA · Statistical Tests for Mixed Linear Models

* KISH· Statistical Design for Research

KLEIBER and KOTZ · Statistical Size Distributions in Economics and Actuarial Sciences

KLEMELÄ · Smoothing of Multivariate Data: Density Estimation and Visualization

KLUGMAN, PANJER, and WILLMOT · Loss Models: From Data to Decisions, *Third Edition*

KLUGMAN, PANJER, and WILLMOT · Solutions Manual to Accompany Loss Models: From Data to Decisions, *Third Edition*

KOSKI and NOBLE · Bayesian Networks: An Introduction

KOTZ, BALAKRISHNAN, and JOHNSON · Continuous Multivariate Distributions, Volume 1, *Second Edition*

KOTZ and JOHNSON (editors) ·Encyclopedia of Statistical Sciences: Volumes 1 to 9 with Index

KOTZ and JOHNSON (editors) ·Encyclopedia of Statistical Sciences: Supplement Volume

KOTZ, READ, and BANKS (editors) ·Encyclopedia of Statistical Sciences: Update Volume 1

KOTZ, READ, and BANKS (editors) ·Encyclopedia of Statistical Sciences: Update Volume 2

KOWALSKI and TU · Modern Applied U-Statistics

KRISHNAMOORTHY and MATHEW · Statistical Tolerance Regions: Theory, Applications, and Computation

* Now available in a lower priced paperback edition in the Wiley Classics Library

† Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

KROESE, TAIMRE, and BOTEV · Handbook of Monte Carlo Methods

KROONENBERG · Applied Multiway Data Analysis

KULINSKAYA, MORGENTHALER, and STAUDTE · Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence

KULKARNI and HARMAN · An Elementary Introduction to Statistical Learning Theory

KUROWICKA and COOKE · Uncertainty Analysis with High Dimensional Dependence Modelling

KVAM and VIDAKOVIC · Nonparametric Statistics with Applications to Science and Engineering

LACHIN · Biostatistical Methods: The Assessment of Relative Risks, *Second Edition*

LAD · Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction

LAMPERTI · Probability: A Survey of the Mathematical Theory, *Second Edition*

LAWLESS · Statistical Models and Methods for Lifetime Data, *Second Edition*

LAWSON · Statistical Methods in Spatial Epidemiology, *Second Edition*

LE · Applied Categorical Data Analysis, *Second Edition*

LE · Applied Survival Analysis

LEE · Structural Equation Modeling: A Bayesian Approach

LEE and WANG · Statistical Methods for Survival Data Analysis, *Third Edition*

LEPAGE and BILLARD · Exploring the Limits of Bootstrap

LESSLER and KALSBEEK · Nonsampling Errors in Surveys

LEYLAND and GOLDSTEIN (editors) ·Multilevel Modelling of Health Statistics

LIAO · Statistical Group Comparison

LIN · Introductory Stochastic Analysis for Finance and Insurance

LITTLE and RUBIN · Statistical Analysis with Missing Data, *Second Edition*

LLOYD · The Statistical Analysis of Categorical Data

LOWEN and TEICH · Fractal-Based Point Processes

MAGNUS and NEUDECKER · Matrix Differential Calculus with Applications in Statistics and Econometrics, *Revised Edition*

MALLER and ZHOU · Survival Analysis with Long Term Survivors

MARCHETTE · Random Graphs for Statistical Pattern Recognition

MARDIA and JUPP · Directional Statistics

MARKOVICH · Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice

MARONNA, MARTIN and YOHAI · Robust Statistics: Theory and Methods

MASON, GUNST, and HESS · Statistical Design and Analysis of Experiments with Applications to Engineering and Science, *Second Edition*

McCULLOCH, SEARLE, and NEUHAUS · Generalized, Linear, and Mixed Models, *Second Edition*

McFADDEN · Management of Data in Clinical Trials, *Second Edition*

* McLACHLAN · Discriminant Analysis and Statistical Pattern Recognition

[*] Now available in a lower priced paperback edition in the Wiley Classics Library

[†] Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

RUBINSTEIN and KROESE · Simulation and the Monte Carlo Method, *Second Edition*

RUBINSTEIN and MELAMED · Modern Simulation and Modeling

RYAN · Modern Engineering Statistics

RYAN · Modern Experimental Design

RYAN · Modern Regression Methods, *Second Edition*

RYAN · Statistical Methods for Quality Improvement, *Third Edition*

SALEH · Theory of Preliminary Test and Stein-Type Estimation with Applications

SALTELLI, CHAN, and SCOTT (editors) ·Sensitivity Analysis

SCHERER · Batch Effects and Noise in Microarray Experiments: Sources and Solutions

* SCHEFFE · The Analysis of Variance

SCHIMEK · Smoothing and Regression: Approaches, Computation, and Application

SCHOTT · Matrix Analysis for Statistics, *Second Edition*

SCHOUTENS · Levy Processes in Finance: Pricing Financial Derivatives

SCOTT · Multivariate Density Estimation: Theory, Practice, and Visualization

* SEARLE· Linear Models

† SEARLE · Linear Models for Unbalanced Data

† SEARLE · Matrix Algebra Useful for Statistics

† SEARLE, CASELLA, and McCULLOCH · Variance Components

SEARLE and WILLETT · Matrix Algebra for Applied Economics

SEBER · A Matrix Handbook For Statisticians

† SEBER · Multivariate Observations

SEBER and LEE · Linear Regression Analysis, *Second Edition*

† SEBER and WILD · Nonlinear Regression

SENNOTT · Stochastic Dynamic Programming and the Control of Queueing Systems

* SERFLING · Approximation Theorems of Mathematical Statistics

SHAFER and VOVK · Probability and Finance: Its Only a Game!

SHERMAN · Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties

SILVAPULLE and SEN · Constrained Statistical Inference: Inequality, Order, and Shape Restrictions

SINGPURWALLA · Reliability and Risk: A Bayesian Perspective

SMALL and McLEISH · Hilbert Space Methods in Probability and Statistical Inference

SRIVASTAVA · Methods of Multivariate Statistics

STAPLETON · Linear Statistical Models, *Second Edition*

STAPLETON · Models for Probability and Statistical Inference: Theory and Applications

STAUDTE and SHEATHER · Robust Estimation and Testing

STOYAN · Counterexamples in Probability, *Second Edition*

* Now available in a lower priced paperback edition in the Wiley Classics Library

† Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

STOYAN, KENDALL, and MECKE · Stochastic Geometry and Its Applications, *Second Edition*

STOYAN and STOYAN · Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics

STREET and BURGESS · The Construction of Optimal Stated Choice Experiments: Theory and Methods

STYAN · The Collected Papers of T. W. Anderson: 1943–1985

SUTTON, ABRAMS, JONES, SHELDON, and SONG · Methods for Meta-Analysis in Medical Research

TAKEZAWA · Introduction to Nonparametric Regression

TAMHANE · Statistical Analysis of Designed Experiments: Theory and Applications

TANAKA · Time Series Analysis: Nonstationary and Noninvertible Distribution Theory

THOMPSON · Empirical Model Building: Data, Models, and Reality, *Second Edition*

THOMPSON · Sampling, *Third Edition*

THOMPSON · Simulation: A Modeler's Approach

THOMPSON and SEBER · Adaptive Sampling

THOMPSON, WILLIAMS, and FINDLAY · Models for Investors in Real World Markets

TIERNEY · LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics

TSAY · Analysis of Financial Time Series, *Third Edition*

UPTON and FINGLETON · Spatial Data Analysis by Example, Volume II: Categorical and Directional Data

† VAN BELLE · Statistical Rules of Thumb, *Second Edition*

VAN BELLE, FISHER, HEAGERTY, and LUMLEY · Biostatistics: A Methodology for the Health Sciences, *Second Edition*

VESTRUP · The Theory of Measures and Integration

VIDAKOVIC · Statistical Modeling by Wavelets

VIERTL · Statistical Methods for Fuzzy Data

VINOD and REAGLE · Preparing for the Worst: Incorporating Downside Risk in Stock Market Investments

WALLER and GOTWAY · Applied Spatial Statistics for Public Health Data

WANG and WANG · Structural Equation Modeling: Applications Using M*plus*

WEISBERG · Applied Linear Regression, *Third Edition*

WEISBERG · Bias and Causation: Models and Judgment for Valid Comparisons

WELSH · Aspects of Statistical Inference

WESTFALL and YOUNG · Resampling-Based Multiple Testing: Examples and Methods for *p*-Value Adjustment

* WHITTAKER · Graphical Models in Applied Multivariate Statistics

WINKER · Optimization Heuristics in Economics: Applications of Threshold Accepting

* Now available in a lower priced paperback edition in the Wiley Classics Library

† Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.