FIFTH EDITION

$R(t) = \exp\left[-\left(\frac{t}{\eta}\right)^{\beta}\right]$

Practical Reliability Engineering

PATRICK O'CONNOR | ANDRE KLEYNER

WILEY

Practical Reliability Engineering

‘The concept of chance enters into the very first steps of scientific activity, by virtue of the fact that no observation is absolutely correct. I think chance is a more fundamental concept than causality, for whether in a concrete case a cause–effect relationship exists can only be judged by applying the laws of chance to the observations.’

Max Born,
Natural Philosophy of Cause and Chance

‘A statistical relationship, however strong and however suggestive, can never establish a causal connection. Our ideas on causation must come from outside statistics, ultimately from some theory.’

Kendall & Stuart,
The Advanced Theory of Statistics

‘Reliability is, after all, engineering in its most practical form.’

James R. Schlesinger
Former US Secretary of State for Defense

Practical Reliability Engineering

Fifth Edition

PATRICK D. T. O'CONNOR

and

ANDRE KLEYNER

WILEY

A John Wiley & Sons, Ltd., Publication

This edition first published 2012

Registered office
John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book. This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

The publisher and the author 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 fitness for a particular purpose. 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 every situation. In view of ongoing research, equipment modifications, changes in governmental regulations, and the constant flow of information relating to the use of experimental reagents, equipment, and devices, the reader is urged to review and evaluate the information provided in the package insert or instructions for each chemical, piece of equipment, reagent, or device for, among other things, any changes in the instructions or indication of usage and for added warnings and precautions. The fact that an organization or Website is referred to in this work as a citation and/or a potential source of further information does not mean that the author or the publisher endorses the information the organization or Website may provide or recommendations it may make. Further, readers should be aware that Internet Websites listed in this work may have changed or disappeared between when this work was written and when it is read. No warranty may be created or extended by any promotional statements for this work. Neither the publisher nor the author shall be liable for any damages arising herefrom.

Library of Congress Cataloging-in-Publication Data

Practical reliability engineering / Patrick D. T. O'Connor, Andre Kleyner. – 5th ed.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-97982-2 (hardback) – ISBN 978-0-470-97981-5 (paper)
1. Reliability (Engineering) I. O'Connor, Patrick D. T. II. Kleyner, Andre.
TS173.O29 2012
620′.00452–dc23
2011032987

A catalogue record for this book is available from the British Library.

Print ISBN: 9780470979822
Paper ISBN: 9780470979815
ePDF ISBN: 9781119961277
oBook ISBN: 9781119961260
ePub ISBN: 9781119964094
Mobi ISBN: 9781119964100

Typeset in 10/12pt Times by Aptara Inc., New Delhi, India

To my wife Lois,
for encouragement and support,
and to the memory of Ina
Patrick O'Connor

To my wife and best friend Faina,
for her patience and unwavering support
Andre Kleyner

Contents

Preface to the First Edition

This book is designed to provide an introduction to reliability engineering and management, both for students and for practising engineers and managers. The emphasis throughout is on practical applications, and the mathematical concepts described are accordingly limited to those necessary for solution of the types of problems covered. Practical approaches to problem-solving, such as the use of probability plotting techniques and computer programs, are stressed throughout. More advanced texts are cited for further reading on the mathematical and statistical aspects. The references given in the Bibliographies are limited to those considered to provide a direct continuation of the chapter material, with the emphasis on practical applications. Tables and charts are provided to complement the analytical methods described, and numerous worked examples are included.

The book describes and comments on the usage of the major national and government standards and specifications covering reliability engineering and management in the USA and the UK. It is considered that this is an important aspect of the practical approach, since so much engineering development work is now governed by such documents. The effects of current engineering, commercial and legislative developments, such as microelectronics, software-based systems, consumerism and product liability, are covered in some detail.

The requirements of the examination syllabi of the American Society for Quality Control, and the Institute of Quality Assurance (UK) in reliability engineering are covered, so the book will be suitable for use in courses leading to these qualifications. The emphasis on practical approaches to engineering and management, the comprehensive coverage of standards and specifications, and the overall layout of the book should make it equally as suitable as a general up to date reference for use in industry and in government agencies.

PATRICK O'CONNOR
1981

Preface to the Second Edition

I have received much helpful criticism of the first edition of my book since it appeared in 1981. Whilst the reviews have generally not been unfavourable, critics have pointed out that, despite the title, the book was not quite practical enough in some areas. I have also come to realize this through my own work, particularly on the application of mathematical modelling and statistics to reliability problems. Consequently, much of the revision for the second edition has been to add to what I consider to be the practical aspects of management and engineering for reliability.

I have added to the sections on reliability prediction, demonstration and measurement, to explain and to stress the fundamental and considerable uncertainty associated with attempts to quantify and forecast a property of engineered products which is inherently non-deterministic. I believe that when people involved in reliability work manage to unshackle themselves from the tyranny of the 'numbers game' the way is cleared for the practical engineering and management approaches that are the only ways to achieve the highly reliable products demanded by the markets of today. I have not removed the descriptions of the methods for quantifying reliability, since I believe that, when these are applied with commonsense and understanding of their inherent limitations, they can help us to solve reliability problems and to design and make better products.

I have added three new chapters, all related to the practical aspects.

The first edition described how to analyse test data, but included little on how to test. I have therefore written a new chapter on reliability testing, covering environmental and stress testing and the integration of reliability and other development testing. I am indebted to Wayne Tustin for suggesting this and for his help and advice on this subject.

The quality of manufacture is obviously fundamental to achieving high reliability. This point was made in the first edition, but was not developed. I have added a complete chapter on quality assurance (QA), as well as new material on integrated management of reliability and QA programmes.

Maintenance also affects reliability, so I have added a new chapter on maintenance and maintainability, with the emphasis on how they affect reliability, how reliability affects maintenance planning and how both affect availability.

I have also added new material on the important topic of reliability analysis for repairable systems. Harry Ascher, of the US Naval Research Laboratory, has pointed out that the reliability literature, including the first edition of my book, has almost totally ignored this aspect, leading to confusion and analytical errors. How many reliability engineers and teachers know that Weibull analysis of repairable system reliability data can be quite misleading except under special, unrealistic conditions? Thanks to Harry Ascher, I know now, and I have tried to explain this in the new edition.

I have also brought other parts of the book up to date, particularly the sections on electronic and software reliability.

The third reprint of the first edition included many corrections, and more corrections are made in this edition.

I am extremely grateful to all those who have pointed out errors and have helped me to correct them. Paul Baird of Hewlett Packard, Palo Alto, was particularly generous. Colleagues at British Aerospace, particularly

Brian Collett, Norman Harris, Chris Gilders and Gene Morgan, as well as many others, also provided help, advice and inspiration.

Finally, my thanks go to my wife Ina for much patience, support and typing.

Patrick O'Connor
1985

Preface to the Third Edition

The new industrial revolution has been driven mainly by the continuing improvements in quality and productivity in nearly all industrial sectors. The key to success in every case has been the complete integration of the processes that influence quality and reliability, in product specification, design, test, manufacture, and support. The other essential has been the understanding and control of variation, in the many ways in which it can affect product performance, cost and reliability. Teachers such as W. E. Deming and G. Taguchi have continued to grow in stature and following as these imperatives become increasingly the survival kit of modern industry.

I tried to stress these factors in the second edition, but I have now given them greater prominence. I have emphasized the use of statistical experimentation for preventing problems, not just for solving them, and the topic is now described as a design and development activity. I have added to the chapter on production quality assurance, to include process improvement methods and more information on process control techniques. These chapters, and the chapter on management, have all been enlarged to emphasize the integration of engineering effort to identify, minimize and reduce variation and its effects. The important work of Taguchi and Shainin is described, for the first time in this book. Chris Gray gave me much valuable help in describing the Taguchi method.

I have updated several chapters, particularly those on electronic systems reliability. I have also added a new chapter on reliability of mechanical components and systems. I would like to thank Professor Dennis Carter for his advice on this chapter.

I have taken the opportunity to restructure the book, to reflect better the main sequence of engineering development, whilst stressing the importance of an integrated, iterative approach.

I have once more been helped by many people who have contributed kind criticisms of the earlier edition, and I have tried to take these into account. I also would like to record with thanks my continuing debt to Norman Harris for his contributions to bridging the gap between engineering and statistics, and for helping me to express his ideas.

Finally, my heartfelt thanks go to my wife and boys for their forebearance, patience, and support. Having an author at home must place severe demands on love and tolerance.

Patrick O'Connor
1990

Preface to the Third Edition Revised

This revised edition has been produced in response to numerous suggestions that the book would be of greater value to students and teachers if it included exercise questions. David Newton and Richard Bromley have therefore teamed up with me to produce exercises appropriate to each chapter of the book.

The exercises cover nearly all of the types of questions that occur in the reliability examinations set by the UK Institute of Quality Assurance (IQA) and by the American Society for Quality Control (ASQC). The ASQC examination questions are of the multiple-choice type, which is not the format used here, but this should make no difference to the value of the exercises in preparing for the ASQC examination.

A solutions manual is available to teachers, free of charge, by writing to John Wiley and Sons Ltd in Chichester.

I would like to thank David Newton and Richard Bromley for their enthusiastic support in preparing this revised edition.

Patrick O'Connor
1995

Preface to the Fourth Edition

It is over ten years since the last major revision and update to my book. Inevitably in that time there have been developments in engineering technology and in reliability methods. In this new edition I have tried to include all of the important changes that affect reliability engineering and management today. In keeping with the original aims of the book, I have emphasised those with practical implications.

The main changes and additions include:

- Updated and more detailed descriptions of how engineering products fail (Chapters 1, 8 and 9).
- More detailed description of the nature of variation in engineering (Chapter 2).
- Descriptions of the Petri net and M(t) methods (Chapters 6 and 12).
- More detailed description of the particular aspects of software in engineering systems, and updated descriptions of design, analysis and test methods (Chapter 10).
- Expanded descriptions of accelerated test methods for development and manufacturing (Chapters 11 and 13).
- Updated and expanded descriptions of test methods for electronics and acceptance sampling (Chapter 13).
- More detailed descriptions of management aspects, including standards, "six sigma", and supplier management (Chapter 15).
- Updated references to standards, and updated and expanded bibliographies.

Some of the new material is adapted from my book "Test Engineering", with permission from the publisher.

The questions and the answers manual (available separately from the publisher) have been augmented to cover the new material.

An Internet homepage has been created for the book, at www.pat-oconnor.co.uk/practicalreliability.htm. The homepage includes listings of suppliers of reliability engineering related services and software.

I would like to express my gratitude to Prof. S.K. Yang for his kind assistance with the description of the Petri net method, Dr. Gregg Hobbs for his teaching and help on HALT/HASS testing, Prof. Jörgen Möltoft for helping with the description of the M(t) method, and Jim McLinn for providing additional material, questions and answers on aspects of accelerated testing and data analysis. I also thank all who have provided suggestions and pointed out errors. Last but certainly not least I thank my wife, Ina, again.

Patrick O'Connor
2001

Preface to the Fifth Edition

Another ten years have elapsed since publication of the fourth edition. In that interval there have been further significant developments in reliability engineering methods, mainly related to the use of software to perform analysis of designs and of reliability data. Of course there have also been developments in engineering that affect reliability. The internet has added a new dimension to the availability of information and tools.

In order to describe many of these developments, Andre Kleyner has taken on the role of joint author and the two of us have worked together to create this new edition. Andre has contributed most of the new material. In particular, he has provided the software-based solutions to many of the examples, supplementing or replacing manual and graphical methods. He has also updated some of the technology aspects and contributed new sections on data analysis and other topics.

The main changes and additions include:

- Software implementation of statistical methods, including probability plotting and a wider use of common software tools such as Microsoft Excel®.
- Expanded description and applications of Monte Carlo simulation methods, in a new chapter.
- More detailed descriptions of reliability prediction methods.
- Expanded treatment of accelerated test data analysis.
- Analysis of warranty data.
- Expanded description of reliability demonstration methods, in a new chapter.
- Course instructors who adapted this book can request the Solutions Mannual at: www.wiley.com/go/oconnor_reliability5.
- General updating of references, including published papers and internet links.
- The Questions sections, originally developed with major contributions from David Newton and Richard Bromley, have been revised and expanded.

A solutions manual for the end-of-chapter questions and instructor's PowerPoint slides are available as a free download, to course tutors only at: www.wiley.com/go/oconnor_reliability5.

We hope that the new edition will maintain the value of *Practical Reliability Engineering* to engineers, managers, teachers and students.

Patrick O'Connor
2011

Acknowledgements

We remain deeply indebted to the people who provided valuable help and advice on the first edition. Their generous efforts still enhance the book. In particular Dr. Ralph Evans, Kenneth Blemel and Norman Harris provided insights and assistance. Professor Dennis Carter was the originator of the load-strength concepts described in Chapter 5. Professor Bev Littlewood helped with the software reliability modelling descriptions in Chapter 10.

The authors would also like to express their gratitude to the people who have contributed to the present edition or helped to review the draft manuscript: Pantelis Vassiliou, Peter Sandborn, Mike Silverman, Vasiliy Krivtsov, Vitali Volovoi, Yizhak Bot, Michael Varnau, Steve McMullen, Andy Foote, Fred Schenkelberg, David Dylis, Craig Hillman, Cheryl Tulkoff, Walt Tomczykowski, Eric Juliot, Joe Boyle and Marina Shapiro.

Patrick O’Connor
pat@pat-oconnor.co.uk
Andre Kleyner
info@andre-kleyner.com
2011

1

Introduction to Reliability Engineering

1.1 What is Reliability Engineering?

No one disputes the need for engineered products to be reliable. The average consumer is acutely aware of the problem of less than perfect reliability in domestic products such as TV sets and automobiles. Organizations such as airlines, the military and public utilities are aware of the costs of unreliability. Manufacturers often suffer high costs of failure under warranty. Argument and misunderstanding begin when we try to quantify reliability values, or try to assign financial or other cost or benefit values to levels of reliability.

The simplest, purely producer-oriented or inspectors' view of quality is that in which a product is assessed against a specification or set of attributes, and when passed is delivered to the customer. The customer, having accepted the product, accepts that it might fail at some future time. This simple approach is often coupled with a warranty, or the customer may have some protection in law, so that he may claim redress for failures occurring within a stated or reasonable time. However, this approach provides no measure of quality over a period of time, particularly outside a warranty period. Even within a warranty period, the customer usually has no grounds for further action if the product fails once, twice or several times, provided that the manufacturer repairs the product as promised each time. If it fails often, the manufacturer will suffer high warranty costs, and the customers will suffer inconvenience. Outside the warranty period, only the customer suffers. In any case, the manufacturer will also probably incur a loss of reputation, possibly affecting future business.

We therefore come to the need for a time-based concept of quality. The inspectors' concept is not time-dependent. The product either passes a given test or it fails. On the other hand, reliability is usually concerned with failures in the time domain. This distinction marks the difference between traditional quality control and reliability engineering.

Whether failures occur or not, and their times to occurrence, can seldom be forecast accurately. Reliability is therefore an aspect of engineering uncertainty. Whether an item will work for a particular period is a question which can be answered as a probability. This results in the usual engineering definition of reliability as:

> *The probability that an item will perform a required function without failure under stated conditions for a stated period of time.*

Reliability can also be expressed as the number of failures over a period.

Durability is a particular aspect of reliability, related to the ability of an item to withstand the effects of time (or of distance travelled, operating cycles, etc.) dependent mechanisms such as fatigue, wear, corrosion,

Practical Reliability Engineering, Fifth Edition. Patrick D. T. O'Connor and Andre Kleyner.

electrical parameter change, and so on. Durability is usually expressed as a minimum time before the occurrence of *wearout* failures. In repairable systems it often characterizes the ability of the product to function after repairs.

The objectives of reliability engineering, in the order of priority, are:

1 To apply engineering knowledge and specialist techniques to prevent or to reduce the likelihood or frequency of failures.
2 To identify and correct the causes of failures that do occur, despite the efforts to prevent them.
3 To determine ways of coping with failures that do occur, if their causes have not been corrected.
4 To apply methods for estimating the likely reliability of new designs, and for analysing reliability data.

The reason for the priority emphasis is that it is by far the most effective way of working, in terms of minimizing costs and generating reliable products.

The primary skills that are required, therefore, are the ability to understand and anticipate the possible causes of failures, and knowledge of how to prevent them. It is also necessary to have knowledge of the methods that can be used for analysing designs and data. The primary skills are nothing more than good engineering knowledge and experience, so reliability engineering is first and foremost the application of good engineering, in the widest sense, during design, development, manufacture and service.

Mathematical and statistical methods can be used for quantifying reliability (prediction, measurement) and for analysing reliability data. The basic methods are described in Chapter 2, to provide an introduction for some of the applications described subsequently. However, because of the high levels of uncertainty involved these can seldom be applied with the kind of precision and credibility that engineers are accustomed to when dealing with most other problems. In practice the uncertainty is often in orders of magnitude. Therefore the role of mathematical and statistical methods in reliability engineering is limited, and appreciation of the uncertainty is important in order to minimize the chances of performing inappropriate analysis and of generating misleading results. Mathematical and statistical methods can make valuable contributions in appropriate circumstances, but practical engineering must take precedence in determining the causes of problems and their solutions. Unfortunately not all reliability training, literature and practice reflect this reality.

Over-riding all of these aspects, though, is the management of the reliability engineering effort. Since reliability (and very often also safety) is such a critical parameter of most modern engineering products, and since failures are generated primarily by the people involved (designers, test engineers, manufacturing, suppliers, maintainers, users), it can be maximized only by an integrated effort that encompasses training, teamwork, discipline, and application of the most appropriate methods. Reliability engineering “specialists” cannot make this happen. They can provide support, training and tools, but only managers can organize, motivate, lead and provide the resources. Reliability engineering is, ultimately, effective management of engineering.

1.2 Why Teach Reliability Engineering?

Engineering education is traditionally concerned with teaching how manufactured products work. The ways in which products fail, the effects of failure and aspects of design, manufacture, maintenance and use which affect the likelihood of failure are not usually taught[1], mainly because it is necessary to understand how a

[1] Mechanical engineering curricula normally include basic failure processes such as fracture mechanics, wear and corrosion.

product works before considering ways in which it might fail. For many products the tendency to approach the failed state is analogous to entropy. The engineer's tasks are to design and maintain the product so that the failed state is deferred. In these tasks he faces the problems inherent in the variability of engineering materials, processes and applications. Engineering education is basically deterministic, and does not usually pay sufficient attention to variability. Yet variability and chance play a vital role in determining the reliability of most products. Basic parameters like mass, dimensions, friction coefficients, strengths and stresses are never absolute, but are in practice subject to variability due to process and materials variations, human factors and applications. Some parameters also vary with time. Understanding the laws of chance and the causes and effects of variability is therefore necessary for the creation of reliable products and for the solution of problems of unreliability.

However, there are practical problems in applying statistical knowledge to engineering problems. These problems have probably deterred engineers in the past from using statistical methods, and texts on reliability engineering and mathematics have generally stressed the theoretical aspects without providing guidance on their practical application. To be helpful a theoretical basis must be credible, and statistical methods which work well for insurance actuaries, market researchers or agricultural experimenters may not work as well for engineers. This is not because the theory is wrong, but because engineers usually have to cope with much greater degrees of uncertainty, mainly due to human factors in production and use.

Some highly reliable products are produced by design and manufacturing teams who practise the traditional virtues of reliance on experience and maintenance of high quality. They do not see reliability engineering as a subject requiring specialist consideration, and a book such as this would teach them little that they did not already practise in creating their reliable products. Engineers and managers might therefore regard a specialist reliability discipline with scepticism. However, many pressures now challenge the effectiveness of the traditional approaches. Competition, the pressure of schedules and deadlines, the cost of failures, the rapid evolution of new materials, methods and complex systems, the need to reduce product costs, and safety considerations all increase the risks of product development. Figure 1.1 shows the pressures that lead to the overall perception of risk. Reliability engineering has developed in response to the need to control these risks.

Later chapters will show how reliability engineering methods can be applied to design, development, manufacturing and maintenance to control the level of risk. The extent to which the methods are applicable must be decided for each project and for each design area. They must not replace normal good practice, such as safe design for components subject to cyclic loading, or application guidelines for electronic components.

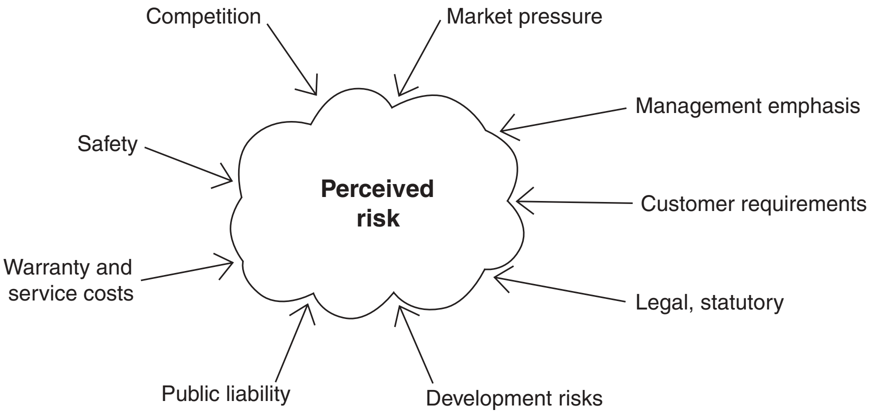


Figure 1.1 *Perception of risk.*

They should be used to supplement good practice. However, there are times when new risks are being taken, and the normal rules and guidelines are inadequate or do not apply. Sometimes we take risks unwittingly, when we assume that we can extrapolate safely from our present knowledge. Designers and managers are often overoptimistic or are reluctant to point out risks about which they are unsure.

It is for these reasons that an understanding of reliability engineering principles and methods is now an essential ingredient of modern engineering.

1.3 Why Do Engineering Products Fail?

There are many reasons why a product might fail. *Knowing, as far as is practicable, the potential causes of failures is fundamental to preventing them*. It is rarely practicable to anticipate all of the causes, so it is also necessary to take account of the uncertainty involved. The reliability engineering effort, during design, development and in manufacture and service should address all of the anticipated and possibly unanticipated causes of failure, to ensure that their occurrence is prevented or minimized.

The main reasons why failures occur are:

1 The design might be *inherently incapable*. It might be too weak, consume too much power, suffer resonance at the wrong frequency, and so on. The list of possible reasons is endless, and every design problem presents the potential for errors, omissions, and oversights. The more complex the design or difficult the problems to be overcome, the greater is this potential.

2 The item might be *overstressed* in some way. If the stress applied exceeds the strength then failure will occur. An electronic component will fail if the applied electrical stress (voltage, current) exceeds the ability to withstand it, and a mechanical strut will buckle if the compression stress applied exceeds the buckling strength. Overstress failures such as these do happen, but fortunately not very often, since designers provide margins of safety. Electronic component specifications state the maximum rated conditions of application, and circuit designers take care that these rated values are not exceeded in service. In most cases they will in fact do what they can to ensure that the in-service worst case stresses remain below the rated stress values: this is called ‘de-rating’. Mechanical designers work in the same way: they know the properties of the materials being used (e.g. ultimate tensile strength) and they ensure that there is an adequate margin between the strength of the component and the maximum applied stress. However, it might not be possible to provide protection against every possible stress application.

3 Failures might be caused by *variation*. In the situations described above the values of strength and load are fixed and known. If the known strength always exceeds the known load, as shown in Figure 1.2, then failure will not occur. However, in most cases, there will be some uncertainty about both. The actual strength values of any population of components will vary: there will be some that are relatively strong, others that are relatively weak, but most will be of nearly average strength. Also, the loads applied will be variable. Figure 1.3 shows this type of situation. As before, failure will not occur so long as the applied load does not exceed the strength. However, if there is an overlap between the distributions of load and strength, and a load value in the high tail of the load distribution is applied to an item in the weak tail of the strength distribution so that there is overlap or *interference* between the distributions (Figure 1.4), then failure will occur. We will discuss load and strength interference in more detail in Chapter 5.

4 Failures can be caused by *wearout*. We will use this term to include any mechanism or process that causes an item that is sufficiently strong at the start of its life to become weaker with age. Well-known examples of such processes are material fatigue, wear between surfaces in moving contact, corrosion, insulation deterioration, and the wearout mechanisms of light bulbs and fluorescent tubes. Figure 1.5 illustrates this kind of situation. Initially the strength is adequate to withstand the applied loads, but as weakening occurs

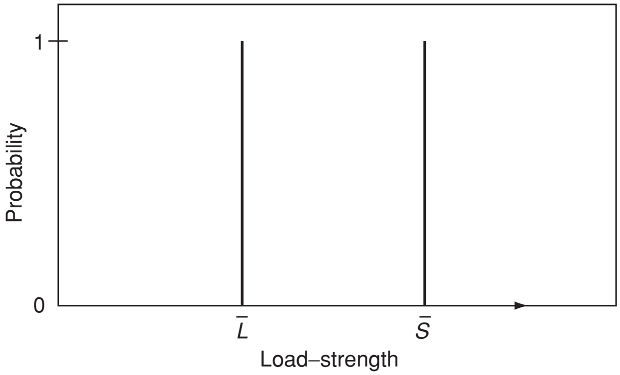


Figure 1.2 *Load–strength – discrete values.*

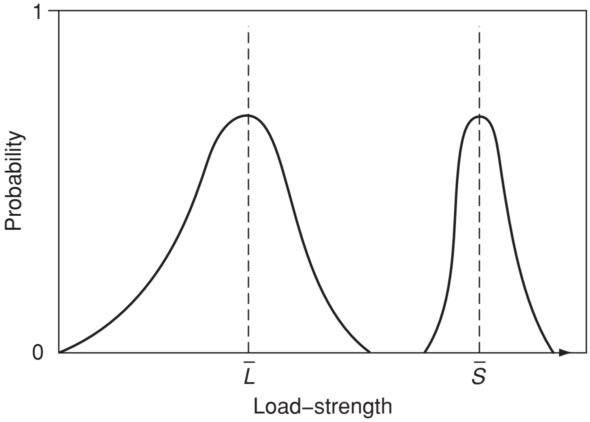


Figure 1.3 *Load–strength – distributed values.*

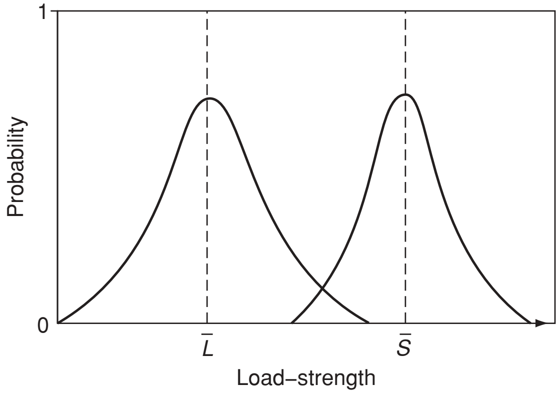


Figure 1.4 *Load–strength – interfering distributions.*

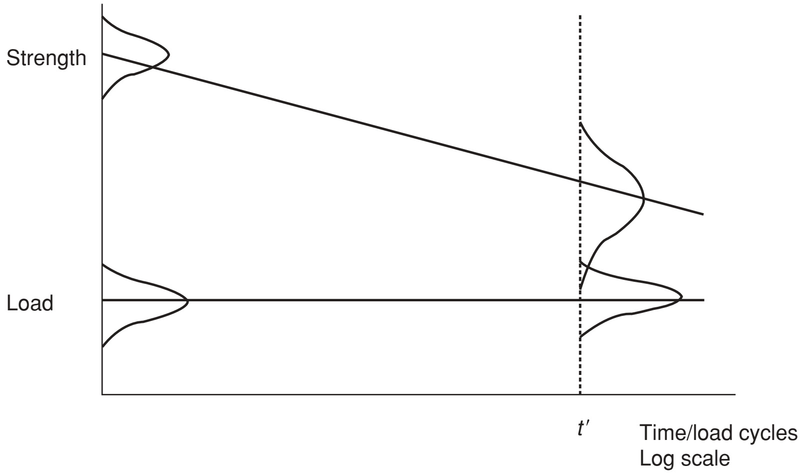


Figure 1.5 *Time-dependent load and strength variation.*

over time the strength decreases. In every case the average value falls and the spread of the strength distribution widens. This is a major reason why it is so difficult to provide accurate predictions of the lives of such items.

5 Failures can be caused by other time-dependent mechanisms. Battery run-down, creep caused by simultaneous high temperature and tensile stress, as in turbine discs and fine solder joints, and progressive drift of electronic component parameter values are examples of such mechanisms.

6 Failures can be caused by *sneaks*. A sneak is a condition in which the system does not work properly even though every part does. For example, an electronic system might be designed in such a way that under certain conditions incorrect operation occurs. The fatal fire in the Apollo spacecraft crew capsule was caused in this way: the circuit design ensured that an electrical short circuit would occur when a particular sequence was performed by the crew. Sneaks can also occur in software designs.

7 Failures can be caused by *errors*, such as incorrect specifications, designs or software coding, by faulty assembly or test, by inadequate or incorrect maintenance, or by incorrect use. The actual failure mechanisms that result might include most of the list above.

8 There are many other potential causes of failure. Gears might be noisy, oil seals might leak, display screens might flicker, operating instructions might be wrong or ambiguous, electronic systems might suffer from electromagnetic interference, and so on.

Failures have many different causes and effects, and there are also different perceptions of what kinds of events might be classified as failures. The burning O-ring seals on the Space Shuttle booster rockets were not classed as failures, until the ill-fated launch of Challenger. We also know that all failures, in principle and almost always in practice, can be prevented.

1.4 Probabilistic Reliability

The concept of reliability as a probability means that any attempt to quantify it must involve the use of statistical methods. An understanding of statistics as applicable to reliability engineering is therefore a necessary basis for progress, except for the special cases when reliability is perfect (we know the item will never fail) or it is zero (the item will never work). In engineering we try to ensure 100 % reliability, but our experience tells us that we do not always succeed. Therefore reliability statistics are usually concerned with probability values which are very high (or very low: the probability that a failure does occur, which is 1 – reliability). Quantifying such numbers brings increased uncertainty, since we need correspondingly more information. Other sources of uncertainty are introduced because reliability is often about people who make and people who use the product, and because of the widely varying environments in which typical products might operate.

Further uncertainty, often of a subjective nature, is introduced when engineers begin to discuss failures. Should a failure be counted if it was due to an error that is hoped will not be repeated? If design action is taken to reduce the risk of one type of failure, how can we quantify our trust in the designer's success? Was the machine under test typical of the population of machines?

Reliability is quantified in other ways. We can specify a reliability as the mean number of failures in a given time (failure rate), or as the *mean time between failures* (MTBF) for items which are repaired and returned to use, or as the *mean time to failure* (MTTF) for items which are not repaired, or as the proportion of the total population of items failing during the mission life.

The application and interpretation of statistics to deal with the effects of variability on reliability are less straightforward than in, say, public opinion polls or measurement of human variations such as IQ or height. In these applications, most interest is centred around the behaviour of the larger part of the population or sample, variation is not very large and data are plentiful. In reliability we are concerned with the behaviour in the extreme tails of distributions and possibly unlikely combinations of load and strength, where variability is often hard to quantify and data are expensive.

Further difficulties arise in the application of statistical theory to reliability engineering, owing to the fact that variation is often a function of time or of time-related factors such as operating cycles, diurnal or seasonal cycles, maintenance periods, and so on. Engineering, unlike most fields of knowledge, is primarily concerned with change, hopefully, but not always, for the better. Therefore the reliability data from any past situation cannot be used to make credible forecasts of the future behaviour, without taking into account non-statistical factors such as design changes, maintainer training, and even imponderables such as unforeseeable production or service problems. The statistician working in reliability engineering needs to be aware of these realities.

Chapter 2 provides the statistical basis of reliability engineering, but it must always be remembered that quality and reliability data contain many sources of uncertainty and variability which cannot be rigorously quantified. It is also important to appreciate that failures and their causes are by no means always clear-cut and unambiguous. They are often open to interpretation and argument. They also differ in terms of importance (cost, safety, other effects). Therefore we must be careful not to apply only conventional scientific, deterministic thinking to the interpretation of failures. For example, a mere count of total reported failures of a product is seldom useful or revealing. It tells us nothing about causes or consequences, and therefore nothing about how to improve the situation. This contrasts with a statement of a physical attribute such as weight or power consumption, which is usually unambiguous and complete. Nevertheless, it is necessary to derive values for decision-making, so the mathematics are essential. The important point is that the reliability engineer or manager is not, like an insurance actuary, a powerless observer of his statistics. Statistical derivations of reliability are not a guarantee of results, and these results can be significantly affected by actions taken by quality and reliability engineers and managers.

1.5 Repairable and Non-Repairable Items

It is important to distinguish between repairable and non-repairable items when predicting or measuring reliability.

For a non-repairable item such as a light bulb, a transistor, a rocket motor or an unmanned spacecraft, reliability is the survival probability over the item's expected life, or for a period during its life, *when only one failure can occur*. During the item's life the instantaneous probability of the first and only failure is called the *hazard rate*. Life values such as the mean life or *mean time to failure* (MTTF), or the expected life by which a certain percentage might have failed (say 10 %.) (*percentile life*), are other reliability characteristics that can be used. Note that non-repairable items may be individual parts (light bulbs, transistors, fasteners) or systems comprised of many parts (spacecraft, microprocessors).

For items which are repaired when they fail, reliability is the probability that failure will not occur in the period of interest, *when more than one failure can occur*. It can also be expressed as the *rate of occurrence of failures* (ROCOF), which is sometimes referred as the *failure rate* (usually denoted as λ). However, the term failure rate has wider meaning and is often applied to both repairable and non-repairable systems expressing the number of failures per unit time, as applied to one unit in the population, when one or more failures can occur in a time continuum. It is also sometimes used as an averaged value or practical metric for the hazard rate.

Repairable system reliability can also be characterized by the *mean time between failures* (MTBF), but only under the particular condition of a constant failure rate. It is often assumed that failures do occur at a constant rate, in which case the failure rate $\lambda = (\text{MTBF})^{-1}$. However, this is only a special case, valuable because it is often true and because it is easy to understand.

We are also concerned with the *availability* of repairable items, since repair takes time. Availability is affected by the rate of occurrence of failures (failure rate) and by maintenance time. Maintenance can be corrective (i.e. repair) or preventive (to reduce the likelihood of failure, e.g. lubrication). We therefore need to understand the relationship between reliability and maintenance, and how both reliability and maintainability can affect availability.

Sometimes an item may be considered as both repairable and non-repairable. For example, a missile is a repairable system whilst it is in store and subjected to scheduled tests, but it becomes a non-repairable system when it is launched. Reliability analysis of such systems must take account of these separate states. Repairability might also be determined by other considerations. For example, whether an electronic circuit board is treated as a repairable item or not will depend upon the cost of repair. An engine or a vehicle might be treated as repairable only up to a certain age.

Repairable system reliability data analysis is covered in Chapter 13 and availability and maintainability in Chapter 16.

1.6 The Pattern of Failures with Time (Non-Repairable Items)

There are three basic ways in which the pattern of failures can change with time. The hazard rate may be decreasing, increasing or constant. We can tell much about the causes of failure and about the reliability of the item by appreciating the way the hazard rate behaves in time.

Decreasing hazard rates are observed in items which become less likely to fail as their survival time increases. This is often observed in electronic equipment and parts. 'Burn-in' of electronic parts is a good example of the way in which knowledge of a decreasing hazard rate is used to generate an improvement in reliability. The parts are operated under failure-provoking stress conditions for a time before delivery. As substandard parts fail and are rejected the hazard rate decreases and the surviving population is more reliable.

A constant hazard rate is characteristic of failures which are caused by the application of loads in excess of the design strength, at a constant average rate. For example, overstress failures due to accidental or transient circuit overload, or maintenance-induced failures of mechanical equipment, typically occur randomly and at a generally constant rate.

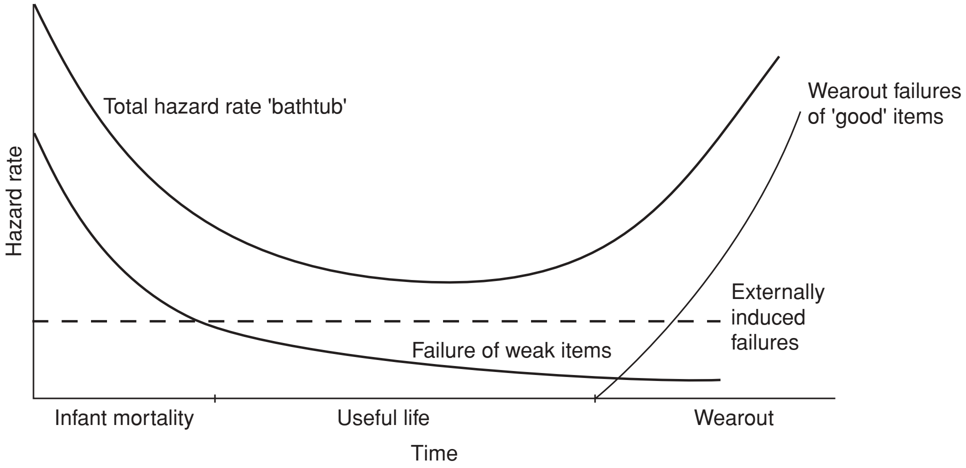


Figure 1.6 *The 'bathtub' curve.*

Wearout failure modes follow an increasing hazard rate. For example, material fatigue brought about by strength deterioration due to cyclic loading is a failure mode which does not occur for a finite time, and then exhibits an increasing probability of occurrence.

The combined effect generates the so-called *bathtub curve* (Figure 1.6). This shows an initial decreasing hazard rate or *infant mortality* period, an intermediate *useful life* period and a final *wearout* period. Death is a good analogy to failure of a non-repairable system, and the bathtub curve model is similar to actuarial statistical models.

1.7 The Pattern of Failures with Time (Repairable Items)

The failure rates (or ROCOF) of repairable items can also vary with time, and important implications can be derived from these trends.

A constant failure rate (CFR) is indicative of externally induced failures, as in the constant hazard rate situation for non-repairable items. A CFR is also typical of complex systems subject to repair and overhaul, where different parts exhibit different patterns of failure with time and parts have different ages since repair or replacement. Repairable systems can show a decreasing failure rate (DFR) when reliability is improved by progressive repair, as defective parts which fail relatively early are replaced by good parts. 'Burn in' is applied to electronic systems, as well as to parts, for this purpose.

An increasing failure rate (IFR) occurs in repairable systems when wearout failure modes of parts begin to predominate.

The pattern of failures with time of repairable systems can also be illustrated by use of the bathtub curve (Figure 1.6), but with the failure rate (ROCOF) plotted against age instead of the hazard rate.

The statistical treatment of failure data is covered in Chapters 2 and 3.

1.8 The Development of Reliability Engineering

Reliability engineering, as a separate engineering discipline, originated in the United States during the 1950s. The increasing complexity of military electronic systems was generating failure rates which resulted in greatly reduced availability and increased costs. Solid state electronics technology offered long term hope, but conversely miniaturization was to lead to proportionately greater complexity, which offset the reliability improvements expected. The gathering pace of electronic device technology meant that the developers of new military systems were making increasing use of large numbers of new component types, involving new manufacturing processes, with the inevitable consequences of low reliability. The users of such equipment were also finding that the problems of diagnosing and repairing the new complex equipment were seriously affecting its availability for use, and the costs of spares, training and other logistics support were becoming excessive. Against this background the US Department of Defense and the electronics industry jointly set up the Advisory Group on Reliability of Electronic Equipment (AGREE) in 1952. The AGREE report concluded that, to break out of the spiral of increasing development and ownership costs due to low reliability, disciplines must be laid down as integral activities in the development cycle for electronic equipment. The report laid particular stress on the need for new equipments to be tested for several thousand hours in high stress cyclical environments including high and low temperatures, vibration and switching, in order to discover the majority of weak areas in a design at an early enough stage to enable them to be corrected before production commenced. Until that time, environmental tests of tens of hours' duration had been considered adequate to prove the suitability of a design. The report also recommended that formal demonstrations of reliability, in terms of statistical confidence that a specified MTBF had been exceeded, be instituted as a condition for acceptance of equipment by the procuring agency. A large part of the report was devoted to providing detailed test plans for various levels of statistical confidence and environmental conditions.

The AGREE report was accepted by the Department of Defense, and AGREE testing quickly became a standard procedure. Companies which invested in the expensive environmental test equipment necessary soon found that they could attain levels of reliability far higher than by traditional methods. It was evident that designers, particularly those working at the fringes of advanced technology, could not be expected to produce highly reliable equipment without it being subjected to a test regime which would show up weaknesses. Complex systems and the components used in them included too many variables and interactions for the human designer to cope with infallibly, and even the most careful design reviews and disciplines could not provide sufficient protection. Consequently it was necessary to make the product speak for itself, by causing it to fail, and then to eliminate the weaknesses that caused the failures. The Department of Defense (DOD) reissued the AGREE report on testing as US Military Standard (MIL-STD) 781, *Reliability Qualification and Production Approval Tests*.

Meanwhile the revolution in electronic device technology continued, led by integrated micro circuitry. Increased emphasis was now placed on improving the quality of devices fitted to production equipments. Screening techniques, in which all production devices are subjected to elevated thermal, electrical and other stresses, were introduced in place of the traditional sampling techniques. With component populations on even single printed circuit boards becoming so large, sampling no longer provided sufficient protection against the production of defective equipment. These techniques were formalized in military standards covering the full range of electronic components. Components produced to these standards were called 'Hi-rel' components. Specifications and test systems for electronic components, based upon the US Military Standards, were developed in the United Kingdom and in continental Europe, and internationally through the International Electrotechnical Commission (IEC).

However, improved quality standards in the electronic components industry resulted in dramatic improvements in the reliability of commercial components. As a result, during the 1980s the US Military began switching from military grade electronic components to "commercial off the shelf" (COTS) parts in order to reduce costs, and this approach has spread to other applications.

Engineering reliability effort in the United States developed quickly, and the AGREE and reliability programme concepts were adopted by NASA and many other major suppliers and purchasers of high technology equipment. In 1965 the DOD issued MIL-STD-785–*Reliability Programs for Systems and Equipment*. This document made mandatory the integration of a programme of reliability engineering activities with the traditional engineering activities of design, development and production, as it was by then realized that such an integrated programme was the only way to ensure that potential reliability problems would be detected and eliminated at the earliest, and therefore the cheapest, stage in the development cycle. Much written work appeared on the cost-benefit of higher reliability, to show that effort and resources expended during early development and during production testing, plus the imposition of demonstrations of specified levels of reliability to MIL-STD-781, led to reductions in in-service costs which more than repaid the reliability programme expenditure. The concept of life cycle costs (LCC), or whole life costs, was introduced.

In the United Kingdom, Defence Standard 00–40, *The Management of Reliability and Maintainability* was issued in 1981. The British Standards Institution issued BS 5760 – *Guide on Reliability of Systems, Equipments and Components*. In the 1990s the series of European Reliability/Dependability[2] standards began to be developed, and became integrated into the International Standards Organization (ISO). For example ISO/IEC 60 300 describes the concepts and principles of dependability management systems. It identifies the generic processes for planning, resource allocation, control, and tailoring necessary to meet dependability objectives. At present, there is a large number of ISO standards regulating testing, validation, reliability analysis, and various other aspects of product development.

Starting in the early 1980s, the reliability of new Japanese industrial and commercial products took Western competitors by surprise. Products such as automobiles, electronic components and systems, and machine tools achieved levels of reliability far in excess of previous experience. These products were also less expensive and often boasted superior features and performance. The 'Japanese quality revolution' had been driven by the lessons taught by American teachers brought in to help Japan's post-war recovery. The two that stand out were J.R. Juran and W. Edwards Deming, who taught the principles of 'total quality management' (TQM) and continuous improvement. Japanese pioneers, particularly K. Ishikawa, also contributed. These ideas were all firmly rooted in the teaching of the American writer on management, Peter Drucker (Drucker, 1995), that people work most effectively when they are given the knowledge and authority to identify and implement improvements, rather than being expected to work to methods dictated by 'management'.

These ideas led to great increases in productivity and quality, and thus in reliability and market penetration, as Drucker had predicted. Many Western companies followed the new path that had been blazed and also made great improvements. The message is now almost universally applied, particularly with the trend to international outsourcing of manufacturing.

The Western approach had been based primarily on formal procedures for design analysis and reliability demonstration testing, whereas the Japanese concentrated on manufacturing quality. Nowadays most customers for systems such as military, telecommunications, transport, power generation and distribution, and so on, rely upon contractual motivation, such as warranties and service support, rather than on imposition of standards that dictate exactly how reliability activities should be performed.

Another aspect of reliability thinking that has developed is the application of statistical methods, primarily to the analysis of failure data and to predictions of reliability and safety of systems. Since reliability can be expressed as a probability, and is affected by variation, in principle these methods are applicable. They form the basis of most teaching and literature on the subject. However, variation in engineering is usually of such an uncertain nature that refined mathematical and quantitative techniques can be inappropriate and misleading. This aspect will be discussed in later chapters.

1.9 Courses, Conferences and Literature

Reliability engineering and management are now taught in engineering courses at a large number of universities, colleges and polytechnics, and on specialist short courses.

[2] In this context dependability is defined as including reliability, maintainability, availability and safety.

Conferences on general and specific reliability engineering and management topics have been held regularly in the United States since the 1960s and in Europe and elsewhere since the 1970s. The best known is the annual US Reliability and Maintainability Symposium (RAMS), sponsored by most of the important engineering associations and institutions in the United States. It is held every year and its conference proceedings contain much useful information and are often cited. The European Safety and Reliability Conference (ESREL) is also held annually and publishes proceedings on a variety of reliability topics, and conferences take place in other countries.

Journals on reliability have also appeared; some are referenced at the end of this chapter. Several books have been published on the subjects of reliability engineering and management; some of these are referenced at the end of other chapters.

Much of the reliability literature has tended to emphasize the mathematical and analytical aspects of the subject, with the result that reliability engineering is often considered by designers and others to be a rather esoteric subject. This is unfortunate, since it creates barriers to communication. It is important to emphasize the more practical aspects and to integrate reliability work into the overall management and engineering process. These aspects are covered in later chapters.

1.10 Organizations Involved in Reliability Work

Several organizations have been created to develop policies and methods in reliability engineering and to undertake research and training. Amid those organizations it is important to mention ASQ (American Society for Quality), which became a truly international organization with members in almost every country in the world. ASQ has many internal organizations including the Reliability Division which is the worldwide professional group with the focus on reliability specific training, education, networking and best practices.

1.11 Reliability as an Effectiveness Parameter

With the increasing cost and complexity of many modern systems, the importance of reliability as an effectiveness parameter, which should be specified and paid for, has become apparent. For example, a radar station, a process plant or an airliner must be available when required, and the cost of non-availability, particularly if it is unscheduled, can be very high. In the weapons field, if an anti-aircraft missile has a less than 100 % probability of functioning correctly throughout its engagement sequence, operational planners must consider deploying the appropriate extra quantity to provide the required level of defence. The Apollo project second stage rocket was powered by six rocket motors; any five would have provided sufficient impulse, but an additional motor was specified to cater for a possible failure of one. As it happened there were no failures, and every launch utilized an ‘unnecessary’ motor. These considerations apply equally to less complex systems, such as vending and copying machines, even if the failure costs are less dramatic in absolute terms.

As an effectiveness parameter, reliability can be ‘traded off’ against other parameters. Reliability generally affects availability, and in this context maintainability is also relevant. Reliability and maintainability are often related to availability by the formula:

$$\text{Availability} = \frac{\text{MTBF}}{\text{MTBF} + \text{MTTR}}$$

where MTTR is the mean time to repair. This is the simplest steady-state situation. It is clear that availability improvements can be achieved by improving either MTBF or MTTR. For example, automatic built-in test equipment can greatly reduce diagnostic times for electronic equipment, at a cost of a slight reduction in overall reliability and an increase in unit costs. Many other parameters can be considered in trade-offs, such as weight, redundancy, cost of materials, parts and processes, or reduction in performance.

The greatest difficulty in estimating relationships for reliability trade-offs derives from the fact that, whereas it is possible to estimate quite accurately such factors as the cost and weight penalties of built-in test equipment, the cost of materials and components, or the worth of a measurable performance parameter, the effect on reliability cannot generally be forecast accurately, and reliability measurements can at best be made only within statistical limits imposed by the amount of data available. Selection of trade-offs must therefore be very much a matter of experience of similar projects in the knowledge that wide margins of error can exist.

1.12 Reliability Programme Activities

What, then, are the actions that managers and engineers can take to influence reliability? One obvious activity already mentioned is quality assurance (QA), the whole range of functions designed to ensure that delivered products are compliant with the design. For many products, QA is sufficient to ensure high reliability, and we would not expect a company mass-producing simple diecastings for non-critical applications to employ reliability staff. In such cases the designs are simple and well proven, the environments in which the products will operate are well understood and the very occasional failure has no significant financial or operational effect. QA, together with craftsmanship, can provide adequate assurance for simple products or when the risks are known to be very low. Risks are low when safety margins can be made very large, as in most structural engineering. Reliability engineering disciplines may justifiably be absent in many types of product development and manufacture. QA disciplines are, however, essential elements of any integrated reliability programme.

A formal reliability programme is necessary whenever the risks or costs of failure are not low. We have already seen how reliability engineering developed as a result of the high costs of unreliability of military equipment, and later in commercial applications. Risks of failure usually increase in proportion to the number of components in a system, so reliability programmes are required for any product whose complexity leads to an appreciable risk.

An effective reliability programme should be based on the conventional wisdom of responsibility and authority being vested in one person. Let us call him or her the reliability programme manager. The responsibility must relate to a defined objective, which may be a maximum warranty cost figure, an MTBF to be demonstrated or a requirement that failure will not occur. Having an objective and the authority, how does the reliability programme manager set about his or her task, faced as he or she is with a responsibility based on uncertainties? This question will be answered in detail in subsequent chapters, but a brief outline is given below.

The reliability programme must begin at the earliest, conceptual phase of the project. It is at this stage that fundamental decisions are made, which can significantly affect reliability. These are decisions related to the risks involved in the specification (performance, complexity, cost, producibility, etc.), development time-scale, resources applied to evaluation and test, skills available, and other factors.

The shorter the project time-scale, the more important is this need, particularly if there will be few opportunities for an iterative approach. The activities appropriate to this phase are an involvement in the assessment of these trade-offs and the generation of reliability objectives. The reliability staff can perform these functions effectively only if they are competent to contribute to the give-and-take inherent in the trade-off negotiations, which may be conducted between designers and staff from manufacturing, marketing, finance, support and customer representatives.

As the project proceeds from initial study to detail design, the reliability risks are controlled by a formal, documented approach to the review of design and to the imposition of design rules relating to selection of components, materials and processes, stress protection, tolerancing, and so on. The objectives at this stage are to ensure that known good practices are applied, that deviations are detected and corrected, and that areas of uncertainty are highlighted for further action. The programme continues through the initial hardware manufacturing and test stages, by planning and executing tests to show up design weaknesses and to demonstrate achievement of specified requirements and by collecting, analysing and acting upon test and failure data. During production, QA activities ensure that the proven design is repeated, and further testing may be applied to eliminate weak items and to maintain confidence. The data collection, analysis and action process continues through the production and in-use phases. Throughout the product life cycle, therefore, the reliability is assessed, first by initial predictions based upon past experience in order to determine feasibility and to set objectives, then by refining the predictions as detail design proceeds and subsequently by recording performance during the test, production and in-use phases. This performance is fed back to generate corrective action, and to provide data and guidelines for future products.

The elements of a reliability programme are outlined in documents such as US MIL-STD-785, UK Defence Standard 00–40 and British Standard 5760 (see Bibliography). The activities are described fully in subsequent chapters.

1.13 Reliability Economics and Management

Obviously the reliability programme activities described can be expensive. Figure 1.7 is a commonly-described representation of the theoretical cost–benefit relationship of effort expended on reliability (and production quality) activities. It shows a U-shaped total cost curve with the minimum cost occurring at a reliability level somewhat lower than 100 %. This would be the optimum reliability, from the total cost point of view.

W.E. Deming presented a different model in his teaching on manufacturing quality (Deming, 1986). This is shown in Figure 1.8. He argued that, since less than perfect quality is the result of failures, all of which

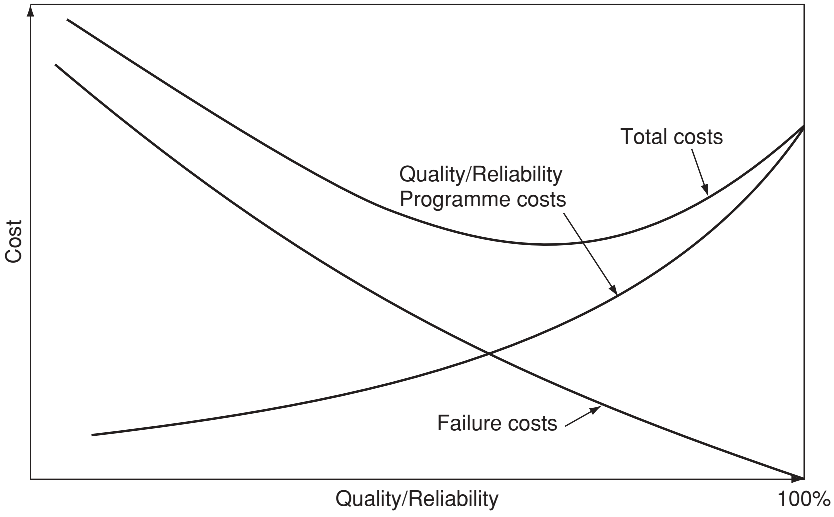


Figure 1.7 Reliability and life cycle costs (traditional view).

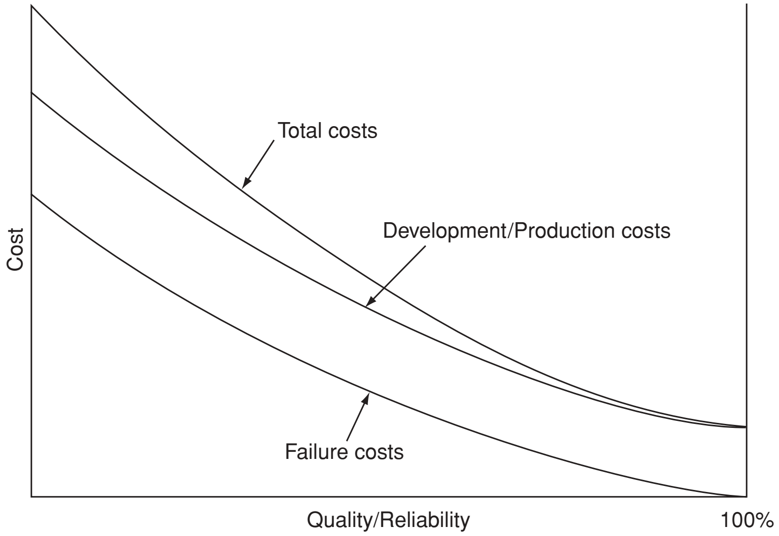


Figure 1.8 *Reliability/Quality and life cycle costs (Based on Deming's quality vs. cost model).*

have causes, we should not be tempted to assume that any level of quality is "optimum", but should ask 'what is the cost of preventing or correcting the causes, on a case by case basis, compared with the cost of doing nothing?' When each potential or actual cause is analysed in this way, it is usually found that it costs less to correct the causes than to do nothing. Thus total costs continue to reduce as quality is improved. This simple picture was the prime determinant of the post-war quality revolution in Japan, and formed the basis for the philosophy of *kaizen* (continuous improvement). 100 % quality was rarely achieved, but the levels that were achieved exceeded those of most Western competitors, and production costs were reduced.

In principle, the same argument applies to reliability: all efforts to improve reliability by identifying and removing potential causes of failures in service should result in cost savings later in the product life cycle, giving a net benefit in the longer term. In other words, an effective reliability programme represents an investment, usually with a large payback over a period of time. Unfortunately it is not easy to quantify the effects of given reliability programme activities, such as additional design analysis or testing, on achieved reliability. The costs (including those related to the effects on project schedules) of the activities are known, and they arise in the short term, but the benefits arise later and are often much less certain. However, achieving levels of reliability close to 100 % is often not realistic for complex products. Recent research on reliability cost modeling (Kleyner, 2010) showed that in practical applications the total cost curve is highly skewed to the right due to the increasing cost and diminishing return on further reliability improvements, as shown in Figure 1.9. The tight timescales and budgets of modern product development can also impact the amount of effort that can be applied. On the other hand there is often strong market pressure to achieve near perfect reliability. See more on cost of reliability in Chapters 14 and 17.

It is important to remember though that while achieving 100 % quality in manufacturing operations, or 100 % reliability in service, is extremely rare in real life applications, especially in high volume production, it should nevertheless be considered as an ultimate goal for any product development and production programme.

Achieving reliable designs and products requires a totally integrated approach, including design, test, production, as well as the reliability programme activities. The integrated engineering approach places high requirements for judgment and engineering knowledge on project managers and team members. Reliability specialists must play their parts as members of the team.

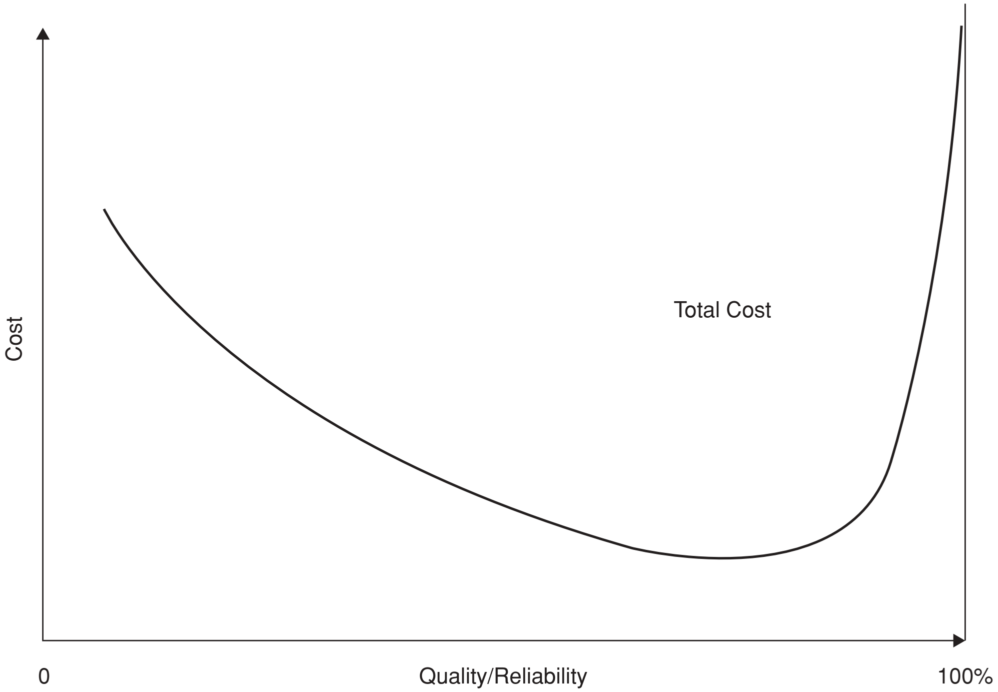


Figure 1.9 *Reliability and life cycle costs (practical applications).*

There are three kinds of engineering product, from the perspective of failure prevention:

1 *Intrinsically reliable components*, which are those that have high margins between their strength and the stresses that could cause failure, and which do not wear out within their practicable life times. Such items include nearly all electronic components (if properly applied), nearly all mechanical non-moving components, and all correct software.
2 *Intrinsically unreliable components*, which are those with low design margins or which wear out, such as badly applied components, light bulbs, turbine blades, parts that move in contact with others, like gears, bearings and power drive belts, and so on.
3 *Systems* which include many components and interfaces, like cars, dishwashers, aircraft, and so on, so that there are many possibilities for failures to occur, particularly across interfaces (e.g. inadequate electrical overstress protection, vibration nodes at weak points, electromagnetic interference, software that contains errors, and so on).

It is the task of design engineers to ensure that all components are correctly applied, that margins are adequate (particularly in relation to the possible extreme values of strength and stress, which are often variable), that wearout failure modes are prevented during the expected life (by safe life design, maintenance, etc.), and that system interfaces cannot lead to failure (due to interactions, tolerance mismatches, etc.). Because achieving all this on any modern engineering product is a task that challenges the capabilities of the very best engineering teams, it is almost certain that aspects of the initial design will fall short of the ‘intrinsically reliable’ criterion. Therefore we must submit the design to analyses and tests in order to show not only that it works, but also to show up the features that might lead to failures. When we find out what these are we must redesign and re-test, until the final design is considered to meet the criterion.

Then the product has to be manufactured. In principle, every one should be identical and correctly made. Of course this is not achievable, because of the inherent variability of all manufacturing processes, whether performed by humans or by machines. It is the task of the manufacturing people to understand and control variation, and to implement inspections and tests that will identify non-conforming product.

For many engineering products the quality of operation and maintenance also influence reliability.

The essential points that arise from this brief and obvious discussion of failures are that:

1 Failures are caused primarily by people (designers, suppliers, assemblers, users, maintainers). Therefore the achievement of reliability is essentially a management task, to ensure that the right people, skills, teams and other resources are applied to prevent the creation of failures.

2 Reliability (and quality) specialists cannot by themselves effectively ensure the prevention of failures. High reliability and quality can be achieved only by effective team working by all involved.

3 There is no fundamental limit to the extent to which failures can be prevented. We can design and build for ever-increasing reliability.

Deming explained how, in the context of manufacturing quality, there is no point at which further improvement leads to higher costs. This is, of course, even more powerfully true when considered over the whole product life cycle, so that efforts to ensure that designs are intrinsically reliable, by good design, thorough analysis and effective development testing, can generate even higher pay-offs than improvements in production quality. The '*kaizen*' (continuous improvement) principle is even more effective when applied to up-front engineering.

The creation of reliable products is, therefore, primarily a management task. Guidance on reliability programme management and costs is covered in Chapter 17.

Questions

1. Define (a) failure rate, and (b) hazard rate. Explain their application to the reliability of components and repairable systems. Discuss the plausibility of the 'bathtub curve' in both contexts.
2. a Explain the theory of component failures derived from the interaction of stress (or load) and strength distributions. Explain how this theory relates to the behaviour of the component hazard function.

   b Discuss the validity of the 'bathtub curve' when used to describe the failure characteristics of non-repairable components.
3. What are the main objectives of a reliability engineering team working on an engineering development project? Describe the important skills and experience that should be available within the team.
4. Briefly list the most common basic causes of failures of engineering products.
5. It is sometimes claimed that increasing quality and reliability beyond levels that have been achieved in the past is likely to be uneconomic, due to the costs of the actions that would be necessary. Present the argument against this belief. Illustrate it with an example from your own experience.
6. Describe the difference between repairable and non-repairable items. What kind of effect might this difference have on reliability? List examples of repairable and non-repairable items in your everyday life.
7. Explain the difference between reliability and durability and how they can be specified in a product development programme.

8. a List the potential economic outcomes of poor reliability, and identify which costs are directly quantifiable and which are intangible. Explain how they can be minimised, and discuss the extent to which very high reliability (approaching zero failures) is achievable in practice.
   b What are the major factors that might limit the achievement of very high reliability?
9. After processing the existing programme cost data and running a regression model on the previous projects, the cost of product development and manufacturing (CDM) has been estimated to follow the equation: CDM = $0.8 million +$3.83 million × R^2 (R is the achieved product reliability at service life and is expected to be above 90 %). The cost of failure (CF) has been estimated as the sum of fixed cost of $40 000 plus variable cost of $150 per failure. The total number of the expected failures is n × (1 − R), where n is the total number of produced units. Considering that the production volume is expected to be 50 000 units, estimate the optimal target reliability and the total cost of the programme.
10. Select an everyday item (coffee maker, lawnmower, bicycle, mobile phone, CD player, computer, refrigerator, microwave oven, cooking stove, etc.).
   a Discuss the ways this item can potentially fail. What can be done to prevent those failures?
   b Based on the Figures 1.3 and 1.4, what would be an example of the load and strength for a critical component within this item? Do you expect load and strength for this component to be time-dependent?

Bibliography

British Standard, BS 4778 (1991) *Glossary of Terms Used in Quality Assurance* (including reliability and maintainability). British Standards Institution, London.

British Standard, BS 5760 (1996) *Reliability of Systems, Equipments and Components*. British Standards Institution, London.

Deming, W. (1986) *Out of the Crisis*, MIT University Press (originally published under the title *Quality, Productivity and Competitive Position*).

Drucker, P. (1995) *The Practice of Management*. Heinemann.

Kleyner, A. (2010) *Determining Optimal Reliability Targets*, Lambert Academic Publishing.

Misra, K. (ed.) (2008) *The Handbook of Performability Engineering*, Springer-Verlag, London.

UK Defence Standard 00–40. *The Management of Reliability and Maintainability*. HMSO.

US MIL-STD-721. *Definitions of Effectiveness Terms for Reliability, Maintainability, Human Factors and Safety*. National Technical Information Service, Springfield, Virginia.

US MIL-STD-785. *Reliability Programs for Systems and Equipment*. National Technical Information Service, Springfield, Virginia (suspended in 1976).

Periodic Publications

International Journal of Performability Engineering. Available at: http://www.ijpe-online.com/

Microelectronics Reliability, Elsevier (published monthly).

Proceedings of the US Reliability and Maintainability Symposium (RAMS). American Society for Quality and IEEE (published annually).

Quality and Reliability Engineering International, Wiley (published quarterly).

Reliability Engineering and Systems Safety, Elsevier (published monthly).

Transactions on Reliability, Institute of Electrical and Electronics Engineers (IEEE) (published quarterly).

2
Reliability Mathematics

2.1 Introduction

The methods used to quantify reliability are the mathematics of probability and statistics. In reliability work we are dealing with uncertainty. As an example, data may show that a certain type of power supply fails at a constant average rate of once per 10^7 h. If we build 1000 such units, and we operate them for 100 h, we cannot say with certainty whether any will fail in that time. We can, however, make a statement about the *probability* of failure. We can go further and state that, within specified statistical *confidence limits*, the probability of failure lies between certain values above and below this probability. If a sample of such equipment is tested, we obtain data which are called *statistics*.

Reliability statistics can be broadly divided into the treatment of *discrete functions, continuous functions* and *point processes*. For example, a switch may either work or not work when selected or a pressure vessel may pass or fail a test—these situations are described by discrete functions. In reliability we are often concerned with two-state discrete systems, since equipment is in either an operational or a failed state. Continuous functions describe those situations which are governed by a continuous variable, such as time or distance travelled. The electronic equipment mentioned above would have a reliability function in this class. The distinction between discrete and continuous functions is one of how the problem is treated, and not necessarily of the physics or mechanics of the situation. For example, whether or not a pressure vessel fails a test may be a function of its age, and its reliability could therefore be treated as a continuous function. The statistics of point processes are used in relation to repairable systems, when more than one failure can occur in a time continuum. The choice of method will depend upon the problem and on the type of data available.

2.2 Variation

Reliability is influenced by variability, in parameter values such as resistance of resistors, material properties, or dimensions of parts. Variation is inherent in all manufacturing processes, and designers should understand the nature and extent of possible variation in the parts and processes they specify. They should know how to measure and control this variation, so that the effects on performance and reliability are minimized.

Variation also exists in the environments that engineered products must withstand. Temperature, mechanical stress, vibration spectra, and many other varying factors must be considered.

Practical Reliability Engineering, Fifth Edition. Patrick D. T. O'Connor and Andre Kleyner.
© 2012 John Wiley & Sons, Ltd. Published 2012 by John Wiley & Sons, Ltd.

Statistical methods provide the means for analysing, understanding and controlling variation. They can help us to create designs and develop processes which are intrinsically reliable in the anticipated environments over their expected useful lives.

Of course, it is not necessary to apply statistical methods to understand every engineering problem, since many are purely deterministic or easily solved using past experience or information available in sources such as databooks, specifications, design guides, and in known physical relationships such as Ohm's law. However, there are also many situations in which appropriate use of statistical techniques can be very effective in optimizing designs and processes, and for solving quality and reliability problems.

2.2.1 A Cautionary Note

Whilst statistical methods can be very powerful, economic and effective in reliability engineering applications, they must be used in the knowledge that variation in engineering is in important ways different from variation in most natural processes, or in repetitive engineering processes such as repeated, in-control machining or diffusion processes. Such processes are usually:

– Constant in time, in terms of the nature (average, spread, etc.) of the variation.
– Distributed in a particular way, describable by a mathematical function known as the normal distribution (which will be described later in this chapter).

In fact, these conditions often do not apply in engineering. For example:

– A component supplier might make a small change in a process, which results in a large change (better or worse) in reliability. The change might be deliberate or accidental, known or unknown. Therefore the use of past data to forecast future reliability, using purely statistical methods, might be misleading.
– Components might be selected according to criteria such as dimensions or other measured parameters. This can invalidate the normal distribution assumption on which much of the statistical method is based. This might or might not be important in assessing the results.
– A process or parameter might vary in time, continuously or cyclically, so that statistics derived at one time might not be relevant at others.
– Variation is often deterministic by nature, for example spring deflection as a function of force, and it would not always be appropriate to apply statistical techniques to this sort of situation.
– Variation in engineering can arise from factors that defy mathematical treatment. For example, a thermostat might fail, causing a process to vary in a different way to that determined by earlier measurements, or an operator or test technician might make a mistake.
– Variation can be discontinuous. For example, a parameter such as a voltage level may vary over a range, but could also go to zero.

These points highlight the fact that variation in engineering is caused to a large extent by people, as designers, makers, operators and maintainers. The behaviour and performance of people is not as amenable to mathematical analysis and forecasting as is, say, the response of a plant crop to fertilizer or even weather patterns to ocean temperatures. Therefore the human element must always be considered, and statistical analysis must not be relied on without appropriate allowance being made for the effects of factors such as motivation, training, management, and the many other factors that can influence reliability.

Finally, it is most important to bear in mind, in any application of statistical methods to problems in science and engineering, that ultimately all cause and effect relationships have explanations, in scientific theory, engineering design, process or human behaviour, and so on. Statistical techniques can be very useful in helping us to understand and control engineering situations. However, they do not by themselves provide explanations. We must always seek to understand causes of variation, since only then can we really be in control. *See the quotations on the flyleaf, and think about them.*

2.3 Probability Concepts

Any event has a probability of occurrence, which can be in the range 0–1. A zero probability means that the event will not occur; a probability of 1 means that it will occur. A coin has a 0.5 (even) probability of landing heads, and a die has a 1/6 probability of giving any one of the six numbers. Such events are *independent*, that is, the coin and the die logically have no memory, so whatever has been thrown in the past cannot affect the probability of the next throw. No 'system' can beat the statistics of these situations; waiting for a run of blacks at roulette and then betting on reds only appears to work because the gamblers who won this way talk about it, whilst those who lost do not.

With coins, dice and roulette wheels we can predict the probability of the outcome from the nominal nature of the system. A coin has two sides, a die six faces, a roulette wheel equal numbers of reds and blacks. Assuming that the coin, die and wheel are fair, these outcomes are also *unbiased*, that is, they are all equally probable. In other words, they occur *randomly*.

With many systems, such as the sampling of items from a production batch, the probabilities can only be determined from the statistics of previous experience.

We can define probability in two ways:

1 If an event can occur in N equally likely ways, and if the event with attribute A can happen in n of these ways, then the probability of A occurring is

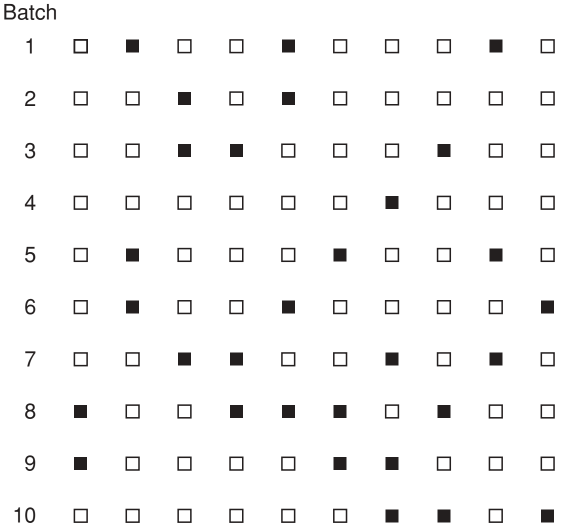
$$P(\mathrm{A}) = \frac{n}{N}$$

2 If, in an experiment, an event with attribute A occurs n times out of N experiments, then as n becomes large, the probability of event A approaches n/N, that is,

$$P(\mathrm{A}) = \lim_{n\to\infty}\left(\frac{n}{N}\right)$$

The first definition covers the cases described earlier, that is, equally likely independent events such as rolling dice. The second definition covers typical cases in quality control and reliability. If we test 100 items and find that 30 are defective, we may feel justified in saying that the probability of finding a defective item in our next test is 0.30, or 30 %.

However, we must be careful in making this type of assertion. The probability of 0.30 of finding a defective item in our next test may be considered as our *degree of belief*, in this outcome, limited by the size of the sample. This leads to a third, subjective, definition of probability. If, in our tests of 100 items, seven of the defectives had occurred in a particular batch of ten and we had taken corrective action to improve the process so that such a problem batch was less likely to occur in future, we might assign some lower probability to the next item being defective. This subjective approach is quite valid, and is very often necessary in quality control and reliability work. Whilst it is important to have an understanding of the rules of probability, there are usually so many variables which can affect the properties of manufactured items that we must always keep an open mind about statistically derived values. We must ensure that the sample from which statistics have been derived represents the new sample, or the overall population, about which we plan to make an assertion based upon our sample statistics.



Batch										
1	□	■	□	□	■	□	□	□	■	□
2	□	□	■	□	■	□	□	□	□	□
3	□	□	■	■	□	□	□	■	□	□
4	□	□	□	□	□	□	■	□	□	□
5	□	■	□	□	□	■	□	□	■	□
6	□	■	□	□	■	□	□	□	□	■
7	□	□	■	■	□	□	■	□	■	□
8	■	□	□	■	■	■	□	■	□	□
9	■	□	□	□	□	■	■	□	□	□
10	□	□	□	□	□	□	■	■	□	■

Figure 2.1 *Samples with defectives (black squares).*

A sample represents a population if all the members of the population have an equal chance of being sampled. This can be achieved if the sample is selected so that this condition is fulfilled. Of course in engineering this is not always practicable; for example, in reliability engineering we often need to make an assertion about items that have not yet been produced, based upon statistics from prototypes.

To the extent that the sample is not representative, we will alter our assertions. Of course, subjective assertions can lead to argument, and it might be necessary to perform additional tests to obtain more data to use in support of our assertions. If we do perform more tests, we need to have a method of interpreting the new data in relation to the previous data: we will cover this aspect later.

The assertions we can make based on sample statistics can be made with a degree of confidence which depends upon the size of the sample. If we had decided to test ten items after introducing a change to the process, and found one defective, we might be tempted to assert that we have improved the process, from 30 % defectives being produced to only 10 %. However, since the sample is now much smaller, we cannot make this assertion with as high confidence as when we used a sample of 100. In fact, the true probability of any item being defective might still be 30 %, that is, the population might still contain 30 % defectives.

Figure 2.1 shows the situation as it might have occurred, over the first 100 tests. The black squares indicate defectives, of which there are 30 in our batch of 100. If these are randomly distributed, it is possible to pick a sample batch of ten which contains fewer (or more) than three defectives. In fact, the smaller the sample, the greater will be the sample-to-sample variation about the population average, and the confidence associated with any estimate of the population average will be accordingly lower. The derivation of confidence limits is covered later in this chapter.

2.4 Rules of Probability

In order to utilize the statistical methods used in reliability engineering, it is necessary to understand the basic notation and rules of probability. These are:

1 The probability of obtaining an outcome A is denoted by P(A), and so on for other outcomes.
2 The *joint* probability that A *and* B occur is denoted by P(AB).

3 The probability that A *or* B occurs is denoted by $P(\mathrm{A}+\mathrm{B})$.

4 The *conditional* probability of obtaining outcome A, *given that B has occurred*, is denoted by P (A | B).

5 The probability of the complement, that is, of A *not* occurring, is $P(\bar{\mathrm{A}}) = 1 - P(\mathrm{A})$

6 If (and only if) events A and B are *independent*, then

$$P(\mathrm{A}|\mathrm{B}) = P(\mathrm{A}|\bar{\mathrm{B}}) = P(\mathrm{A})$$

and

$$P(\mathrm{B}|\mathrm{A}) = P(\mathrm{B}|\bar{\mathrm{A}}) = P(\mathrm{B}) \tag{2.1}$$

that is, P(A) is unrelated to whether or not B occurs, and vice versa.

7 The joint probability of the occurrence of two independent events A and B is the product of the individual probabilities:

$$P(\mathrm{AB}) = P(\mathrm{A})P(\mathrm{B}) \tag{2.2}$$

This is also called the *product rule* or *series rule*. It can be extended to cover any number of independent events. For example, in rolling a die, the probability of obtaining any given sequence of numbers in three throws is

$$\frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} = \frac{1}{216}$$

8 If events A and B are *dependent*, then

$$P(\mathrm{AB}) = P(\mathrm{A})P(\mathrm{B}|\mathrm{A}) = P(\mathrm{B})P(\mathrm{A}|\mathrm{B}) \tag{2.3}$$

that is, the probability of A occurring times the probability of B occurring given that A has already occurred, or vice versa.

If $P(\mathrm{A}) \neq 0$, (2.3) can be rearranged to

$$P(\mathrm{B}|\mathrm{A}) = \frac{P(\mathrm{AB})}{P(\mathrm{A})} \tag{2.4}$$

9 The probability of any one of two events A or B occurring is

$$P(\mathrm{A}+\mathrm{B}) = P(\mathrm{A}) + P(\mathrm{B}) - P(\mathrm{AB}) \tag{2.5}$$

10 The probability of A or B occurring, if A and B are independent, is

$$P(\mathrm{A}+\mathrm{B}) = P(\mathrm{A}) + P(\mathrm{B}) - P(\mathrm{A})P(\mathrm{B}) \tag{2.6}$$

The derivation of this equation can be shown by considering the system shown in Figure 2.2, in which either A or B, or A and B, must work for the system to work. If we denote the system success probability as P_s, then the failure probability, $P_f = 1 - P_s$. The system failure probability is the joint probability of A and B failing, that is,

$$\begin{aligned} P_f &= [1 - P(\mathrm{A})][1 - P(\mathrm{B})] \\ &= 1 - P(\mathrm{A}) - P(\mathrm{B}) + P(\mathrm{A})P(\mathrm{B}) \\ P_s &= 1 - P_f = P(\mathrm{A}+\mathrm{B}) = P(\mathrm{A}) + P(\mathrm{B}) - P(\mathrm{A})P(\mathrm{B}) \end{aligned}$$

11 If events A and B are *mutually exclusive*, that is, A and B cannot occur simultaneously, then

$$P(\mathrm{AB}) = 0$$

and

$$P(\mathrm{A}+\mathrm{B}) = P(\mathrm{A}) + P(\mathrm{B}) \tag{2.7}$$

12 If multiple, mutually exclusive probabilities of outcomes B_i jointly give a probability of outcome A, then

$$P(\mathrm{A}) = \sum_i P(\mathrm{AB}_i) = \sum_i P(\mathrm{A}|\mathrm{B}_i)P(\mathrm{B}_i) \tag{2.8}$$

13 Rearranging Eq. (2.3)

$$P(\mathrm{AB}) = P(\mathrm{A})P(\mathrm{B}|\mathrm{A}) = P(\mathrm{B})P(\mathrm{A}|\mathrm{B})$$

we obtain

$$P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{A})P(\mathrm{B}|\mathrm{A})}{P(\mathrm{B})} \tag{2.9}$$

This is a simple form of *Bayes' theorem*. A more general expression is

$$P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{A})P(\mathrm{B}|\mathrm{A})}{\sum_i P(\mathrm{B}|\mathrm{E}_i)P(\mathrm{E}_i)} \tag{2.10}$$

where E_i is the ith event.

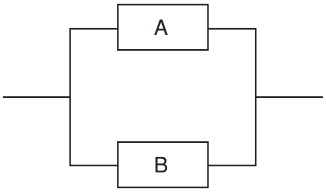


Figure 2.2 *Dual redundant system.*

Example 2.1

The reliability of a missile is 0.85. If a salvo of two missiles is fired, what is the probability of at least one hit? (Assume independence of missile hits.)

Let A be the event 'first missile hits' and B the event 'second missile hits'. Then

$$P(\mathrm{A}) = P(\mathrm{B}) = 0.85$$
$$P(\bar{\mathrm{A}}) = P(\bar{\mathrm{B}}) = 0.15$$

There are four possible, mutually exclusive outcomes, AB, A$\bar{\mathrm{B}}$, $\bar{\mathrm{A}}$B; $\bar{\mathrm{A}}$ $\bar{\mathrm{B}}$. The probability of both missing, from Eq. (2.2), is

$$\begin{aligned} P(\bar{\mathrm{A}})P(\bar{\mathrm{B}}) &= P(\bar{\mathrm{A}}\,\bar{\mathrm{B}}) \\ &= 0.15^2 = 0.0225 \end{aligned}$$

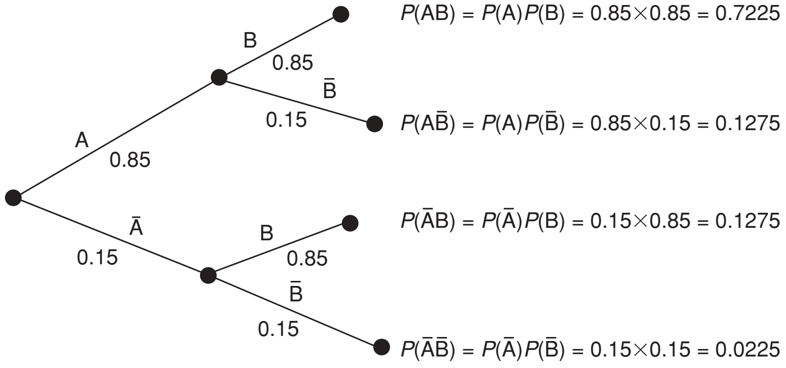
Therefore the probability of at least one hit is

$$P_{\mathrm{s}} = 1 - 0.0225 = 0.9775$$

We can derive the same result by using Eq. (2.6):

$$\begin{aligned} P(\mathrm{A}+\mathrm{B}) &= P(\mathrm{A}) + P(\mathrm{B}) - P(\mathrm{A})P(\mathrm{B}) \\ &= 0.85 + 0.85 - 0.85^2 = 0.9775 \end{aligned}$$

Another way of deriving this result is by using the *sequence tree diagram*:


A
0.85
Ā
0.15
B
0.85
B̄
0.15
B
0.85
B̄
0.15
P(AB) = P(A)P(B) = 0.85×0.85 = 0.7225
P(AB̄) = P(A)P(B̄) = 0.85×0.15 = 0.1275
P(ĀB) = P(Ā)P(B) = 0.15×0.85 = 0.1275
P(ĀB̄) = P(Ā)P(B̄) = 0.15×0.15 = 0.0225

The probability of a hit is then derived by summing the products of each path which leads to at least one hit. We can do this since the events defined by each path are mutually exclusive.

$$P(\mathrm{AB}) + P(\mathrm{A}\bar{\mathrm{B}}) + P(\bar{\mathrm{A}}\mathrm{B}) = 0.9775$$

(Note that the sum of all the probabilities is unity.)

Example 2.2

In Example 2.1 the missile hits are not independent, but are dependent, so that if the first missile fails the probability that the second will also fail is 0.2. However, if the first missile hits, the hit probability of the second missile is unchanged at 0.85. What is the probability of at least one hit?

$$\begin{aligned} P(\mathrm{A}) &= 0.85 \\ P(\mathrm{B}|\mathrm{A}) &= 0.85 \\ P(\bar{\mathrm{B}}|\mathrm{A}) &= 0.15 \\ P(\bar{\mathrm{B}}|\bar{\mathrm{A}}) &= 0.2 \\ P(\mathrm{B}|\bar{\mathrm{A}}) &= 0.8 \end{aligned}$$

The probability of at least one hit is

$$P(\mathrm{AB}) + P(\bar{\mathrm{A}}\mathrm{B}) + P(\bar{\mathrm{B}}\mathrm{A})$$

Since A, B and A$\bar{\mathrm{B}}$ are independent,

$$\begin{aligned} P(\mathrm{AB}) &= P(\mathrm{A})P(\mathrm{B}) \\ &= 0.85 \times 0.85 = 0.7225 \end{aligned}$$

and

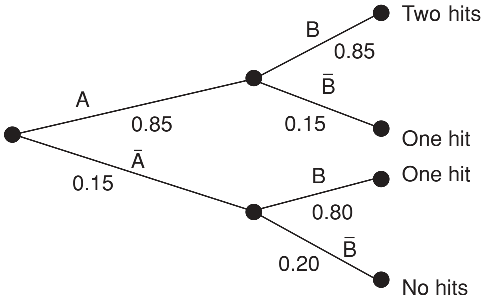
$$\begin{aligned} P(\mathrm{A}\bar{\mathrm{B}}) &= P(\mathrm{A})P(\bar{\mathrm{B}}) \\ &= 0.85 \times 0.15 = 0.1275 \end{aligned}$$

Since $\bar{\mathrm{A}}$ and B are dependent, from Eq. (2.3),

$$\begin{aligned} P(\bar{\mathrm{A}}\mathrm{B}) &= P(\bar{\mathrm{A}})P(\mathrm{B}|\bar{\mathrm{A}}) \\ &= 0.15 \times 0.8 = 0.12 \end{aligned}$$

and the sum of these probabilities is 0.97.

This result can also be derived by using a sequence tree diagram:



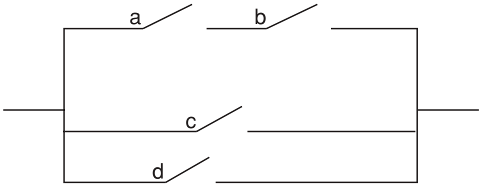
As in Example 2.1, the probability of at least one hit is calculated by adding the products of each path leading to at least one hit, that is,

$$\begin{aligned}&P(\mathrm{A})P(\mathrm{B}) + P(\mathrm{A})P(\bar{\mathrm{B}}) + P(\bar{\mathrm{A}})P(\mathrm{B})\\ &\quad = (0.85 \times 0.85) + (0.85 \times 0.15) + (0.15 \times 0.80) = 0.97\end{aligned}$$

Example 2.3

In the circuit shown, the probability of any switch being closed is 0.8 and all events are independent. (a) What is the probability that a circuit will exist? (b) Given that a circuit exists, what is the probability that switches a and b are closed?

Let the events that a, b, c and d are closed be A, B, C and D. Let X denote the event that the circuit exists.



(a) $\mathrm{X} = \mathrm{AB} + (\mathrm{C} + \mathrm{D})$

$$\begin{aligned}P(\mathrm{X}) &= P(\mathrm{AB}) + P(\mathrm{C}+\mathrm{D}) - P(\mathrm{AB})P(\mathrm{C}+\mathrm{D})\\ P(\mathrm{AB}) &= P(\mathrm{A})P(\mathrm{B})\\ &= 0.8 \times 0.8 = 0.64\\ P(\mathrm{C}+\mathrm{D}) &= P(\mathrm{C}) + P(\mathrm{D}) - P(\mathrm{C})P(\mathrm{D})\\ &= 0.8 + 0.8 - 0.64 = 0.96\end{aligned}$$

Therefore

$$P(\mathrm{X}) = 0.64 + 0.96 - (0.96 \times 0.64) = 0.9856$$

(b) From Eq. (2.4),

$$P(\mathrm{AB}|\mathrm{X}) = \frac{P(\mathrm{ABX})}{P(\mathrm{X})}$$

A and B jointly give X. Therefore, from Eq. (2.8),

$$P(\mathrm{ABX}) = \mathrm{P}(\mathrm{AB})$$

So

$$\begin{aligned}P(\mathrm{AB}|\mathrm{X}) &= \frac{P(\mathrm{AB})}{P(\mathrm{X})} = \frac{P(\mathrm{A})P(\mathrm{B})}{P(\mathrm{X})}\\ &= \frac{0.8 \times 0.8}{0.9856} = 0.6494\end{aligned}$$

Example 2.4

A test set has a 98 % probability of correctly classifying a defective item and a 4 % probability of classifying a good item as defective. If in a batch of items tested 3 % are actually defective, what is the probability that when an item is classified as defective, it is truly defective?

Let D represent the event that an item is defective and C represent the event that an item is classified defective. Then

$$\begin{aligned} P(\mathrm{D}) &= 0.03 \\ P(\mathrm{C}|\mathrm{D}) &= 0.98 \\ P(\mathrm{C}|\bar{\mathrm{D}}) &= 0.04 \end{aligned}$$

We need to determine P(D|C). Using Eq. (2.10),

$$\begin{aligned} P(\mathrm{D}|\mathrm{C}) &= \frac{P(\mathrm{D})P(\mathrm{C}|\mathrm{D})}{P(\mathrm{C}|\mathrm{D})P(\mathrm{D}) + P(\mathrm{C}|\bar{\mathrm{D}})P(\bar{\mathrm{D}})} \\ &= \frac{(0.03)(0.98)}{(0.98)(0.03) + (0.04)(0.97)} = 0.43 \end{aligned}$$

This indicates the importance of a test equipment having a high probability of correctly classifying good items as well as bad items.

More practical applications of the Bayesian statistical approach to reliability can be found in Martz and Waller (1982) or Kleyner *et al.* (1997).

2.5 Continuous Variation

The variation of parameters in engineering applications (machined dimensions, material strengths, transistor gains, resistor values, temperatures, etc.) are conventionally described in two ways. The first, and the simplest, is to state maximum and minimum values, or tolerances. This provides no information on the nature, or shape, of the actual distribution of values. However, in many practical cases, this is sufficient information for the creation of manufacturable, reliable designs.

The second approach is to describe the nature of the variation, using data derived from measurements. In this section we will describe the methods of statistics in relation to describing and controlling variation in engineering.

If we plot measured values which can vary about an average (e.g. the diameters of machined parts or the gains of transistors) as a histogram, for a given sample we may obtain a representation such as Figure 2.3(a).

In this case 30 items have been measured and the frequencies of occurrence of the measured values are as shown. The values range from 2 to 9, with most items having values between 5 and 7. Another random sample of 30 from the same population will usually generate a different histogram, but the general shape is likely to be similar, for example, Figure 2.3(b). If we plot a single histogram showing the combined data of many such samples, but this time show the values in measurement intervals of 0.5, we get Figure 2.3(c). Note that now we have used a percentage frequency scale. We now have a rather better picture of the distribution of values, as we have more information from the larger sample. If we proceed to measure a large number and we further reduce the measurement interval, the histogram tends to a curve which describes the population *probability density function* (pdf) or simply the *distribution* of values. Figure 2.4 shows a general *unimodal* probability distribution, f(x) being the probability density of occurrence, related to the variable x.

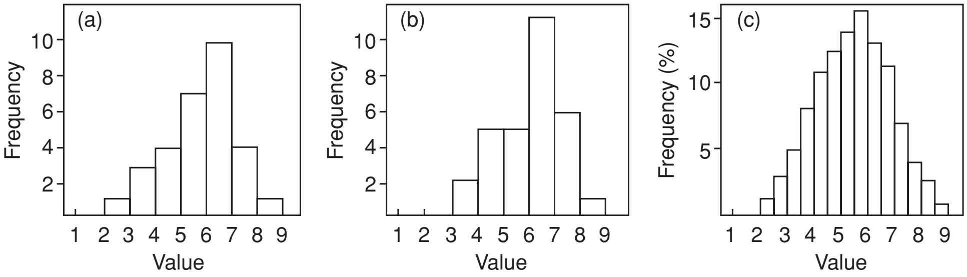


Figure 2.3 *(a) Frequency histogram of a random sample, (b) frequency histogram of another random sample from the same population, (c) data of many samples shown with measurement intervals of 0.5.*

The value of x at which the distribution peaks is called the *mode*. Multimodal distributions are encountered in reliability work as well as unimodal distributions. However, we will deal only with the statistics of unimodal distributions in this book, since multimodal distributions are usually generated by the combined effects of separate unimodal distributions.

The area under the curve is equal to unity, since it describes the total probability of all possible values of x, as we have defined a probability which is a certainty as being a probability of one. Therefore

$$\int_{-\infty}^{\infty} \mathrm{f}(x)\,\mathrm{d}x = 1 \tag{2.11}$$

The probability of a value falling between any two values x_1 and x_2 is the area bounded by this interval, that is,

$$P(x_1 < x < x_2) = \int_{x_1}^{x_2} \mathrm{f}(x)\,\mathrm{d}x \tag{2.12}$$

To describe a pdf we normally consider four aspects:

1 The *central tendency*, about which the distribution is grouped.
2 The *spread*, indicating the extent of variation about the central tendency.
3 The *skewness*, indicating the lack of symmetry about the central tendency. Skewness equal to zero is a characteristic of a symmetrical distribution. Positive skewness indicates that the distribution has a longer tail to the right (see for example Figure 2.5) and negative skewness indicates the opposite.

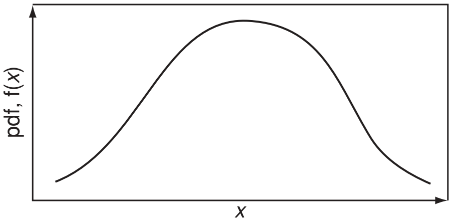


Figure 2.4 *Continuous probability distribution.*

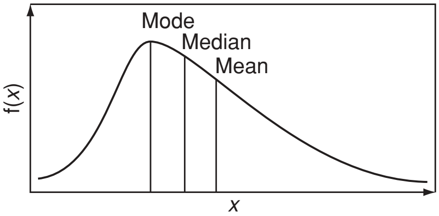


Figure 2.5 Measures of central tendency.

4 The *kurtosis*, indicating the 'peakedness' of the pdf In general terms kurtosis characterizes the relative peakedness or flatness of a distribution compared to the normal distribution. Positive kurtosis indicates a relatively peaked distribution. Negative kurtosis indicates a relatively flat distribution.

2.5.1 Measures of Central Tendency

For a sample containing *n* items the sample *mean* is denoted by x̄:

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} \tag{2.13}$$

The sample mean can be used to *estimate* the population mean, which is the average of all possible outcomes. For a continuous distribution, the mean is derived by extending this idea to cover the range $-\infty$ to $+\infty$.

The mean of a distribution is usually denoted by μ. The mean is also referred to as the *location parameter, average value* or *expected value*, $E(x)$.

$$\mu = \int_{-\infty}^{\infty} x\mathrm{f}(x)\,\mathrm{d}x \tag{2.14}$$

This is analogous to the centre of gravity of the pdf The *estimate* of a population mean from sample data is denoted by $\widehat{\mu}$.

Other measures of central tendency are the *median*, which is the mid-point of the distribution, that is, the point at which half the measured values fall to either side, and the *mode*, which is the value (or values) at which the distribution peaks. The relationship between the mean, median and mode for a right-skewed distribution is shown in Figure 2.5. For a symmetrical distribution, the three values are the same, whilst for a left-skewed distribution the order of values is reversed.

2.5.2 Spread of a Distribution

The spread, or dispersion, that is, the extent to which the values which make up the distribution vary, is measured by its *variance*. For a sample size *n* the variance, Var (x) or $E(x - \bar{x})^2$, is given by

$$\mathrm{Var}(x) = E(x - \bar{x})^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} \tag{2.15}$$

Where sample variance is used to estimate the population variance, we use (*n* – 1) in the denominator of Eq. (2.15) instead of *n*, as it can be shown to provide a better estimate. The *estimate of population variance* from a sample is denoted $\hat{\sigma}^2$ where

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{(x - \bar{x})^2}{n - 1} \tag{2.16}$$

The *population* variance σ^2, for a finite population *N*, is given by

$$\sigma^2 = \frac{\sum\limits_{i=1}^{N} (x_i - \mu)^2}{N} \tag{2.17}$$

For a continuous distribution it is:

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 \mathrm{f}(x)\,\mathrm{d}x \tag{2.18}$$

σ is called the *standard deviation* (SD) and is frequently used in practice instead of the variance. It is also referred to as the *scale parameter*. σ^2 is the second moment about the mean and is analogous to a radius of gyration.

The third and fourth moments about the mean give the skewness and kurtosis mentioned before. Since we will not make use of these parameters in this book, the reader is referred to more advanced statistical texts for their derivation (e.g. Hines and Montgomery, 1990).

2.5.3 The Cumulative Distribution Function

The cumulative distribution function (cdf), F(*x*), gives the probability that a measured value will fall between –∞ and *x*:

$$\mathrm{F}(x) = \int_{-\infty}^{x} \mathrm{f}(x)\,\mathrm{d}x \tag{2.19}$$

Figure 2.6 shows the typical ogive form of the cdf with $\mathrm{F}(x) \to 1$ as $x \to \infty$.

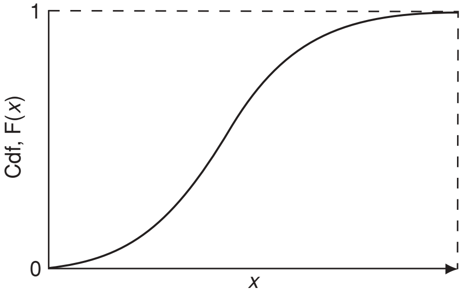


Figure 2.6 *Typical cumulative distribution function (cdf).*

2.5.4 Reliability and Hazard Functions

In reliability engineering we are concerned with the probability that an item will survive for a stated interval (e.g. time, cycles, distance, etc.), that is, that there is no failure in the interval (0 to x). This is the *reliability*, and it is given by the reliability function $\mathrm{R}(x)$. From this definition, it follows that

$$\mathrm{R}(x) = 1 - \mathrm{F}(x) = \int_{x}^{\infty} \mathrm{f}(x)\,\mathrm{d}x = 1 - \int_{-\infty}^{x} \mathrm{f}(x)\,\mathrm{d}x \tag{2.20}$$

The *hazard function* or *hazard rate* $\mathrm{h}(x)$ is the conditional probability of failure in the interval x to $(x + \mathrm{d}x)$, given that there was no failure by x:

$$\mathrm{h}(x) = \frac{\mathrm{f}(x)}{\mathrm{R}(x)} = \frac{\mathrm{f}(x)}{1 - \mathrm{F}(x)} \tag{2.21}$$

The *cumulative hazard function* $\mathrm{H}(x)$ is given by

$$\mathrm{H}(x) = \int_{-\infty}^{x} \mathrm{h}(x)\,\mathrm{d}x = \int_{-\infty}^{x} \frac{\mathrm{f}(x)}{1 - \mathrm{F}(x)}\,\mathrm{d}x \tag{2.22}$$

Figure 2.7 illustrates the relationship between the failure probability density function (pdf), reliability R(t), and failure function F(t). At any point of time the area under the curve left of t would represent the fraction of the population expected to fail F(t) and area to the right the fraction expected to survive R(t).

Please note, that in engineering we do not usually encounter measured values below zero and the lower limit of the definite integral is then 0.

2.5.5 Calculating Reliability Using Microsoft Excel® Functions

In the past decades the Microsoft Excel® spreadsheet software became a widely utilized tool to perform a multitude of engineering and non-engineering tasks including statistical calculations. This book will illustrate how to perform some statistical analysis including reliability calculations utilizing Excel spreadsheet functions. Excel applications will cover both continuous and discrete statistical distributions.

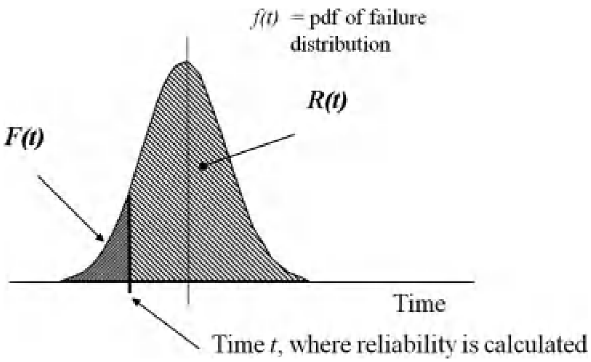

f(t) = pdf of failure distribution
R(t)
F(t)
Time
Time t, where reliability is calculated

Figure 2.7 *Probability Density Function (pdf) and its application to reliability.*

2.6 Continuous Distribution Functions

2.6.1 The Normal (or Gaussian) Distribution

By far the most widely used ‘model’ of the nature of variation is the mathematical function known as the *normal (or Gaussian) distribution*. The *normal* data distribution pattern occurs in many natural phenomena, such as human heights, weather patterns, and so on. However, there are limitations inherent in using this model in many engineering applications (see the comments in Section 2.8.1).

The normal pdf is given by

$$\mathrm{f}(x) = \frac{1}{\sigma(2\pi)^{1/2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{2.23}$$

where μ is the location parameter, equal to the mean. The mode and the median are coincident with the mean, as the pdf is symmetrical. σ is the scale parameter, equal to the SD.

A population which conforms to the normal distribution has variations which are symmetrically disposed about the mean (Figure 2.8) (i.e. the skewness is zero). Since the tails of the normal distribution are symmetrical, a given spread includes equal values in the left-hand and right-hand tails.

For normally distributed variables, about 68 % of the population fall between ± 1 SD. About 95 % fall between ± 2 SD, and about 99.7 % between ± 3 SD.

An important reason for the wide applicability of the normal distribution is the fact that, whenever several random variables are added together, the resulting sum tends to normal regardless of the distributions of the variables being added. This is known as the *central limit theorem*. It justifies the use of the normal distribution in many engineering applications, including quality control. The normal distribution is a close fit to most quality control and some reliability observations, such as the sizes of machined parts and the lives of items subject to wearout failures. Appendix 1 gives values for Φ*(z)*, the *standardized normal* cdf, that is, $\mu = 0$

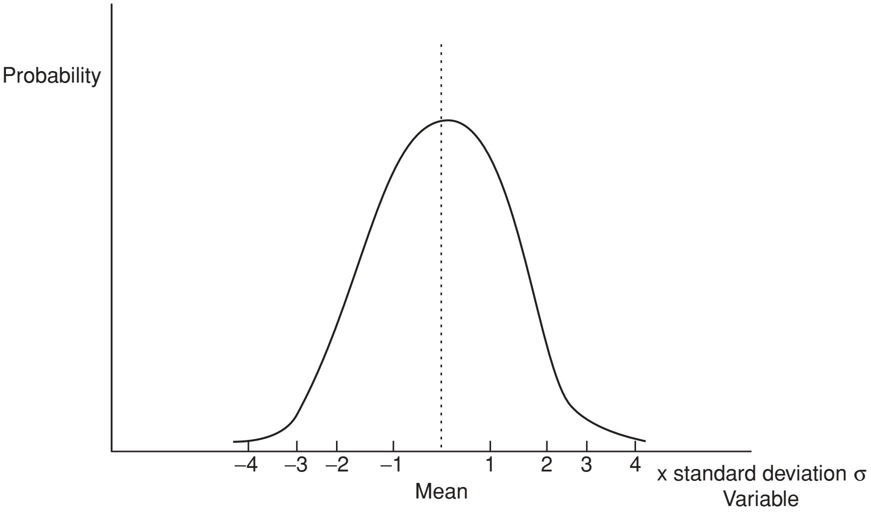


Figure 2.8 *The normal (Gaussian) distribution.*

and $\sigma = 1$. z represents the number of SDs displacement from the mean. Any normal distribution can be evaluated from the standardized normal distribution by calculating the standardized normal variate z, where

$$z = \frac{x - \mu}{\sigma}$$

and finding the appropriate value of Φ (z).

The pdf of a normal distribution with parameters μ and σ can be calculated using Excel as $f(x)$ = NORMDIST(x, μ, σ, FALSE) and reliability as $R(x)$ =1-NORMDIST(x, μ, σ, TRUE). The standardized normal cdf can be calculated as Φ (z) = NORMSDIST(z).

Example 2.5

The life of an incandescent lamp is normally distributed, with mean 1200 h and SD 200 h. What is the probability that a lamp will last (a) at least 800 h? (b) at least 1600 h?

a $z = (x – \mu)/\sigma$, that is, the distance of x from μ expressed as a number of SDs. Then

$$z = \frac{800 - 1200}{200} = -2\,\text{SD}$$

Appendix 1 shows that the probability of a value not exceeding 2 SD is 0.977. Figure 2.9(a) shows this graphically, on the pdf (the shaded area).

b The probability of a lamp surviving more than 1600 h is derived similarly:

$$z = \frac{1600 - 1200}{200} = 2\,\text{SD}$$

This represents the area under the pdf curve beyond the +2 SD point. (Figure 2.9(a)) or 1− (area under the curve to the left of +2 SD) on the cdf (Figure 2.9(b)). Therefore the probability of surviving beyond 1600 h is (1 - 0.977) = 0.023.

The answers can also be obtained using Excel functions as follows:

Solution for (a): R(800 hours) = 1-NORMDIST(800,1200,200,TRUE) = 0.9772

Solution for (b): R(1600) = 1-NORMDIST(1600,1200,200,TRUE) = 0.0228

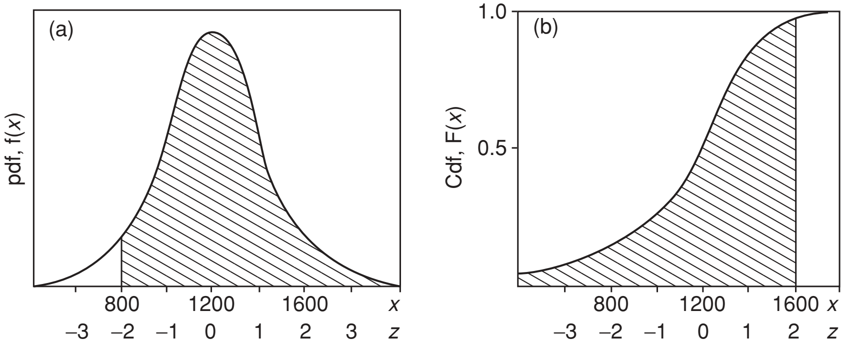


Figure 2.9 (a) The pdf f(x) versus x; (b) the cdf F(x) versus x (see Example 2.5).

2.6.2 The Lognormal Distribution

The lognormal distribution is based on the normal distribution. A random variable is lognormally distributed if the logarithm of the random variable is normally distributed. The lognormal distribution is a more versatile distribution than the normal as it has a range of shapes, and therefore is often a better fit to reliability data, such as for populations with wearout characteristics. Also, it does not have the normal distribution's disadvantage of extending below zero to $-\infty$. Outside reliability applications the lognormal is often used to model usage data, such as vehicle mileage per year, count of switch operations, repair time of a maintained system, and so on. The lognormal pdf is

$$\mathrm{f}(x) = \begin{cases} \dfrac{1}{\sigma x (2\pi)^{1/2}} \exp\left[-\dfrac{1}{2}\left(\dfrac{\ln x - \mu}{\sigma}\right)^2\right] & (\text{for } x \geq 0) \\ 0 & (\text{for } x < 0) \end{cases} \tag{2.24}$$

As mentioned before, it is the normal distribution with $\ln x$ as the variate. The mean and SD of the lognormal distribution are given by

$$\text{Mean} = \exp\left(\mu + \frac{\sigma^2}{2}\right)$$
$$\text{SD} = [\exp(2\mu + 2\sigma^2) - \exp(2\mu + \sigma^2)]^{1/2}$$

where μ and σ are the mean and SD of the ln data.

When $\mu \gg \sigma$, the lognormal distribution approximates to the normal distribution. The normal and lognormal distributions describe reliability situations in which the hazard rate increases from $x = 0$ to a maximum and then decreases.

The cdf and reliability of the system following lognormal distribution with parameters μ and σ can also be calculated using Excel functions.

For example, R(x) = 1-LOGNORMDIST(x, μ, σ).

2.6.3 The Exponential Distribution

The exponential distribution describes the situation wherein the hazard rate is constant. A Poisson process (Section 2.10.2) generates a constant hazard rate. The pdf is

$$\mathrm{f}(x) = \begin{cases} a \exp(-ax) & (\text{for } x \geq 0) \\ 0 & (\text{for } x < 0) \end{cases} \tag{2.25}$$

This is an important distribution in reliability work, as it has the same central limiting relationship to life statistics as the normal distribution has to non-life statistics. It describes the constant hazard rate situation. As the hazard rate is often a function of time, we will denote the independent variable by t instead of x. The constant hazard rate is denoted by λ. The mean life, or mean time to failure (MTTF), is $1/\lambda$. The pdf is then written as

$$\mathrm{f}(t) = \lambda \exp(-\lambda t) \tag{2.26}$$

The probability of no failures occurring before time t is obtained by integrating Eq. (2.26) between 0 and t and subtracting from 1:

$$\mathrm{R}(t) = 1 - \int_0^t \mathrm{f}(t)\mathrm{d}t = \exp(-\lambda t) \tag{2.27}$$

The Excel functions for the exponential distribution are: pdf f(t) = EXPONDIST(t, λ, FALSE) and reliability R(t) =1-EXPONDIST(t, λ, TRUE).

R(t) is the *reliability function* (or survival probability). For example, the reliability of an item with an MTTF of 500 h over a 24 h period is

$$\mathrm{R}(24) = \exp\left(\frac{-24}{500}\right) = 0.953 \text{ or } = \text{1-EXPONDIST}(24,\ 1/500,\ \text{TRUE})$$

For items which are repaired, λ is called the *failure rate*, and 1/λ is called the *mean time between failures* (MTBF) (also referred as θ). Please note from Eq. (2.27) that 63.2 % of items will have failed by t = MTBF.

2.6.4 The Gamma Distribution

In statistical terms the gamma distribution represents the sum of n exponentially distributed random variables. The gamma distribution is a flexible life distribution model that may offer a good fit to some sets of failure data. In reliability terms, it describes the situation when partial failures can exist, that is, when a given number of partial failure events must occur before an item fails, or the time to the ath failure when time to failure is exponentially distributed. The pdf is

$$\begin{aligned}
\mathrm{f}(x) &= \begin{cases} \dfrac{\lambda}{\Gamma(a)}(\lambda x)^{a-1}\exp(-\lambda x) & (\text{for } x \geq 0) \\ 0 & (\text{for } x < 0) \end{cases} \\
\mu &= \frac{a}{\lambda} \\
\sigma &= \frac{a^{1/2}}{\lambda}
\end{aligned} \tag{2.28}$$

where λ is the failure rate (complete failures) and a the number of partial failures per complete failure, or events to generate a failure. $\Gamma(a)$ is the *gamma function*:

$$\Gamma(a) = \int_0^\infty x^{a-1}\exp(-x)\mathrm{d}x \tag{2.29}$$

When (a – 1) is a positive integer, $\Gamma(a) = (a - 1)!$ This is the case in the partial failure situation. The exponential distribution is a special case of the gamma distribution, when $a = 1$, that is,

$$\mathrm{f}(x) = \lambda \exp(-\lambda x)$$

The gamma distribution can also be used to describe a decreasing or increasing hazard rate. When $a < 1$, h(x) will decrease whilst for $a > 1$, h(x) increases.

Utilizing Excel functions, pdf f(x) = GAMMADIST(x, a, λ,FALSE) and reliability

R(x) = 1-GAMMADIST(x, a, l, TRUE)

2.6.5 The $\chi 2$ Distribution

The χ^2 (chi-square) distribution is a special case of the gamma distribution, where $\lambda = \frac{1}{2}$, and $\nu = a/2$, where ν is called the number of *degrees of freedom* and must be a positive integer. This permits the use of the χ^2 distribution for evaluating reliability situations, since the number of failures, or events to failure, will always be positive integers. The χ^2 distribution is really a family of distributions, which range in shape from that of the exponential to that of the normal distribution. Each distribution is identified by the degrees of freedom.

In statistical theory, the χ^2 distribution is very important, as it is the distribution of the sums of squares of n, independent, normal variates. This allows it to be used for statistical testing, goodness-of-fit tests and evaluating confidence. These applications are covered later. The cdf for the χ^2 distribution is tabulated for a range of degrees of freedom in Appendix 2. The Excel function corresponding to Appendix 2 tables is = CHIINV(α, ν) with α being a risk factor.

2.6.6 The Weibull Distribution

The Weibull distribution is arguably the most popular statistical distribution used by reliability engineers. It has the great advantage in reliability work that by adjusting the distribution parameters it can be made to fit many life distributions. When Walloddi Weibull delivered his famous American paper in 1951, the first reaction to his statistical distribution was negative, varying from skepticism to rejection. However the US Air Force recognized the merit of Weibull's method and funded his research until 1975.

The Weibull pdf is (in terms of time t)

$$\mathrm{f}(t) = \begin{cases} \dfrac{\beta}{\eta^{\beta}} t^{\beta-1} \exp\left[-\left(\dfrac{t}{\eta}\right)^{\beta}\right] & (\text{for } t \geq 0) \\ 0 & (\text{for } t < 0) \end{cases} \tag{2.30}$$

The corresponding reliability function is

$$\mathrm{R}(t) = \exp\left[-\left(\frac{t}{\eta}\right)^{\beta}\right] \tag{2.31}$$

The hazard rate is

$$\frac{\beta}{\eta^{\beta}} t^{\beta-1}$$

Mean or MTTF: $\mu = \eta\Gamma\left(\dfrac{1}{\beta} + 1\right)$

Standard deviation: $\sigma = \eta\sqrt{\Gamma\left(\frac{2}{\beta}+1\right) - \Gamma\left(\frac{1}{\beta}+1\right)^{2}}$

β is the *shape parameter* and η is the *scale parameter*, or *characteristic life*—it is the life at which 63.2 % of the population will have failed (see Eq. (2.31) substituting $t = \eta$).

When $\beta = 1$, the exponential reliability function (constant hazard rate) results, with

$$\eta = \text{mean life } (1/\lambda).$$

When $\beta < 1$, we get a *decreasing* hazard rate reliability function.

When $\beta > 1$, we get an *increasing* hazard rate reliability function.

When $\beta = 3.5$, for example, the distribution approximates to the normal distribution. Thus the Weibull distribution can be used to model a wide range of life distributions characteristic of engineered products. The Excel function for pdf is f(t) = WEIBULL(*t, β, η,* FALSE) and reliability R(t) = 1-WEIBULL(*t, β, η,* TRUE).

So far we have dealt with the two-parameter Weibull distribution. If, however, failures do not start at $t = 0$, but only after a finite time γ, then the Weibull reliability function takes the form

$$\mathrm{R}(t) = \exp\left[-\left(\frac{t-\gamma)}{\eta}\right)^{\beta}\right] \tag{2.32}$$

that is, a three-parameter distribution. γ is called the *failure free time, location parameter* or *minimum life*. More on the Weibull distribution will be presented in Chapter 3.

2.6.7 The Extreme Value Distributions

In reliability work we are often concerned not with the distribution of variables which describe the bulk of the population but only with the extreme values which can lead to failure. For example, the mechanical properties of a semiconductor wire bond are such that under normal operating conditions good wire bonds will not fracture or overheat. However, extreme high values of electrical load or extreme low values of bond strength can result in failure. In other words, we are concerned with the implications of the tails of the distributions in load–strength interactions. However, we often cannot assume that, because a measured value appears to be, say, normally distributed, that this distribution necessarily is a good model for the extremes. Also, few measurements are likely to have been made at these extremes. Extreme value statistics are capable of describing these situations asymptotically.

Extreme value statistics are derived by considering the lowest or highest values in each of a series of equal samples. For example, consider the sample data in Table 2.1, taken randomly from a common population. The overall data can be plotted as shown in Figure 2.10 as f(x). However, if we plot separately the lowest values and the highest values in each sample, they will appear as $\mathrm{g}_{\mathrm{L}}(x)$ and $g_{\mathrm{H}}(x)$. $g_{\mathrm{L}}(x)$ is the extreme value distribution of the lowest extreme whilst $g_H(x)$ is the extreme value distribution of the highest extreme in each sample. For many distributions the distribution of the extremes will be one of three types:

Type I—also known as the extreme value or Gumbel distribution.
Type II—also known as the log extreme value distribution.
Type III—for the lowest extreme values. This is the Weibull distribution.

Table 2.1 *Sample data taken randomly from a common population.*

Sample	Data							
1	30	31	41	29	39	36	38	30
2	31	34	23	27	29	32	35	35
3	26	33	35	32	34	29	30	34
4	27	33	30	31	31	36	28	40
5	18	39	25	32	31	34	27	37
6	22	36	42	27	33	27	31	31
7	39	35	32	39	32	27	28	32
8	33	34	32	30	34	35	33	28
9	32	32	37	25	33	35	35	19
10	28	32	36	37	17	31	42	32
11	26	22	32	23	33	36	36	31
12	36	31	45	24	30	27	24	27

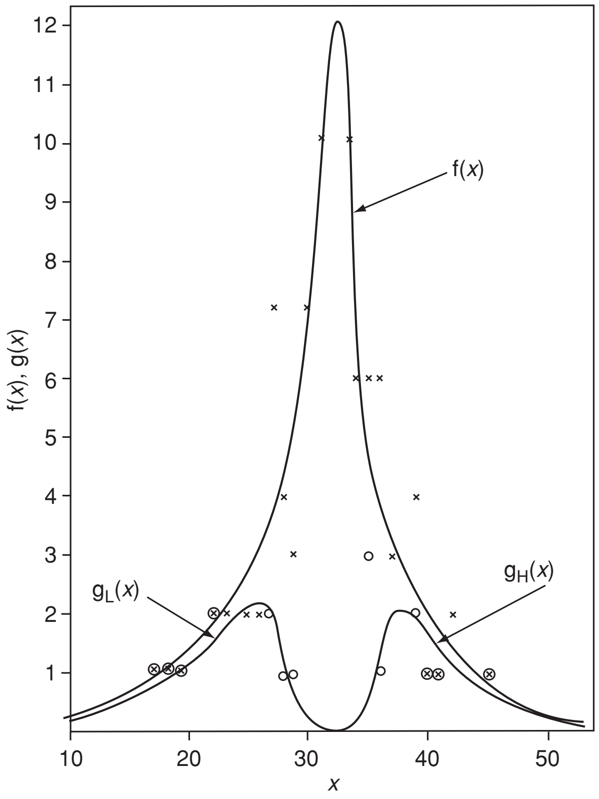


Figure 2.10 *Extreme value distributions.*

2.6.7.1 Extreme Value Type I

The type I extreme value distributions for maximum and minimum values are the limiting models for the right and left tails of the exponential types of distribution, where this is defined as any distribution whose cumulative probability approaches unity at a rate which is equal to or greater than that for the exponential distribution. This includes most reliability distributions, such as the normal, lognormal and exponential distributions.

The probability density functions for maximum and minimum values, respectively, are

$$\mathrm{f}(x) = \frac{1}{\sigma}\exp\left\{-\frac{1}{\sigma}(x-\mu) - \exp\left[-\frac{1}{\sigma}(x-\mu)\right]\right\} \tag{2.33}$$

$$\mathrm{f}(x) = \frac{1}{\sigma}\exp\left\{\frac{1}{\sigma}(x-\mu) - \exp\left[\frac{1}{\sigma}(x-\mu)\right]\right\} \tag{2.34}$$

The *reduced variate* is given by

$$y = \frac{x-\mu}{\sigma}$$

Substituting in Eqs. (2.33) and (2.34), we can derive the cdf in terms of the reduced variate y.

For maximum values:

$$\mathrm{F}(y) = \int_{-\infty}^{y} \exp\{-[x+\exp(-x)]\}\,\mathrm{d}x = \exp[-\exp(-y)] \tag{2.35}$$

For minimum values:

$$\mathrm{F}(y) = 1 - \exp[-\exp(y)] \tag{2.36}$$

The distribution of maximum values is right-skewed and the distribution of minimum values is left-skewed. The hazard function of maximum values approaches unity with increasing x, whilst that for minimum values increases exponentially.

For maximum values:

$$\mu_{\mathrm{ev}_{\max}} = \mu + 0.577\sigma$$

For minimum values:

$$\mu_{\mathrm{ev}_{\min}} = \mu - 0.577\sigma$$

The standard deviation μ_{ev} is 1.283σ in both cases.

2.6.7.2 Extreme Value Type II

The extreme type II distribution does not play an important role in reliability work. If the logarithms of the variables are extreme value distributed, then the variable is described by the extreme value type II distribution. Thus its relationship to the type I extreme value distribution is analogous to that of the lognormal to the normal distribution.

2.6.7.3 Extreme Value Type III

The type III extreme value distribution for minimum values is the limiting model for the left-hand tail for distributions which are bounded at the left. In fact, the Weibull distribution is the type III extreme value distribution for minimum values, and although it was initially derived empirically, its use for describing the strength distribution of materials has been justified using extreme value theory.

2.6.7.4 The Extreme Value Distributions Related to Load and Strength

The type I extreme value distribution for maximum values is often an appropriate model for the occurrence of load events, when these are not bounded to the right, that is, when there is no limiting value.

It is well known that engineering materials possess strengths well below their theoretical capacity, mainly due to the existence of imperfections which give rise to non-uniform stresses under load. In fact, the strength will be related to the effect of the imperfection which creates the greatest reduction in strength, and hence the extreme value distribution for minimum values suggests itself as an appropriate model for strength.

The strength, and hence the time to failure, of many types of product can be considered to be dependent upon imperfections whose extent is bounded, since only very small imperfections will escape detection by inspection or process control, justifying use of a type III (Weibull) model. On the other hand, a type I model might be more representative of the strength of an item which is mass-produced and not 100 % inspected, or in which defects can exist whose extent is not bounded, but which are not detected, for example, a long wire, whose strength will then be a function of length.

For a system consisting of many components in series, where the system hazard rate is decreasing from $t = 0$ (i.e. bounded) a type III (Weibull) distribution will be a good model for the system time to failure.

2.7 Summary of Continuous Statistical Distributions

Figure 2.11 is a summary of the continuous distributions described above.

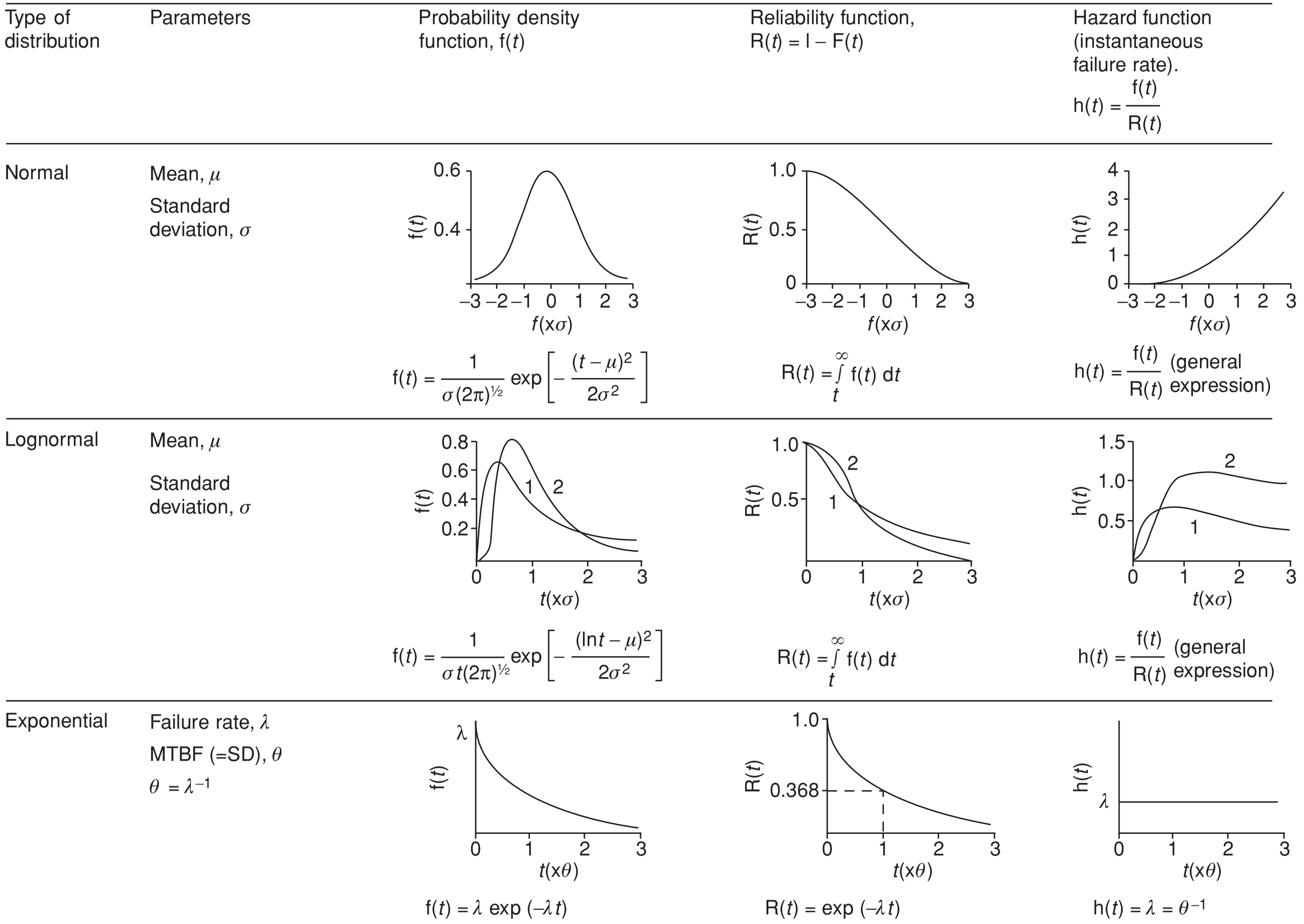
2.8 Variation in Engineering

Every practical engineering design must take account of the effects of the variation inherent in parameters, environments, and processes. Variation and its effects can be considered in three categories:

1 *Deterministic*, or *causal*, which is the case when the relationship between a parameter and its effect is known, and we can use theoretical or empirical formulae, for example, we can use Ohm's law to calculate the effect of resistance change on the performance of a voltage divider. No statistical methods are required. The effects of variation are calculated by inserting the expected range of values into the formulae.

2 *Functional*, which includes relationships such as the effect of a change of operating procedure, human mistakes, calibration errors, and so on. There are no theoretical formulae. In principle these can be allowed for, but often are not, and the cause and effect relationships are not always easy to identify or quantify.

3 *Random*. These are the effects of the inherent variability of processes and operating conditions. They can be considered to be the variations that are left unexplained when all deterministic and functional causes have been eliminated. For example, a machining process that is in control will nevertheless produce



Type of distribution	Parameters	Probability density function, f(t)	Reliability function, R(t) = I – F(t)	Hazard function (instantaneous failure rate). $h(t) = \frac{f(t)}{R(t)}$
Normal	Mean, μ Standard deviation, σ	$f(t) = \frac{1}{\sigma(2\pi)^{½}} \exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right]$	$R(t) = \int_t^\infty f(t)\,dt$	$h(t) = \frac{f(t)}{R(t)}$ (general expression)
Lognormal	Mean, μ Standard deviation, σ	$f(t) = \frac{1}{\sigma t(2\pi)^{½}} \exp\left[-\frac{(\ln t-\mu)^2}{2\sigma^2}\right]$	$R(t) = \int_t^\infty f(t)\,dt$	$h(t) = \frac{f(t)}{R(t)}$ (general expression)
Exponential	Failure rate, λ MTBF (=SD), θ $\theta = \lambda^{-1}$	$f(t) = \lambda \exp(-\lambda t)$	$R(t) = \exp(-\lambda t)$	$h(t) = \lambda = \theta^{-1}$

Figure 2.11 *Shapes of common failure distributions, reliability and hazard rate functions (shown in relation to t).*

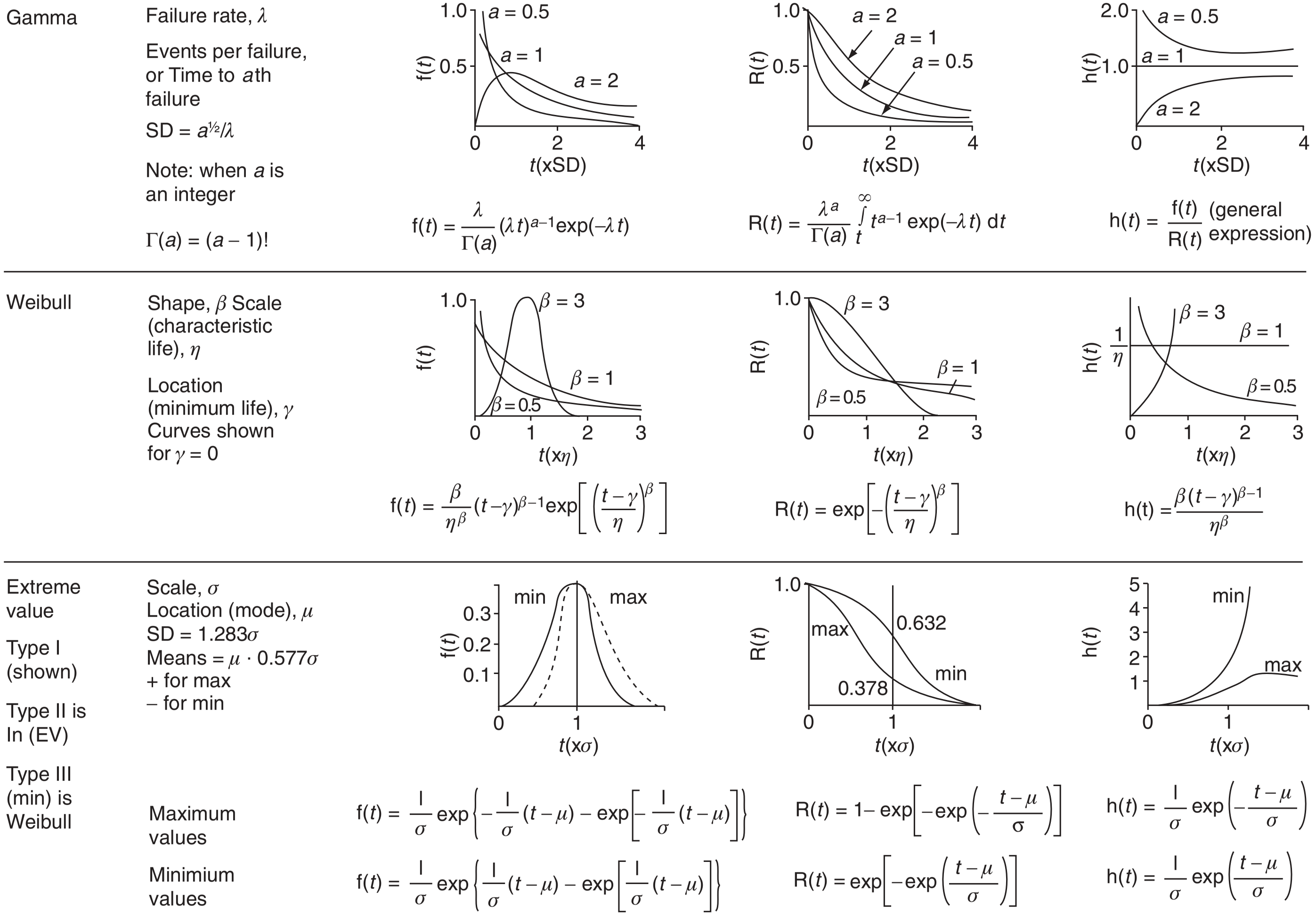


Figure 2.11 *(Continued).*

parts with some variation in dimensions, and random voltage fluctuations can occur on power supplies due to interference. Note that the random variations have causes. However, it is not always possible or practicable to predict how and when the cause will arise. The statistical models described above can be used to describe random variations, subject to the limitations discussed later.

Variation can also be *progressive*, for example due to wear, material fatigue, change of lubricating properties, or electrical parameter drift.

2.8.1 Is the Variation Normal?

The central limit theorem, and the convenient properties of the normal distribution, are the reasons why this particular function is taught as the basis of nearly all statistics of continuous variation. It is a common practice, in most applications, to assume that the variation being analysed is normal, then to determine the mean and SD of the normal distribution that best fits the data.

However, at this point we must stress an important limitation of assuming that the normal distribution describes the real variation of any process. The normal pdf has values between $+\infty$ and $-\infty$. Of course a machined component dimension cannot vary like this. The machine cannot add material to the component, so the dimension of the stock (which of course will vary, but not by much) will set an upper limit. The nature of the machining process, using gauges or other practical limiting features, will set a lower limit. Therefore the variation of the machined dimension would more realistically look something like Figure 2.12. Only the central part might be approximately normal, and the distribution will have been *curtailed*. In fact all variables, whether naturally occurring or resulting from engineering or other processes, are curtailed in some way, so the normal distribution, while being mathematically convenient, is actually misleading when used to make inferences well beyond the range of actual measurements, such as the probability of meeting an adult who is 1 m tall.

There are other ways in which variation in engineering might not be normal. These are:

- There might be other kinds of selection process. For example, when electronic components such as resistors, microprocessors, and so on are manufactured, they are all tested at the end of the production process. They are then ‘binned’ according to the measured values. Typically, resistors that fall within ±2 % of the nominal resistance value are classified as precision resistors, and are labelled, binned and sold

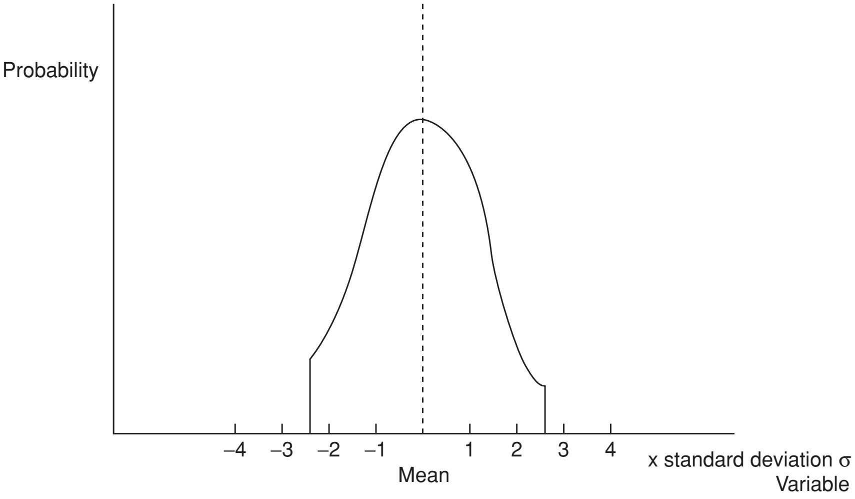


Figure 2.12 *Curtailed normal distribution.*

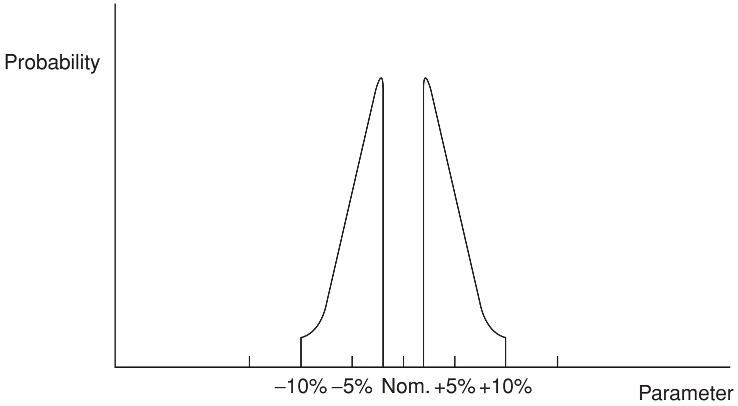


Figure 2.13 *Effect of selection.*

as such. Those that fall outside these limits, but within ±10 % become non-precision resistors, and are sold at a lower price. Those that fall outside ±10 % are scrapped. Because those sold as ±10 % will not include any that are ±2 %, the distribution of values is as shown in Figure 2.13. Similarly, microprocessors are sold with different operating speeds depending on the maximum speed at which they function correctly on test, having all been produced on the same process. The different maximum operating speeds are the result of the variations inherent in the process of manufacturing millions of transistors and capacitors and their interconnections, on each chip on each wafer. The technology sets the upper limit for the design and the process, and the selection criteria the lower limits. Of course, the process will also produce a proportion that will not meet other aspects of the specification, or that will not work at all.

– The variation might be unsymmetrical, or *skewed*, as shown in Figure 2.14. Distribution functions such as the lognormal and the Weibull can be used to model unsymmetrical variation. However, it is still important to remember that these mathematical models will still represent only approximations to the true variations, and the further into the tails that we apply them the greater will be the scope for uncertainty and error.

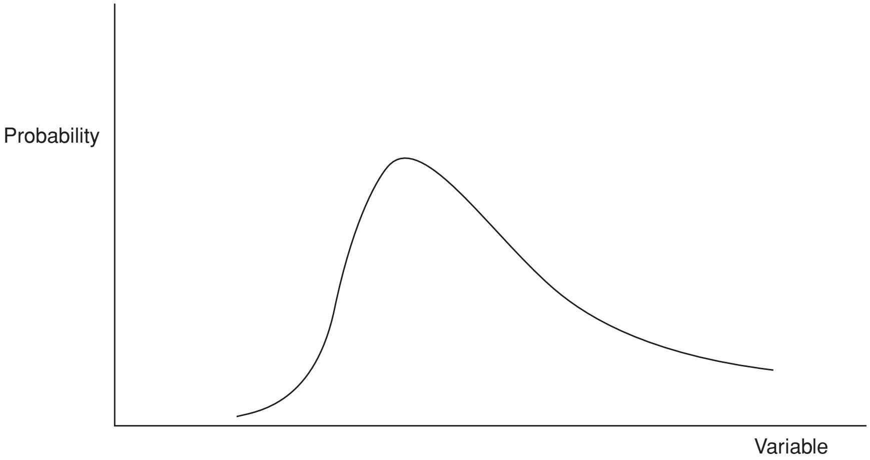


Figure 2.14 *Skewed distribution.*

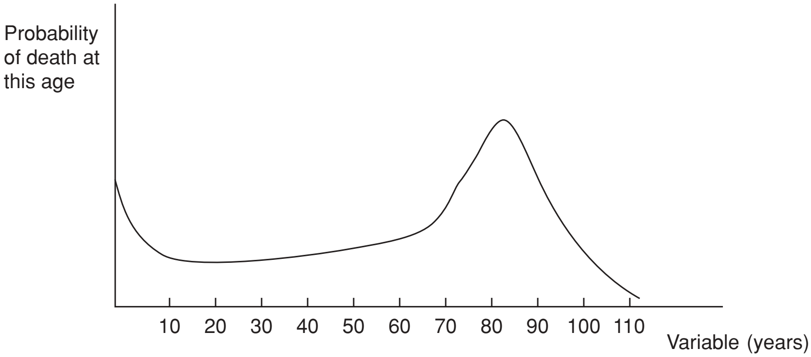


Figure 2.15 *Bi-modal distribution.*

– The variation might be multimodal (Figure 2.15), rather than unimodal as represented by distribution functions like the normal, lognormal and Weibull functions. For example, a process might be centred on one value, then an adjustment moves this nominal value. In a population of manufactured components this might result in a total variation that has two peaks, or a bi-modal distribution. A component might be subjected to a pattern of stress cycles that vary over a range in typical applications, and a further stress under particular conditions, for example resonance, lightning strike, and so on.

Variation of engineering parameters is, to a large extent, the result of human performance. Factors such as measurements, calibrations, accept/reject criteria, control of processes, and so on are subject to human capabilities, judgements, and errors. People do not behave normally.

Walter Shewhart, in 1931 was the first to explain the nature of variation in manufacturing processes. Figure 2.16 illustrates four very different kinds of variation, which, however all have the same means and

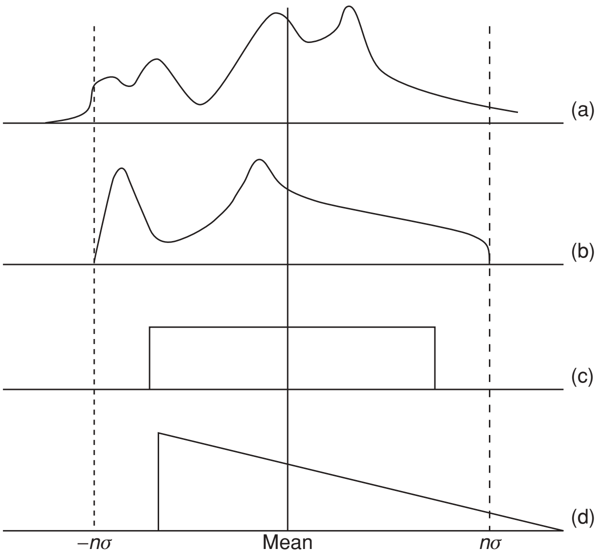


Figure 2.16 *Four distributions with the same means and SDs (after W. A. Shewhart).*

Figure 2.16 SDs. These show clearly how misleading it can be to assume that any manufacturing variation is normal and then to make assertions about the population based upon the assumption.

2.8.2 Effects and Causes

In engineering (and in many other applications) we are really much more concerned with the effects of variation than with the properties and parameters. If, for example, the output of a process varied as in Figure 2.16(c), and the '$\pm\sigma$' lines denoted the allowable tolerance, 100 % would be in tolerance. If, however, the process behaved as in (a) or (d), a proportion would be outside tolerance (only at the high end in the case of (d)). Variation can have other effects. A smaller diameter on a shaft might lead to higher oil loss or reduced fatigue life. Higher temperature operation might make an electronic circuit shut down. A higher proportion of fast microprocessors in a production batch would result in higher profit. We must therefore first identify the effects of variation (they are often starkly apparent), and determine whether and to what extent the effects can be reduced. This is not simply a matter of 'reducing SD'.

The effects of variation can be reduced in two ways:

1 We can compensate for the variation, by methods such as gauging or 'select on test' (this curtails the original variation), by providing temperature compensating devices, and so on.
2 We can reduce the variation.

In both cases we must understand the cause, or causes, of the variation. Shewhart categorized manufacturing variation into '*assignable*' and '*non-assignable*' causes. (These are also referred to as '*special causes*' and '*common causes*'.) Assignable variation is any whose cause can be practically and economically identified and reduced: deterministic and functional variation fall into this category. Non-assignable variation is that which remains when all of the assignable variation has been removed. A process in this state is '*in control*', and will have minimal, random, variation. Note that these are practical criteria, with no strict mathematical basis. Shewhart developed the methods of *statistical process control* (SPC) around this thinking, with the emphasis on using the data and charting methods to identify and reduce assignable variation, and to keep processes in control. SPC methods are described in detail in Chapter 13.

2.8.3 Tails

People such as life insurance actuaries, clothing manufacturers, and pure scientists are interested in averages and SDs: the behaviour of the bulk of the data. Since most of the sample data, in any situation, will represent this behaviour, they can make credible assertions about these population parameters. However, the further we try to extend the assertions into the tails, the less credible they become, particularly when the assertions are taken beyond any of the data. In engineering we are usually more concerned about the behaviour of variation at the extremes, than that near the average. We are concerned by high stresses, high and low temperatures, slow processors, weak components, and so on. In other words, it is the tails of the distributions that concern us. We often have only small samples to measure or test. Therefore, using conventional mathematical statistics to attempt to understand the nature, causes and effects of variation at the extremes can be misleading. However, these situations can be analysed using the extreme value distributions presented earlier in this chapter.

2.9 Conclusions

These are the aspects that matter in engineering, and they transcend the kind of basic statistical theory that is generally taught and applied. Later teachers, particularly W. E. Deming (see Chapter 1) and Genichi

Taguchi (Chapter 11) extended the ideas by demonstrating how reducing variation reduces costs and increases productivity, and by emphasizing the management implications.

Despite all of these reasons why conventional statistical methods can be misleading if used to describe and deal with variation in engineering, they are widely taught and used, and their limitations are hardly considered. Examples are:

- Most textbooks and teaching on SPC emphasise the use of the normal distribution as the basis for charting and decision-making. They emphasize the mathematical aspects, such as probabilities of producing parts outside arbitrary 2σ or 3σ limits, and pay little attention to the practical aspects discussed above.
- Typical design rules for mechanical components in critical stress application conditions, such as aircraft and civil engineering structural components, require that there must be a factor of safety (say 2) between the maximum expected stress and the lower σ value of the expected strength. This approach is really quite arbitrary, and oversimplifies the true nature of variations such as strength and loads, as described above. Why, for example, select 3σ? If the strength of the component were truly normally distributed, about 0.1 % of components would be weaker than the 3-σ value. If few components are made and used, the probability of one failing would be very low. However, if many are made and used, the probability of a failure among the larger population would increase proportionately. If the component is used in a very critical application, such as an aircraft engine suspension bolt, this probability might be considered too high to be tolerable. Of course there are often other factors that must be considered, such as weight, cost, and the consequences of failure. The criteria applied to design of a domestic machine might sensibly be less conservative than for a commercial aircraft application.
- The '*six sigma*' approach to achieving high quality is based on the idea that, if any process is controlled in such a way that only operations that exceed plus or minus 6σ of the underlying distribution will be unacceptable, then fewer than 3.4 per million operations will fail. The exact quantity is based on arbitrary and generally unrealistic assumptions about the distribution functions, as described above. (The 'six sigma' approach entails other features, such as the use of a wide range of statistical and other methods to identify and reduce variations of all kinds, and the training and deployment of specialists. It is described in Chapter 17.)

2.10 Discrete Variation

2.10.1 The Binomial Distribution

The binomial distribution describes a situation in which there are only two outcomes, such as pass or fail, and the probability remains the same for all trials. (Trials which give such results are called *Bernoulli trials*.) Therefore, it is obviously very useful in QA and reliability work. The pdf for the binomial distribution is

$$\mathrm{f}(x) = \frac{n!}{x!(n-x)!} p^x q^{(n-x)} \tag{2.37}$$

$$\frac{n!}{x!(n-x)!} \quad \text{may be written} \quad \binom{n}{x}$$

This is the probability of obtaining x good items and $(n - x)$ bad items, in a sample of n items, when the probability of selecting a good item is p and of selecting a bad item is q. The mean of the binomial distribution (from Eq. 2.13) is given by

$$\mu = np \tag{2.38}$$

and the SD from (Eq. 2.17)

$$\sigma = (npq)^{1/2} \tag{2.39}$$

The binomial distribution can only have values at points where x is an integer. The cdf of the binomial distribution (i.e. the probability of obtaining r or fewer successes in n trials) is given by

$$\mathrm{F}(r) = \sum_{x=0}^{r} \binom{n}{x} p^x q^{(n-x)} \tag{2.40}$$

Excel functions for the binomial distribution are: pdf $f(x)$ =BINOMDIST(x, n, p, FALSE) and cdf F(r) =BINOMDIST(r, n, p, TRUE).

Example 2.6

A frequent application of the cumulative binomial distribution is in quality control acceptance sampling. For example, if the acceptance criterion for a production line is that not more than 4 defectives may be found in a sample of 20, we can determine the probability of acceptance of a lot if the production process yields 10 % defectives.

From Eq. (2.40),

$$\begin{aligned}\mathrm{F}(4) &= \sum_{x=0}^{4} \binom{20}{x} 0.1^x 0.9^{(20-x)}\\ &= 0.957\end{aligned}$$

Utilising Excel spreadsheet F(4) = BINOMDIST(4, 20, 0.1, TRUE) = 0.9568.

Example 2.7

An aircraft landing gear has 4 tyres. Experience shows that tyre bursts occur on average on 1 landing in 1200. Assuming that tyre bursts occur independently of one another, and that a safe landing can be made if not more than 2 tyres burst, what is the probability of an unsafe landing?

If n is the number of tyres and p is the probability of a tyre bursting,

$$\begin{aligned}n &= 4\\ p &= \frac{1}{1200} = 0.000\,83\\ q &= (1-p) = 0.999\,17\end{aligned}$$

The probability of a safe landing is the probability that not more than 2 tyres burst.

$$\begin{aligned}F(2) &= \binom{4}{2}(0.000\,83)^2(0.999\,17)^2 + \binom{4}{1}(0.000\,83)^1(0.999\,17)^3 + \binom{4}{0}(0.000\,83)^0(0.999\,17)^4\\ &= 0.000\,004\,159\,7 + 0.003\,325\,006\,9 + 0.996\,670\,831\\ &= 0.999\,999\,997\,7\end{aligned}$$

Again, utilizing Excel function: F(2) = BINOMDIST(2, 4, 0.000 83, TRUE) = 0.999 999 997 714
Therefore the probability of an unsafe landing is

$$1 - 0.999\,999\,997\,7 = 2.3 \times 10^{-9}$$

2.10.2 The Poisson Distribution

If events are Poisson-distributed they occur at a *constant average rate*, with only one of the two outcomes countable, for example, the number of failures in a given time or defects in a length of wire:

$$\mathrm{f}(x) = \frac{\mu^x}{x!}\exp(-\mu) \quad (x = 0, 1, 2, \ldots) \tag{2.41}$$

where μ is the mean rate of occurrence. Excel functions for the Poisson distribution are: pdf f(x) = POISSON(x, MEAN, FALSE) and cdf = POISSON(x, MEAN, TRUE). In this case, MEAN = μ × Duration (in time, length, etc.)

For example, if we need to know the probability of not more than 3 failures occurring in 1000 h of operation of a system, when the mean rate of failures is 1 per 1000 h, ($\mu = 1/1000$, $x = 3$) we can calculate the MEAN = $\mu \times 1000$ h = 1.0.

Therefore P(x ≤ 3) = POISSON(3, 1, TRUE) = 0.981.

The Poisson distribution can also be considered as an extension of the binomial distribution, in which n is considered infinite or very large. Therefore it gives a good approximation to the binomial distribution, when p or q are small and n is large. This is useful in sampling work where the proportion of defectives is low (i.e. $p < 0.1$).

The Poisson approximation is

$$\begin{gathered}\mathrm{f}(x) = \frac{(np)^x}{x!}\exp(-np) \\ \left[\mu = np;\ \sigma = (np)^{1/2} = \mu^{1/2}\right]\end{gathered} \tag{2.42}$$

This approximation allows us to use Poisson tables or charts in appropriate cases and also simplifies calculations. However the applications of Poisson approximation became somewhat limited after computerized applications, such as Excel and various statistical programs became available.

It is also important to note, that if times to failure are exponentially distributed (see exponential distribution earlier this chapter), the probability of x failures is Poisson-distributed. For example, if the MTBF is 100 h, the probability of having more than 15 failures in 1000 h is derived as:

$$\text{Expected number of failures} = \frac{1000}{100} = 10$$

Probability of having 15 failures or less can be calculated using the Poisson Excel formula POISSON (15,10,TRUE) = 0.9513. Thus the probability of having more than 15 failures is 1 – 0.9513 = 0.0487.

Example 2.8

If the probability of an item failing is 0.001, what is the probability of 3 failing out of a population of 2000?

The binomial solution is

$$\binom{2000}{3} 0.999^{1997} 0.001^{3} = 0.1805$$

or utilizing Excel: BINOMDIST(3, 2000, 0.001, FALSE) = 0.180 53.

As an alternative, the Poisson approximation can be applied. The Poisson approximation is evaluated as follows:

$$\begin{aligned} \mu &= np \\ &= 2000 \times 0.001 = 2 \\ P(x = 3) &= \frac{2^3}{3!} \exp(-2) = 0.1804 \end{aligned}$$

As the normal distribution represents a limiting case of the binomial and Poisson distributions, it can be used to provide a good approximation to these distributions. For example, it can be used when 0.1 > p > 0.9 and n is large.

Then

$$\begin{aligned} \mu &= np \\ \sigma &= (npq)^{1/2} \end{aligned}$$

Example 2.9

What is the probability of having not more than 20 failures if $n = 100$, $p = 0.14$?

Using the binomial distribution,

$$P_{20} = 0.9640$$

Using the normal approximation

$$\begin{aligned} \mu &= np = 14 \\ \sigma &= (npq)^{1/2} = 3.470 \\ z &= \frac{20 - 14}{3.47} = 1.73 \end{aligned}$$

Referring to Appendix 1, $P_{20} = 0.9582$ or Excel: = NORMSDIST(1.73) = 0.958 18.

As $p \to 0.5$, the approximation improves, and we can then use it with smaller values of n. Typically, if $p = 0.4$, we can use the approximation with $n = 50$.

2.11 Statistical Confidence

Earlier in this chapter we mentioned the problem of statistical confidence. Confidence is the exact fraction of times the *confidence interval* will include the true value, if the experiment is repeated many times. The confidence interval is the interval between the *upper* and *lower confidence limits*. Confidence intervals are used in making an assertion about a population given data from a sample. Clearly, the larger the sample the greater will be our intuitive confidence that the estimate of the population parameter will be close to the true value. To illustrate this point, let's use the hypothetical example where we test 10 samples out of large population and 1 sample fails with 9 surviving. In this case we may infer a non-parametric reliability of 90 %. If we test 100 samples from the same population and experience 10 failures, we may again similarly infer 90 % reliability. However our confidence in that number will be much higher than that in the first case due to the larger sample in the second case.

Statistical confidence and engineering confidence must not be confused; statistical confidence takes no account of engineering or process knowledge or changes which might make sample data unrepresentative. Derived statistical confidence values must always be interpreted in the light of engineering knowledge, which might serve to increase or decrease our engineering confidence.

2.11.1 Confidence Limits on Continuous Variables

If the population value x follows a normal distribution, it can be shown that the means, $\bar{x}$, of samples drawn from it are also normally distributed, with variance σ^2/n (SD $= \sigma/n^{1/2}$). The SD of the sample means is also called the *standard error of the estimate*, and is denoted S_x.

If x is not normally distributed, provided that n is large (> 30), $\bar{x}$ will tend to a normal distribution. If the distribution of x is not excessively skewed (and is unimodal) the normal approximation for $\bar{x}$ at values of n as small as 6 or 7 may be acceptable.

These results are derived from the central limit theorem, mentioned in Section 2.6.1. They are of great value in deriving confidence limits on population parameters, based on sample data. In reliability work it is not usually necessary to derive exact confidence limits and therefore the approximate methods described are quite adequate.

Example 2.10

A sample of 100 values has a mean of 27.56, with a standard deviation of 1.10. Derive 95 % confidence limits for the population mean. (Assume that the sample means are normally distributed.)

In this case, the SD of the sample means, or standard error of the estimate, is

$$\frac{\sigma}{n^{1/2}} = \frac{1.1}{(100)^{1/2}} = 0.11$$

We can refer to the table of the normal cdf (Appendix 1) to obtain the 95 % single-sided confidence limits. The closest tabulated value of z is 1.65.

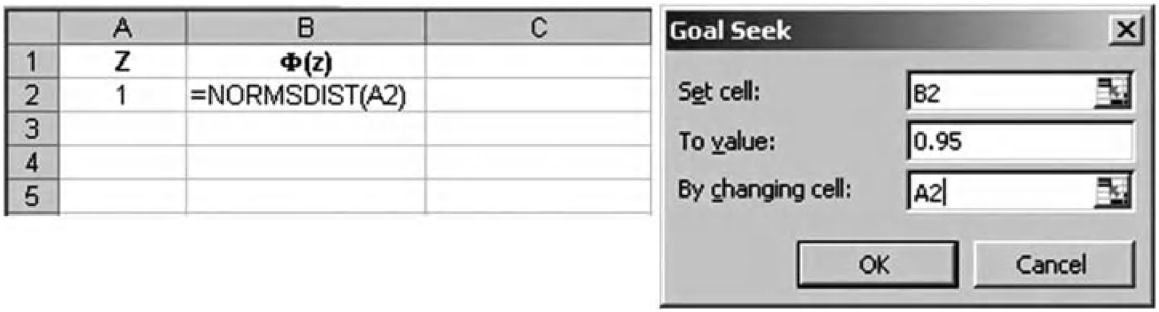
Alternatively we can run Excel's Tools – Goal Seek for Z in NORMSDIST(Z) = 0.95 (see Figure 2.17) to calculate the Z-value approaching 1.65.

Therefore, approximately ±1.65 SDs are enclosed within the 95 % single-sided confidence limits. Since the normal distribution is symmetrical, the 90 % *double-sided* confidence interval will exclude 5 % of values at either limit.

In the example, 1.65 SDs = 0.18. Therefore the 95 % confidence limits on the population mean are 27.56 ± 0.18, and the 90 % confidence interval is (27.56 − 0.18) to (27.56 + 0.18).

As a guide in confidence calculations, assuming a normal distribution see Figure 2.18:

± 1.65 SDs enclose approximately 90 % confidence limits (i.e. 5 % lie in each tail).
± 2.0 SDs enclose approximately 95 % confidence limits (i.e. 2.5 % lie in each tail).
± 2.5 SDs enclose approximately 99 % confidence limits (i.e. 0.5 % lie in each tail).



	A	B	C
1	Z	Φ(z)	
2	1	=NORMSDIST(A2)	
3			
4			
5			

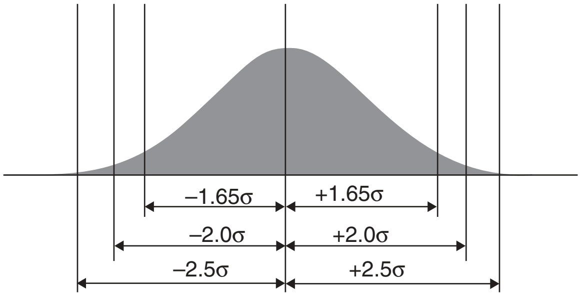
Goal Seek

Set cell: B2

To value: 0.95

By changing cell: A2

Figure 2.17 *Utilizing Excel's Goal Seek to find Z-value corresponding to the 95 % confidence interval.*



±1.65 SD Confidence (area under the curve): C ≅ 90%
±2.0 SD Confidence (area under the curve): C ≅ 95%
±2.5 SD Confidence (area under the curve): C ≅ 99%

Figure 2.18 *Confidence levels for normal distribution.*

2.12 Statistical Hypothesis Testing

It is often necessary to determine whether observed differences between the statistics of a sample and prior knowledge of a population, or between two sets of sample statistics, are statistically significant or due merely to chance. The variation inherent in sampling makes this distinction itself subject to chance. We need, therefore, to have methods for carrying out such tests. Statistical hypothesis testing is similar to confidence estimation, but instead of asking the question *How confident are we that the population parameter value is within the given limits?* (On the assumption that the sample and the population come from the same distribution), we ask *How significant is the deviation of the sample?*

In statistical hypothesis testing, we set up a *null hypothesis*, that is, that the two sets of information are derived from the same distribution. We then derive the *significance* to which this inference is tenable. As in confidence estimation, the significance we can attach to the inference will depend upon the size of the sample. Many significance test techniques have been developed for dealing with the many types of situation which can be encountered.

In this section we will cover a few of the simpler methods commonly used in reliability work. However, the reader should be aware that the methods described and the more advanced techniques are readily accessible on modern calculators and as computer programs. The texts listed in the Bibliography should be used to identify appropriate tests and tables for special cases.

2.12.1 Tests for Differences in Means (z Test)

A very common significance test is for the hypothesis that the mean of a set of data is the same as that of an assumed normal population, with known μ and σ. This is the *z test*. The *z-statistic* is given by

$$z = \frac{|\mu - \bar{x}|}{S_{\bar{x}}} = \frac{|\mu - \bar{x}|}{\sigma n^{-1/2}} \tag{2.43}$$

where n is the sample size, μ the population mean, $\bar{x}$ the sample mean and σ the population SD. We then derive the significance level from the normal cdf table.

Example 2.11

A type of roller bearing has a time to failure which is normally distributed, with a mean of 6000h and an SD of 450h. A sample of nine, using a changed lubricant, gave a mean life of 6400h. Has the new lubricant resulted in a statistically significant change in mean life?

$$z = \frac{|6000 - 6400|}{450 \times 9^{-1/2}} = 2.67$$

From Appendix 1, $z = 2.67$ indicates a cumulative probability of 0.996. This indicates that there is only 0.004 probability of observing this change purely by chance, that is, the change is significant at the 0.4 % level. Thus we reject the null hypothesis that the sample data are derived from the same normal distribution as the population, and infer that the new lubricant does provide an increased life.

Significance is denoted by α. In engineering, a significance level of less than 5 % can usually be considered to be sufficient evidence upon which to reject a null hypothesis. A significance of greater than 10 % would not normally constitute sufficient evidence, and we might either reject the null hypothesis or perform further trials to obtain more data. The significance level considered sufficient will depend upon the importance of the decision to be made based on the evidence. As with confidence, significance should also be assessed in the light of engineering knowledge.

Instead of testing a sample against a population, we may need to determine whether there is a statistically significant difference between the means of two samples. The SD of the distribution of the difference in the means of the samples is

$$S_{(\bar{x}_1 - \bar{x}_2)} = \frac{\sigma_1}{n_1^{1/2}} + \frac{\sigma_2}{n_2^{1/2}} \tag{2.44}$$

The SD of the distribution of the difference of the sampling means is called the *standard error of the difference*. This test assumes that the SDs are the population SDs. Then

$$z = \frac{\text{difference in sample means}}{\text{standard error of the difference}}$$

Example 2.12

In Example 2.11, if the mean value of 6000 and SD of 450 were in fact derived from a sample of 60, does the mean of 6400, with an SD of 380 from a sample of 9 represent a statistically significant difference?

The difference in the means is

$$6400 - 6000 = 400$$

The standard error of the difference is

$$\begin{aligned} S_\mathrm{d} &= \frac{\sigma_1}{n_1^{1/2}} + \frac{\sigma_2}{n_2^{1/2}} \\ &= \frac{450}{60^{1/2}} + \frac{380}{9^{1/2}} = 185 \\ z &= \frac{400}{185} = 2.16 \\ a &= 1 - \Phi(z) = 0.015(1.5 \text{ percent}) \end{aligned}$$

We can therefore say that the difference is highly significant, a similar result to that of Example 2.11.

2.12.2 Use of the z Test for Binomial Trials

We can also use the z test for testing the significance of binomial data. Since in such cases we are concerned with both extremes of the distribution, we use a two-sided test, that is, we use 2α instead of α.

Example 2.13

Two sets of tests give the results in Table 2.2. We need to know if the differences in test results are statistically significant.

The null hypothesis that the tests are without difference is examined by combining the test results:

$$P = \frac{\text{total failed}}{\text{total tested}} = \frac{30}{527} = 0.057$$

The standard error of the difference in proportions is

$$\begin{aligned} S_\mathrm{d} &= \left[pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]^{1/2} \\ &= \left[0.057 \times 0.943\left(\frac{1}{217} + \frac{1}{310}\right)\right]^{1/2} \\ &= 0.02 \end{aligned}$$

Table 2.2 Results for tests in Example 2.13.

Test	Number tested, *n*	Number failed
1	217	16
2	310	14

The proportion failed in test 1 is 16/217 = 0.074. The proportion failed in test 2 is 14/310 = 0.045. The difference in proportions is 0.074 – 0.045 = 0.029. Therefore $z = 0.029/0.020 = 1.45$, giving

$$\begin{aligned}\alpha &= 1 - \Phi(z) = 7.35 \text{ per cent}\\ 2\alpha &= 14.7 \text{ per cent}\end{aligned}$$

With such a result, we would be unable to reject the null hypothesis and would therefore infer that the difference between the tests is not very significant.

2.12.3 χ2 Test for Significance

The χ^2 test for the significance of differences is used when we can make no assumptions about the underlying distributions. The value of the χ^2 statistic is calculated by summing the terms

$$\frac{(x_i - E_i)^2}{E_i}$$

where x_i and E_i are the ith observed and expected values, respectively. This value is compared with the χ^2 value appropriate to the required significance level.

Example 2.14

Using the data of Example 2.13, the χ^2 test is set up as follows:

Test	Failure		Success		Totals
1	16	12.35	201	204.65	217
2	14	17.65	296	292.35	310
Totals	30		497		527

The first number in each column is the observed value and the second number is the expected value based upon the totals of the observations (e.g. expected failures in test 1 = 30/527 × 217 = 12.35).

$$\chi^2 = \frac{(16 - 12.35)^2}{12.35} + \frac{(201 - 204.65)^2}{204.65} + \frac{(14 - 17.65)^2}{17.65} + \frac{(296 - 292.35)^2}{292.35} = 1.94$$

The number of DF is one less than the number of different possibilities which could exist. In this case there is only one DF, since there are two possibilities – pass and fail. The value of χ^2 of 1.94 for 1 DF (from Appendix 2) occurs between 0.1 and 0.2 (alternatively, CHIDIST(1.94,1) = 0.1636). Therefore a cumulative probability is between 80 and 90 % (1- CHIDIST(1.94,1) = 0.8363). The difference between the observed data sets is therefore not significant. This inference is the same as that derived in Example 2.13.

2.12.4 Tests for Differences in Variances. Variance Ratio Test (F Test)

The significance tests for differences in means described above have been based on the assumption in the null hypothesis that the samples came from the same normal distribution, and therefore should have a

Table 2.3 Life test data on two items.

	Sample size, n	Sample standard deviation, σ	Sample variance, σ^2
Item 1	20	37	1369
Item 2	10	31	961

common mean. We can also perform significance tests on the differences of variances. The *variance ratio, F*, is defined as

$$F = \frac{\text{greater estimate of population variance}}{\text{lesser estimate of population variance}}$$

Values of the F distribution are well tabulated against the number of degrees of freedom in the two variance estimates (for a sample size n, DF $= n - 1$) and can easily be found on the Internet (see for example NIST, 2011).

The Excel® function for the values of F-distribution is FINV(P, DF1, DF2). Where P is a probability (significance level), DF1 is degrees of freedom for the first population (numerator) and DF2 for the second population (denominator). When the value of F-distribution and degrees of freedom are known, the probability can be calculated using the other Excel function FDIST(F, DF1, DF2). The use of the F test is illustrated by Example 2.15.

Example 2.15

Life test data on two items give the results in Table 2.3.

$$F = \frac{1369}{961} = 1.42$$

Entering the tables of F values at 19 DF for the greater variance estimate and 9 DF for the lesser variance estimate, we see that at the 5 % level our value for F is less than the tabulated value. Therefore the difference in the variances is not significant at the 5 % level. The Excel solution would involve the Goal Seek function (similar to the example in Figure 2.17) for the value P as FINF(P, 19, 9) = 1.42 and would produce P = 0.3, which is much higher than the required 5 % risk level.

2.13 Non-Parametric Inferential Methods

Methods have been developed for measuring and comparing statistical variables when no assumption is made as to the form of the underlying distributions. These are called non-parametric (or distribution-free) statistical methods. They are only slightly less powerful than parametric methods in terms of the accuracy of the inferences derived for assumed normal distributions. However, they are more powerful when the distributions are not normal. They are also simple to use. Therefore they can be very useful in reliability work provided that the data being analysed are independently and identically distributed (IID). The implications of data not independently and identically distributed are covered in Section 2.15 and in the next chapter.

Table 2.4 *Critical values of r for the sign test. Reproduced by permission of McGraw-Hill.*

	Significance level per cent		
n	10	5	1
8	1	0	0
10	1	1	0
12	2	2	1
14	3	2	1
16	4	3	2
18	5	4	3
20	5	5	3
25	7	7	5
30	10	9	7
35	12	11	9
40	14	13	11
45	16	15	13
50	18	17	15
55	20	19	17
60	23	21	19
75	29	28	25
100	41	39	36

2.13.1 Comparison of Median Values

2.13.1.1 The Sign Test

If a null hypothesis states that the median values of two samples are the same, then about half the values of each sample should lie on either side of the median. Therefore about half the values of $(x_i - \bar{x})$ should be positive and half negative. If the null hypothesis is true and *r* is the number of differences with one sign, then *r* has a binomial distribution with parameters *n* and p =1/2. We can therefore use the binomial distribution to determine critical values of *r* to test whether there is a statistically significant difference between the median values. Table 2.4 gives critical values for *r* for the sign test where *r* is the number of less frequent signs. If the value of *r* is equal to or less than the tabulated value the null hypothesis is rejected.

Example 2.16

Ten items are tested to failure, with lives

98, 125, 141, 72, 119, 88, 64, 187, 92, 114

Do these results indicate a statistically significant change from the previous median life of 125?

The sign test result is

$$-0+----+--$$

that is, $r = 2$, $n = 9$ (since one difference $= 0$, we discard this item).

Table 2.4 shows that r is greater than the critical value for $n = 9$ at the 10 % significance level, and therefore the difference in median values is not statistically significant at this level.

2.13.1.2 *The Weighted Sign Test*

We can use the sign test to determine the likely magnitude of differences between samples when differences in medians are significant. The amount by which the samples are believed to differ are added to (or subtracted from) the values of one of the samples, and the sign test is then performed as described above. The test then indicates whether the two samples differ significantly by the weighted value.

2.13.1.3 *Tests for Variance*

Non-parametric tests for analysis of variance are given in Chapter 11.

2.13.1.4 *Reliability Estimates*

Non-parametric methods for estimating reliability values are given in Chapter 13.

2.14 Goodness of Fit

In analysing statistical data we need to determine how well the data fit an assumed distribution. The goodness of fit can be tested statistically, to provide a level of significance that the null hypothesis (i.e. that the data do fit the assumed distribution) is rejected. Goodness-of-fit testing is an extension of significance testing in which the sample cdf is compared with the assumed true cdf.

A number of methods are available to test how closely a set of data fits an assumed distribution. As with significance testing, the power of these tests in rejecting incorrect hypotheses varies with the number and type of data available, and with the assumption being tested.

2.14.1 The χ^2 Goodness-of-Fit Test

A commonly used and versatile test is the χ^2 goodness-of-fit test, since it is equally applicable to any assumed distribution, provided that a reasonably large number of data points is available. For accuracy, it is desirable to have at least three data classes, or *cells*, with at least five data points in each cell.

The justification for the χ^2 goodness-of-fit test is the assumption that, if a sample is divided into n cells (i.e. we have ν degrees of freedom where $\nu = n - 1$), then the values within each cell would be normally distributed about the expected value, if the assumed distribution is correct, that is, if x_i and E_i are the observed and expected values for cell i:

$$\sum_{i}^{n} \frac{(x_i - E_i)^2}{E_i} = \chi^2 \quad \text{(with } n - 1 \text{ degrees of freedom)}$$

High values of χ^2 cast doubt on the null hypothesis. The null hypothesis is usually rejected when the value of χ^2 falls outside the 90th percentile. If χ^2 is below this value, there is insufficient information to reject the hypothesis that the data come from the supposed distribution. If we obtain a very low χ^2 (e.g. less than the

Table 2.5 *Data from an overstress life test of transistors.*

Cell (h)	Number in cell	Cell (h)	Number in cell
0–999	18	3000–3999	12
1000–1999	14	4000–4999	6
2000–2999	10		

10th percentile), it suggests that the data correspond more closely to the supposed distribution than natural sampling variability would allow (i.e. perhaps the data have been ‘doctored’ in some way).

The application can be described by use of an example.

Example 2.17

Failure data of transistors are given in Table 2.5. What is the likelihood that failures occur at a constant average rate of 12 failures/1000 hours?

$$\chi^2 = \frac{(18-12)^2}{12} + \frac{(14-12)^2}{12} + \frac{(10-12)^2}{12} + \frac{(12-12)^2}{12} + \frac{(6-12)^2}{12} = 6.67$$

Referring to Appendix 2 for values of χ^2 with χ^2 with $(n-1) = 4$ degrees of freedom, 6.67 lies between the 80th and 90th percentiles of the χ^2 distribution (risk factors between 0.1 and 0.2). CHIDIST (6.67, 4) = 0.1543. Therefore the null hypothesis that the data are derived from a constant hazard rate process cannot be rejected at the 90 % level (risk factor needs to be less than 0.1).

If an assumed distribution gave expected values of 20, 15, 12, 10, 9 (i.e. a decreasing hazard rate), then

$$\chi^2 = \frac{(18-20)^2}{20} + \frac{(14-15)^2}{15} + \frac{(10-12)^2}{12} + \frac{(12-10)^2}{10} + \frac{(8-9)^2}{9} = 1.11$$

$\chi^2 = 1.11$ lies close to the 10th percentile (CHIDIST(1.11, 4) = 0.8926). Therefore we cannot reject the null hypothesis of the decreasing hazard rate distribution at the 90 % level.

Note that the E_i values should always be at least 5. Cells should be amalgamated if necessary to achieve this, with the degrees of freedom reduced accordingly. Also, if we have estimated the parameters of the distribution we are fitting to, the degrees of freedom should be reduced by the number of parameters estimated.

2.14.2 The Kolmogorov–Smirnov Test

Another goodness-of-fit test commonly used in reliability work is the Kolmogorov–Smirnov (K–S) test. It is rather simpler to use than the χ^2 test and can give better results with small numbers of data points. It is also convenient to use in conjunction with probability plots (see Chapter 3), since it is based upon cumulative ranked data, that is, the sample cdf. The procedure is:

1 Tabulate the ranked failure data. Calculate the values of $|x_i - E_i|$ where x_i is the ith cumulative rank value and E_i the expected cumulative rank value for the assumed distribution.
2 Determine the highest single value.
3 Compare this value with the appropriate K–S value (Appendix 3).

Table 2.6 Failure data with ranked values of x_i.

Event	Time to failure (h)	x_i	E_i	$\lvert x_i - E_i \rvert$
1	12.2	0.056	0.035	0.021
2	13.1	0.136	0.115	0.021
3	14.0	0.217	0.29	0.073
4	14.1	0.298	0.32	0.022
5	14.6	0.379	0.44	0.061
6	14.7	0.459	0.46	0.001
7	14.7	0.54	0.46	0.08*
8	15.1	0.621	0.58	0.041
9	15.7	0.702	0.73	0.028
10	15.8	0.783	0.75	0.033
11	16.3	0.864	0.85	0.014
12	16.9	0.94	0.95	0.006

Example 2.18

Table 2.6 shows failure data with the ranked values of x_i. We wish to test the null hypothesis that the data do not fit a normal distribution with parameters which give the tabulated cumulative values of E_i. Therefore, in the E_i column we list the expected value of proportion failed at each failure time.

The largest value of $|x_i - E_i|$ is 0.08 (shown in Table 2.6 by*). The Kolmogorov–Smirnov table (Appendix 3) shows that, for $n = 12$, the critical value of $|x_i - E_i|$ is 0.338 at the 10 % significance level. Therefore the null hypothesis is not rejected at this level, and we can accept the data as coming from the hypothesized normal distribution.

Example 2.18 shows quite a large difference between the critical K–S value and the largest value of $|x_i - E_i|$. When the parameters of the assumed cdf are being estimated from the sample data, as in this example, the critical K–S values are too large and give lower significance levels than are appropriate in the circumstances. In order to correct for this, the critical values should be multiplied by the factors:

$$\begin{aligned} &0.70 \ \ (\beta > 3.0) \\ &0.70 \ \ (1.5 \leq \beta \leq 3.0) \\ &0.70 \ \ (\beta < 1.5) \end{aligned}$$

where β is the Weibull shape parameter. Therefore, in Example 2.18, since the Weibull β value appropriate to the normal distribution is > 3.0, the corrected K–S critical value is $0.338 \times 0.70 = 0.237$.

2.15 Series of Events (Point Processes)

Situations in which discrete events occur randomly in a continuum (e.g. time) cannot be truly represented by a single continuous distribution function. Failures occurring in repairable systems, aircraft accidents and vehicle traffic flow past a point are examples of series of discrete events. These situations are called *stochastic point processes*. They can be analysed using the statistics of *event series*.

The Poisson distribution function (Eq. 2.41) describes the situation in which events occur randomly and at a constant average rate. This situation is described by a *homogeneous Poisson process* (HPP). A HPP is a *stationary* point process, since the distribution of the number of events in an interval of fixed length does not vary, regardless of when (where) the interval is sampled.

The Poisson distribution function is (from (2.41))

$$\mathrm{f}(x) = \frac{(\lambda x)^n}{n!} \exp(-\lambda x) \quad (\text{for } n = 0, 1, 2, \ldots) \tag{2.45}$$

where λ is the mean rate of occurrence, so that λx is the expected number of events in (0, x).

In a *non-homogeneous* Poisson process (NHPP) the point process is non-stationary (rate of occurrence is a function of time), so that the distribution of the number of events in an interval of fixed length changes as x increases. Typically, the discrete events (e.g. failures) might occur at an increasing or decreasing rate.

Note that an essential condition of any homogeneous Poisson process is that the probabilities of events occurring in any period are independent of what has occurred in preceding periods. A HPP describes a sequence of independently and identically exponentially distributed (IIED) random variables. A NHPP describes a sequence of random variables which is neither independently nor identically distributed.

2.15.1 Trend Analysis (Time Series Analysis)

When analysing data from a stochastic point process it is important to determine whether the process has a trend, that is, to know whether a failure rate is increasing, decreasing or constant. We can test for trends by analysing the *arrival values* of the event series. The arrival values $x_1, x_2, \ldots, x_n$ are the values of the independent variables (e.g. time) from $x = 0$ at which each event occurs. The *interarrival values* $\mathrm{X}_1, \mathrm{X}_2, \ldots \mathrm{X}_n$ are the intervals between successive events $1, 2, \ldots, n$, from $x = 0$. Figure 2.19 shows the distinction between arrival and interarrival values.

If x_0 is the period of observation, then the test statistic for trend is

$$U = \frac{\Sigma x_i/n - x_0/2}{x_0\sqrt{1/(12n)}} \tag{2.46}$$

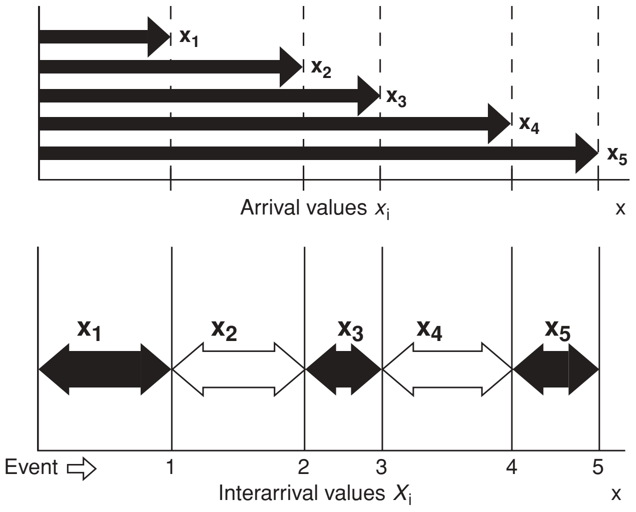


Figure 2.19 Arrival and interarrival values.

This is called the *centroid test* or the *Laplace test*. It compares the centroid of the observed arrival values with the mid-point of the period of observation. If $U = 0$ there is no trend, that is, the process is stationary. If $U < 0$ the trend is decreasing, that is, the interarrival values are tending to become larger. Conversely, when $U > 0$ the trend is increasing, that is, interarrival values are tending to become progressively smaller.

If the period of observation ends at an event, use $(n - 1)$ instead of n and exclude the time to the last event from the summation Σx_i.

We can test the null hypothesis that there is no trend in the chronologically ordered data by testing the value of U against the values of the standard normal variate, z. For example, using Appendix 1 or Excel function, if $U = 1.65$, for $z = 1.65$, $\Phi(z) = 0.95$. Therefore we can reject the null hypothesis at the 5 % significance level.

The centroid test is theoretically adequate if $n \geq 4$, when the observation interval ends with an event, and if $n \geq 3$, when the interval is terminated at a predetermined time.

The method is also called *time series analysis* (TSA).

Example 2.19

Arrival values (x_i) and interarrival values (X_i) between 12 successive failures of a component are as follows (observation ends at the last failure):

x_i	X_i	x_i	X_i
175	175	618	102
196	21	641	23
304	108	679	38
415	111	726	47
504	89	740	14
516	12	791	51

$$
\begin{aligned}
\Sigma x_i &= 5514 \text{ (excluding 791)}\\
n - 1 &= 11\\
\frac{\Sigma x_i}{n-1} &= 501.3\\
\frac{x_0}{2} &= 395.5\\
U &= \frac{501.3 - 395.5}{791\sqrt{1/(12 \times 11)}} = 1.54 \text{ Referring to Appendix 1 for z} = 1.54,\ \Phi(z) = 0.94
\end{aligned}
$$

Therefore we can reject the null hypothesis that there is no trend at the 6 % significance level. The interarrival times are becoming shorter, that is, the failure rate is increasing.

The existence of a trend in the data, as in Example 2.19, indicates that the interarrival values *are not independently and identically distributed* (IID). This is a very important point to consider in the analysis of failure data, as will be explained in Chapter 13.

2.15.2 Superimposed Processes

If a number of separate stochastic point process combine to form an overall process, for example, failure processes of individual components (or sockets) in a system, these are called *superimposed processes*. If the

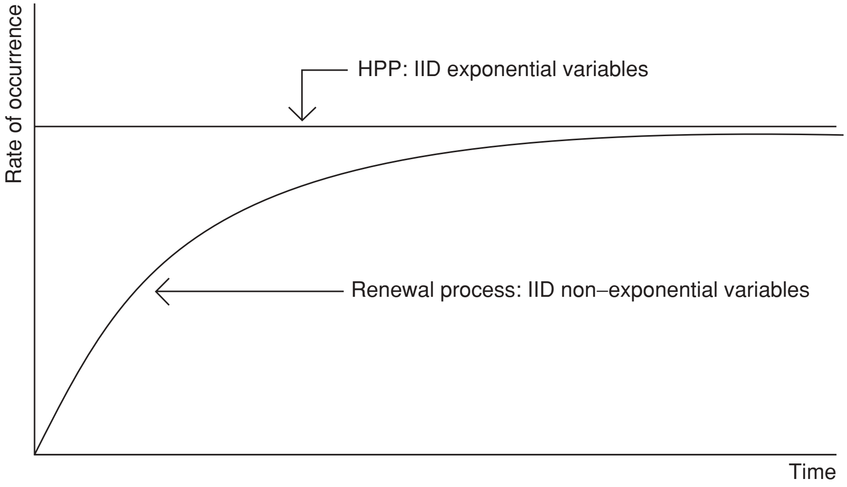


Figure 2.20 *Rate of occurrence for superimposed processes.*

individual random variables are IID exponential then the overall process variable is also IID exponential and the process is HPP.

If the individual variables are IID non-exponential, the overall process will tend to a HPP. Such a process is called a *renewal process or ordinary renewal process* (*ORD*). Figure 2.20 shows these processes. Renewal process analytics is often applied to describe the behaviour of a repairable system, where initial failures follow an exponential or any other statistical distribution and failed parts can be repaired to "as good as new" condition, returned to operation and experience secondary failures. None of the traditional statistical distribution covered in this chapter can be applied due to the fact that the failed units are not taken out of the total population, therefore the cdf can theoretically be greater than 1.0 in the cases where the total number of failures exceeds the size of the population. The renewal process can be described by the fundamental renewal equation:

$$\Lambda(t) = F(t) + \int_0^t \Lambda(t-\tau)dF(\tau)$$

Where $\Lambda(t)$ is the renewal function, which would represent the number of replacements (repairs) per unit and $F(t)$ is the cdf of the primary failures (as if there were no replacement of the failed parts).

2.16 Computer Software for Statistics

Computer software is available which can be used to carry out the analytical techniques described in this chapter and in later chapters which describe particular applications. As mentioned before, Microsoft Excel has a wide variety of statistical functions, covering most of the equations presented in this chapter. Among the software packages specializing in statistical analysis Minitab® statistical software and SAS® are probably most widely used around the world. Minitab is a comprehensive statistical package covering various aspects of data analysis, quality, design of experiments, and other engineering and non-engineering applications. SAS has more emphasis on business applications.

2.17 Practical Conclusions

Whilst the mathematical methods described in this chapter can be useful for providing insights and for forecasting, it is important that their limitations are appreciated. They are mathematical models, and they do not necessarily reflect reality in the way that deterministic, physics-based formulae do. The important points that must be borne in mind when applying statistical methods to engineering are:

- Real variation is seldom normal.
- The most important variation, as far as reliability is concerned, is usually that in the tails, where there is inevitably less (or no) data, the data are more uncertain, and where conventional statistical models can be most misleading.
- Variation can change over time, so that the patterns measured at one time might not represent the true situation at another. We will cover this aspect in more detail in later chapters.
- There might be interaction effects between variables, causing combined effects that are more significant than those of individual variations. This aspect will be covered in the later chapters.
- Variation in engineering is usually made or influenced by people. People do not behave in accordance with any credible mathematical models.
- Most engineering education in statistics covers only the mathematics, and few statisticians understand the practical aspects of the engineering problems they help to solve. This leads to inappropriate analyses and conclusions, and to a distrust of statistical methods among engineers.

We must strike an appropriate balance between using deterministic and statistical methods. For example, if we conduct a test in which an item is released from a height, and if it drops the test is a success, with results as follows:

items tested (0 failures): 0 1 10 2 0

then we could infer that the 80 % confidence that the reliability is at least 0.9 would be:

0 0.90 0.98 0.99

This assumes that the data are entirely statistical, that is, we have no prior knowledge of the physics or engineering. On the other hand, if we are confident that we have such knowledge (in this case, that the force of gravity will always act on the released items), then we will have 100 % confidence in 100 % reliability, even without performing any tests. In such deterministic cases statistical tests and interpretations are inappropriate.

However, many engineering situations can range from deterministic to statistical. For example, there might be cases when the release mechanism fails to open properly. The causes and effects of variations are often uncertain (particularly when interactions exist), so we must make the best use of our knowledge and use the best methods to explore the uncertainties.

Statistical tests can, by themselves, generate misleading results. We have discussed this in the context of extrapolations beyond the range of measured data. Another example might be a series of tests that indicate that an item operated at high stress is more reliable than when operated at low stress. The results might be due to the items on the high-stress tests being manufactured using a better process, or the high stress may actually improve the reliability (e.g. a higher temperature might improve the performance of a seal), or the results might be due to chance and unrepresentative of future tests. The cause of the observed reliability difference must be ascertained and understood in engineering terms. Sometimes this can be difficult.

Ultimately, all of our understanding should be based upon real knowledge (scientific, human, etc.). The statistical methods can provide clues to help us to gain such knowledge. The quotation from Kendall and Stuart on the inside of the front cover should be the motto for all statistics applications.

Questions

1. In the test firing of a missile, there are some events that are known to cause the missile to fail to reach its target. These events are listed below; together with their approximate probabilities of occurrence during a flight:

Event	Probability
(A_1) Cloud reflection	0.0001
(A_2) Precipitation	0.005
(A_3) Target evasion	0.002
(A_4) Electronic countermeasures	0.04

The probabilities of failure if these events occur are:

$$P(\mathrm{F}/\mathrm{A}_1) = 0.3;\ P(\mathrm{F}/\mathrm{A}_2) = 0.01;\ P(\mathrm{F}/\mathrm{A}_3) = 0.005;\ P(\mathrm{F}/\mathrm{A}_4) = 0.0002.$$

Use Bayes' theorem (Eq. 2.10) to calculate the probability of each of these events being the cause in the event of a missile failing to reach its target.

2. For a device with a failure probability of 0.02 when subjected to a specific test environment, use the binomial distribution to calculate the probabilities that a test sample of 25 devices will contain (a) no failures; (b) one failure; (c) more than one failure.
3. Repeat question 2 for a failure probability of 0.2.
4. Repeat questions 2 and 3 using the Poisson approximation to the binomial, and comment on the answers.
5. One of your suppliers has belatedly realized that about 10 % of the batches of a particular component recently supplied to you have a manufacturing fault that has reduced their reliability. There is no external or visual means of identifying these substandard components. Batch identity has, however, been maintained, so your problem is to sort batches that have this fault ('bad' batches) from the rest ('good' batches). An accelerated test has been devised such that components from good batches have a failure probability of 0.02 whereas those from bad batches have a failure probability of 0.2. A sampling plan has been devised as follows:

   1 Take a random sample of 25 items from each unknown batch, and subject them to the test.
   2 If there are 0 or 1 failed components, decide that the batch is a good one.
   3 If there are two or more failures, decide that the batch is a bad one.

   There are risks in this procedure. In particular, there are (i) the risk of deciding that a good batch is bad; and (ii) the risk of deciding that a bad batch is good. Use Bayes' theorem and your answers to questions 2 and 3 to evaluate these risks.
6. a Explain the circumstances in which you would expect observed failure times to conform to an exponential distribution.

   b Explain the relationship between the exponential and Poisson distributions in a reliability context.

c For equipment with an MTBF of 350 h calculate the probability of surviving a 200 h mission without failure.

7. A railway train is fitted with three engine/transmission units that can be assumed to exhibit a constant hazard with a mean life of 200 h. In a 15 h working day, calculate the probability of a train having: (a) no failed engine/transmission units, (b) not more than one failed unit, (c) not more than two failed units.

8. Assuming the exponential failure distribution, calculate the probability of a system surviving an operation time equal to twice the duration of the MTBF.

9. a Explain, using sketches where necessary, the meanings of the following terms used in describing the reliability behaviour of components; and show clearly how they are related to each other: (i) lifetime probability density function; (ii) cumulative distribution function; (iii) reliability function; (iv) hazard function.

   b Write down the expression for the cumulative distribution function (cdf) of the two-parameter Weibull distribution. Define its parameters and produce sketches to show how changing their values influences the cdf and the hazard function.

10. Ten components were tested to failure. The ordered ages at failure (hours) were: 70.9; 87.2; 101.7; 104.2; 106.2; 111.4; 112.6; 116.7; 143.0; 150.9.

    a On the assumption that these times to failure are normally distributed, estimate the component reliability and the hazard function (i) at age 100 h; and (ii) at age 150 h.

    b Use a Kolmogorov–Smirnov test to see whether it is reasonable to assume normality.

11. A flywheel is retained on a shaft by five bolts, which are each tightened to a specified torque of 50 ± 5 Nm. A sample of 20 assemblies was checked for bolt torque. The results from the 100 bolts had a mean of 47.2 Nm and a standard deviation of 1.38 Nm.

    a Assuming that torques are normally distributed, estimate the proportion below 45 Nm.

    b For a given assembly, what is the probability of (i) there being no bolts below 45 Nm; (ii) there being at least one bolt below 45 Nm; (iii) there being fewer than two bolts above 45 Nm; (iv) all five bolts being below 45 Nm.

    c In the overall sample of 100 bolts, four were actually found with torques below 45 Nm. (i) Comment on the comparison between this result and your answer to (a) above. (ii) Use this result to obtain a 90 % two-sided confidence interval for the proportion below 45 Nm.

    d Explain the meaning of the confidence interval in c (ii) above as you would to an intelligent, but non-technically-minded, manager.

    e The lowest torque bolt in each assembly was identified. For these 20 bolts, the mean torque was 45.5 Nm and the standard deviation 0.88 Nm. Assuming an appropriate extreme-value distribution, calculate the probability that *on a given assembly* the lowest torque will be (i) below 45 Nm; (ii) below 44 Nm.

12. The following data are the times (hours) between successive failures in a machining centre: 96; 81; 105; 34; 92; 81; 89; 138; 75; 156; 205; 111; 177.

    Calculate the trend statistic (Eq. 2.46) and test its significance.

13. Describe four ways in which the variation of an engineering parameter might be different from that based upon assuming that the normal distribution function is the correct model. Give an example in each case.

14. In most statistical applications the results that matter most relate to the behaviour of the majority of the population being studied. Why is this not the case in most engineering applications?

15. Ten items are put on test, until all have failed. The first failure occurs at 35 000 operating cycles. Regression analysis of the times to failure shows a good fit to a two-parameter Weibull distribution, and the distribution parameters are derived. The specification states that the item should have a B_{10}

life of 30 000 cycles. Discuss the practical implications assuming that the test shows compliance with the specification.

16. In question 15, how might your judgment be influenced if the item is:
    a A steel bolt subject to fatigue loading?
    b A plastic component in a child's toy?
    c A lighting unit?
    d A bearing in a gearbox?
    e A light-emitting diode?
    (You might like to try this question after studying Chapters 8 and 9.)
17. Explain what is meant by 'statistical confidence'. How might statistical confidence derived from an experiment be modified in the light of engineering knowledge?
18. In Example 2.7 describe three factors that could invalidate the assumption that tyre failures occur independently of one another. In what ways might the assumption be more valid for car tyres?
19. What is the probability of having not more than 20 but not less than 10 failures if $n = 100$, $p = 0.14$ (Section 2.10)?
    a Using the binomial distribution
    b Using the normal approximation
20. The life of an electronic controller is distributed lognormally, with the parameters $\mu = 20$ and $\sigma = 10$. What is the probability that a controller will last (a) at least 50 h? (b) at least 200 h?
21. A commercial washing machine has a non-repairable motor with a constant failure rate of 0.08 failure per year. The service organization has purchased two spare motors. If the design life of the washing machine is 7 years, what is the probability that two spares will be adequate. Hint: assume Poisson distribution.
22. Compare reliability values for the two products, Product A with exponentially distributed life and product B with Weibull distributed life. The parameters are $\text{MTBF}_A = \eta_B = 1000$ h. Compare the reliabilities at 300 hours for:
    $\beta_B = 0.5$
    $\beta_B = 1.0$
    $\beta_B = 3.0$
    How would you describe the effect of the Weibull shape parameter β on the reliability if the scale parameter remains the same?
23. Show the derivations of the hazard rate and cumulative hazard function for the Weibull distribution
24. Calculate the cumulative hazard rate for the Weibull distribution with the parameters $\beta = 2.5$ $\eta = 200$ h at time t = 100 h.

Bibliography

Helpful introductory sources

Chatfield, C. (1983) *Statistics for Technology*, 2nd edn, Chapman & Hall.

Conover, W. (1998) *Practical Non-parametric Statistics*, J. Wiley.

Hines, W. and Montgomery, D. (1990) *Probability and Statistics in Engineering and Management Science*, 3rd edn, Wiley.

Langley, R. (1979) *Practical Statistics Simply Explained*, 2nd edn, Pan.

Modarres, M. Kaminskiy, M. Krivtsov V. (1999) *Reliability Engineering and Risk Analysis*, 2nd edn, Marcel Dekker.

Montgomery, D., Runger, G., Hubele, M. (2006) *Engineering Statistics*, 4th edn, J. Wiley.

NIST (2011) Section 1.3.6.7.3 *Upper Critical Values of the F Distribution*. Engineering Statistics Handbook, published by NIST. Available at: http://www.itl.nist.gov/div898/handbook/eda/section3/eda3673.htm.

Ryan, T. (2000) *Statistical Methods for Quality Improvement*, Wiley.

More advanced works

Abernethy, R. (2003) *The New Weibull Handbook*, 5th edn, Dr. Robert Abernethy.

Duncan, A. (1986) *Quality Control and Industrial Statistics*, 5th edn, Irwin.

Hahn G., Meeker, W. (1991) *Statistical Intervals, a Guide for Practitioners*, J. Wiley.

Hollander, M., Wolfe, D. (1999) *Non-parametric Statistical Methods*, 2nd edn, J. Wiley.

Kleyner, A., Bhagath, S., Gasparini, M. *et al.* (1997) Bayesian techniques to reduce the sample size in automotive electronics attribute testing. *Microelectronics and Reliability*, **37**(6), 879–883.

Martz, H. and Waller, A. (1982) *Bayesian Reliability Analysis*, J. Wiley.

Modarres, M., Kaminskiy, M. and Krivtsov, V. (1999) Reliability Engineering and Risk MINITAB (2011) (General purpose statistics training and application software). Minitab Inc. 3081 Enterprise Drive, PA 16801, USA. (*The MINITAB Handbook* (Duxbury Press, Boston) is an excellent introduction to basic statistics), www.mintab.com.

National Institute of Standards and Technology (NIST). NIST/SEMATECH e-Handbook of Statistical Methods http://www.itl.nist.gov/div898/handbook/October 2010.

Nelson, W. (2003) *Applied Life Data Analysis*, J. Wiley.

SAS Business Analytics Software Manual. http://www.sas.com/resources/.

Seitano, S. (2001) *Engineering Uncertainty and Risk Analysis*, Hydroscience Inc.

Wassreman, G. (2003) *Reliability Verification Testing, and Analysis in Engineering Design*, Marcel Dekker.

3
Life Data Analysis and Probability Plotting

3.1 Introduction

It is frequently useful in reliability engineering to determine which distribution best fits a set of data and to derive estimates of the distribution parameters. The mathematical basis for the approaches to these problems was covered in Chapter 2.

3.1.1 General Approach to Life Data Analysis and Probability Plotting

The methods described in this chapter can be used to analyse any appropriate data, such as dimensional or parameter measurements. However, their use for analysing reliability time-to-failure (life data) will be emphasized.

General-purpose statistical software such as Minitab® includes capabilities for probability plotting. Since the Weibull distribution is the most commonly applied in reliability life data analysis, computer software packages have been developed specifically for this purpose, such as ReliaSoft Weibull++®, SuperSMITH Weibull® and few others. This chapter makes use of ReliaSoft Weibull++® software to illustrate how to perform these tasks.

Note that probability plotting methods to derive time-to-failure distribution parameters are only applicable when the data are independently and identically distributed (IID). This is usually the case for non-repairable components and systems but may *not* be the case with failure data from repairable systems. The reason is that repaired systems can have secondary failures, which are dependent on the primary failures. Also due to successive repairs the population of repairable systems can experience more failures than the size of that population causing cdf > 1.0, which is mathematically impossible. Reliability modelling and data analysis for repairable systems will be covered later in Chapters 6 and 13.

3.1.2 Statistical Data Analysis Methods

The process of finding the best statistical distribution based on the observed failure data, can be graphically illustrated by Figure 3.1. Based on the available data comprising the shaded segment of the pdf the rest of $f(t)$ can be 'reconstructed'. The goal of this process is to find the best fitting statistical distribution and to derive estimates of that distribution's parameters and consequently the reliability function $R(t)$. However in

Practical Reliability Engineering, Fifth Edition. Patrick D. T. O'Connor and Andre Kleyner.

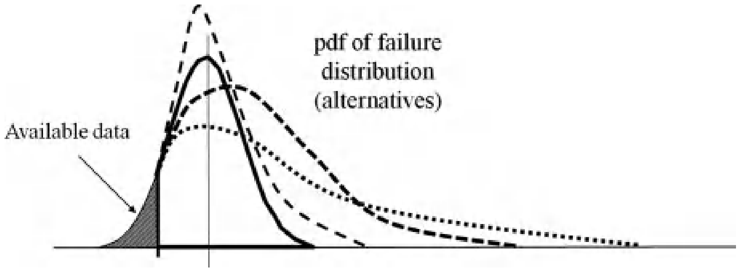


Figure 3.1 Probability plotting alternatives in regards to the possible pdf of failure distribution.

practical terms, this procedure is typically done based on constructing the cdf curve which has the best fit to the existing data.

The least mathematically intensive method for parameter estimation is the method of probability plotting. As the term implies, probability plotting in general involves a physical plot of the data on specially constructed probability plotting paper (different for each statistical distribution). The axes of probability plotting papers are transformed in such a way that the true cdf plots as a straight line. Therefore if the plotted data can be fitted by a straight line, the data fit the appropriate distribution (see Figure 3.2 fitting the normal probability plot). Further constructions permit the distribution parameters to be estimated. This method is easily implemented by hand, given that one can obtain the appropriate probability plotting paper. Probability plotting papers exist for all the major distribution including normal, lognormal, Weibull, exponential, extreme value, and so on and can be downloaded from the internet (see, e.g. ReliaSoft, 2011) However most of probability plotting these days is done with the use of computer software, which is covered later in this chapter.

The Weibull distribution (see Chapter 2) is a popular distribution for analysing life data, so the process is often referred to as *Weibull analysis*. The Weibull model can be applied based on 2-parameter, 3-parameter or mixed distributions. Other commonly used life distributions include the exponential, extreme value, lognormal and normal distributions. The analyst chooses the life distribution, that is most appropriate to model each particular data set based on goodness-of-fit tests, past experience and engineering judgement. The life data analysis process would require the following steps:

1 Gather life data for the product.
2 Select a lifetime distribution against which to test the data.
3 Generate plots and results that estimate the life characteristics of the product, such as the reliability, failure rate, mean life, or any other appropriate metrics.

This chapter will discuss theoretical and practical aspects of performing probability plotting and life data analysis.

3.2 Life Data Classification

In reliability work, life data can be time, distance travelled, on/off switches, cycles, and so on to failure. The accuracy and credibility of any parameter estimations are highly dependent on the quality, accuracy and completeness of the supplied data. Good data, along with the appropriate model choice, usually results in good parameter estimations.

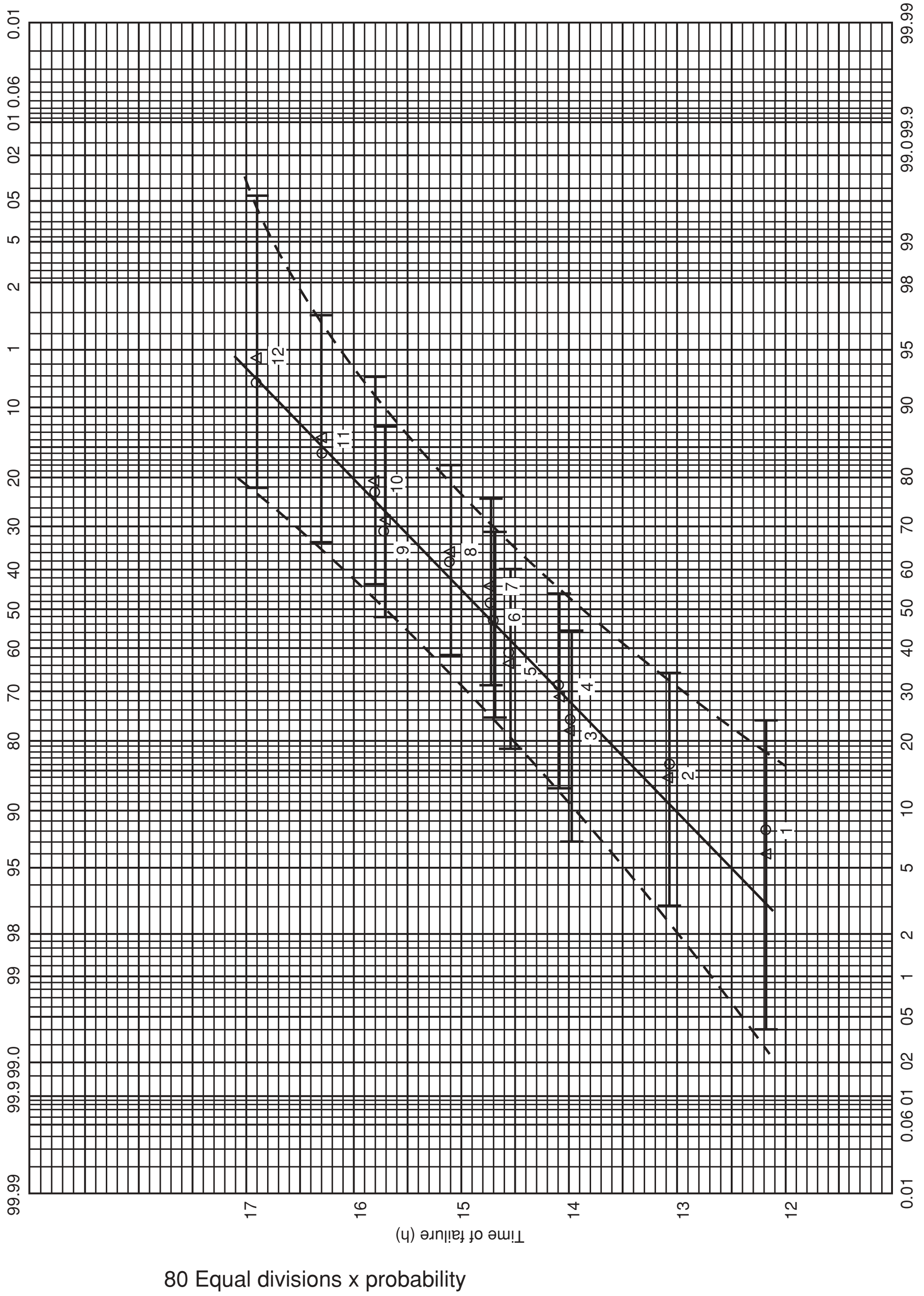


Figure 3.2 *Normal probability plot.*

In using life data analysis (as well as general statistics), one must be very cautious in qualifying the data. The first and foremost assumption that must be satisfied is that the collected data, or the sample, are truly representative of the population of interest. Most statistical analyses assume that the data are drawn at random from the population of interest. For example, if our job was to estimate the average life of humans, we would expect our sample to have the same make-up as the general population, that is equal numbers of men and women, a representative percentage of smokers and non-smokers, and so on. If we used a sample of ten male smokers to estimate life expectancy, the resulting analysis and prediction would most likely be biased and inaccurate. The assumption that our sample is truly representative of the population and that the test or use conditions are truly representative of the use conditions in the field must be satisfied in all analyses. Bad, or insufficient data, will almost always result in bad estimations, which has been summed up as 'Garbage in, garbage out.'

3.2.1 Complete Data

Complete data means that the value of each sample unit is observed or known. For example, if we had to compute the average test score for a sample of ten students, complete data would consist of the known score for each student. Likewise in the case of life data analysis, our data set if complete would be composed of the times-to-failure of all units in our sample. For example, if we tested five units and they all failed and their times-to-failure were recorded (see Figure 3.3) we would then have complete information as to the time of each failure in the sample.

3.2.2 Censored Data

In many cases when life data are analysed, all of the units in the sample may not have failed (i.e. the event of interest was not observed) or the exact times-to-failure of all the units are not known. This type of data is commonly called censored data. There are three types of possible censoring schemes, right censored (also called suspended data), interval censored and left censored.

3.2.3 Right Censored (Suspended)

The most common case of censoring is what is referred to as right censored data, or *suspended data*. In the case of life data, these data sets are composed of units that did not fail. For example, if we tested five units

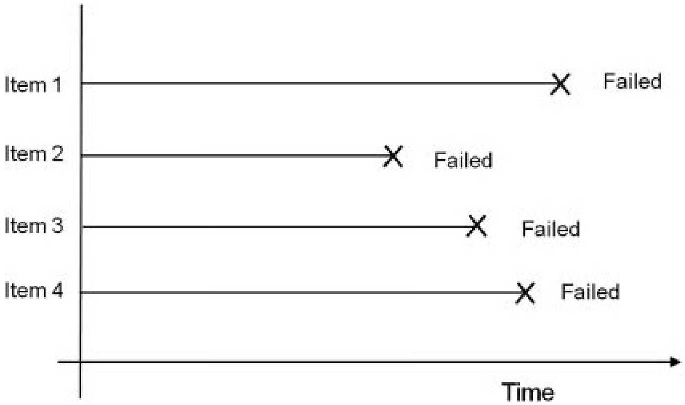


Figure 3.3 Complete data set.

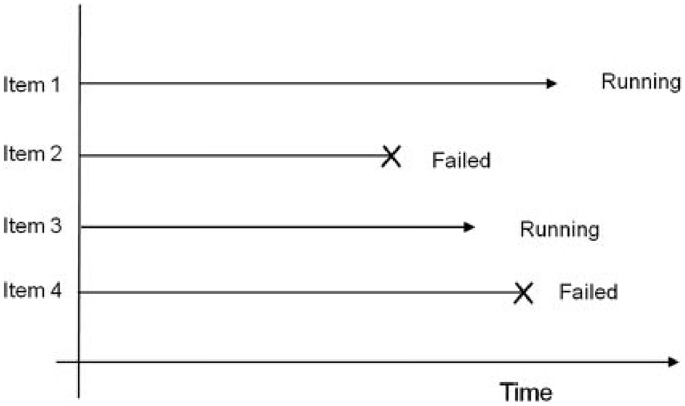


Figure 3.4 *Right censored data.*

and only three had failed by the end of the test, we would have suspended data (or right censored data) for the two non-failed units. The term 'right censored' implies that the event of interest (i.e. the time-to-failure) is to the right of our data point. In other words, if the units were to keep on operating, the failure would occur at some time after our data point (or to the right on the time scale), see Figure 3.4.

3.2.4 Interval Censored

The second type of censoring is commonly called *interval censored data*. Interval censored data reflects uncertainty as to the exact times the units failed within an interval. This type of data frequently comes from tests or situations where the objects of interest are not constantly monitored. If we are running a test on five units and inspecting them every 100 hours, we only know that a unit failed or did not fail between inspections. More specifically, if we inspect a certain unit at 100 hours and find it is operating and then perform another inspection at 200 hours to find that the unit is no longer operating, we know that a failure occurred in the interval between 100 and 200 hours. In other words, the only information we have is that it failed in a certain interval of time (see Figure 3.5). This is also often referred to as *inspection data*.

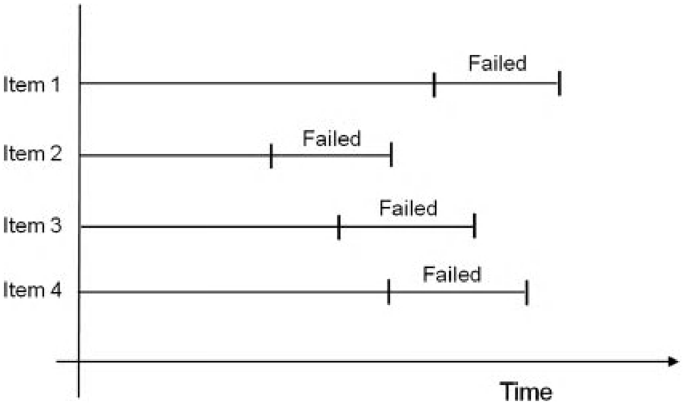


Figure 3.5 *Interval censored data.*

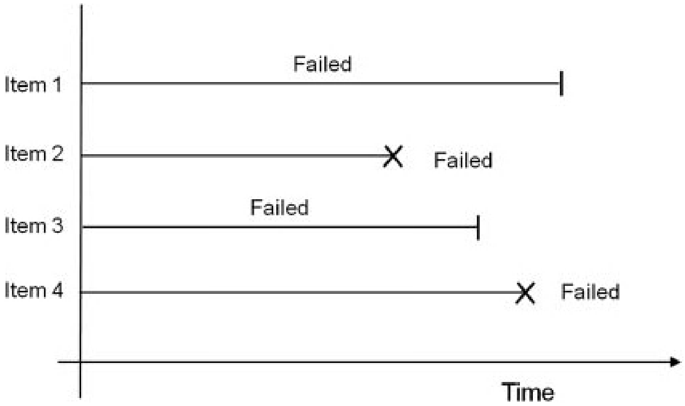


Figure 3.6 *Left censored data.*

3.2.5 Left Censored

The third type of censoring is similar to the interval censoring and is called left censored data. In left censored data, a failure time is only known to be before a certain time (see Figure 3.6) or to the left of our data point. For instance, we may conduct the first inspection at 100 hours and find that the part has already failed. In other words, it could have failed any time between 0 and 100 hours. This is identical to interval censored data in which the starting time for the interval is zero.

Complete data is much easier to work with than any type of censored data. While complete data sets and right censored data can often be analysed using graphical methods the left and interval censored data require more sophisticated approaches involving software tools. Some of these methods will be covered in this chapter.

3.3 Ranking of Data

Probability plotting (manual or computerized) is often based upon charting the variable of interest (time, miles, cycles, etc.) against the cumulative percentage probability. The data therefore need to be ordered and the cumulative probability of each data point calculated. This section will cover the methods used to rank various types of data and make them suitable for analysis and plotting.

3.3.1 Concept of Ranking

Data ranking provides an estimate of what percentage of population is represented by the particular test sample. Ranking presents an alternative to submitting data in frequency-histogram form due to the fact that in engineering applications often only small samples are available. For example, if we test five items and observe failures at 100, 200, 300, 400 and 500 hours respectively, then the rank of the first data point at 100 hours would be 20 % $\left(\frac{1}{5}\right)$, the rank of the second 40 % $\left(\frac{2}{5}\right)$, and so on, which is sometimes referred as *naïve rank estimator*. This, however, would statistically imply that 20 % of the population would have shorter life than 100 hours. By the same token assuming that the fifth sample represents 100 % of the population we concede to the assumption that all of the units in the field will fail within 500 hours. However, for probability plotting, it is better to make an adjustment to allow for the fact that each failure represents a point on a distribution.

To overcome this, and thus to improve the accuracy of the estimation mean and median ranking were introduced for probability plotting.

3.3.2 Mean Rank

Mean ranks are based on the distribution-free model and are used mostly to plot symmetrical statistical distributions, such as the normal. The usual method for mean ranking is to use $(N + 1)$ in the denominator, instead of N, when calculating the cumulative percentage position:

$$mean\ rank = \frac{i}{N+1} \tag{3.1}$$

3.3.3 Median Rank

Median ranking is the method most frequently used in probability plotting, particularly if the data are known not to be normally distributed. Median rank can be defined as the cumulative percentage of the population represented by a particular data sample with 50 % confidence. For example if the median rank of the second sample out of 5 is 31.47 % (see Table 3.1), that means that those two samples represent 31.47 % of the total population with 50 % confidence. There are different techniques which can be employed to calculate the median rank. The most common methods include *cumulative binomial* and its algebraic approximation.

3.3.4 Cumulative Binomial Method for Median Ranks

According to the cumulative binomial method, median rank can be calculated by solving the cumulative binomial distribution for Z (rank for the j^{th} failure) (Nelson 1982):

$$P = \sum_{k=j}^{N} \binom{N}{k} Z^k (1-Z)^{N-k} \tag{3.2}$$

where N is the sample size and j is the order number.

The median rank would be obtained by solving the following equation for Z:

$$0.50 = \sum_{k=j}^{N} \binom{N}{k} Z^k (1-Z)^{N-k} \tag{3.3}$$

The same methodology can then be repeated by changing P from 0.50 (50 %) to our desired confidence level. For $P = 95\,\%$ one would formulate the equation as:

$$0.95 = \sum_{k=j}^{N} \binom{N}{k} Z^k (1-Z)^{N-k} \tag{3.4}$$

Table 3.1 *Median rank for the sample size of 5.*

k	1	2	3	4	5
Median rank, if n = 5	12.94 %	31.47 %	50.0 %	68.53 %	87.06 %

As it will be shown in this chapter, the concept of ranking is widely utilised in both graphical plotting and computerized data analysis methods.

3.3.5 Algebraic Approximation of the Median Rank

The median ranks are well tabulated and published, also most statistical software packages have the option to calculate them (Minitab®, SAS®, etc.). For example, Weibull++® software has a 'Quick Calculator Pad' allowing the user to calculate any rank for any combination of sample size and number of failures. However when neither software nor tables are available or when the sample is beyond the range covered by the available tables the approximation formula (3.5), known as Benard's approximation, can be used. The *j*th rank value is approximated by:

$$\text{Median rank } r_j = \frac{j - 0.3}{N + 0.4} \tag{3.5}$$

Where: $j =$ failure order number and $N =$ sample size.

This approximation formula is widely utilized in manual probability plotting employing graphical methods with distribution papers, such as Weibull, Normal, Lognormal, Extreme Value and others.

3.3.6 Ranking Censored Data

When dealing with censored data, the probability plotting procedure becomes more complicated. The concept of censored data analysis is easier to explain with right censored data. Suspended items are not plotted as data points on the graph, but their existence affects the ranks of the remaining data points, therefore the ranks get adjusted. This is done to reflect the uncertainty associated with the unknown failure time for the suspended items. The derivation of adjusted median ranks for censored data is carried out as follows:

1 List order number (i) of failed items ($i = 1, 2, \ldots$).
2 List increasing ordered sequence of life values (t_i) of failed items.
3 Against each failed item, list the number of items which have survived to a time between that of the previous failure and this failure (or between $t = 0$ and the first failure).
4 For each failed item, calculate the *mean order number* i_{t_i} using the formula

$$i_{t_\mathrm{i}} = i_{t_{\mathrm{i}-1}} + N_{t_\mathrm{i}} \tag{3.6}$$

where

$$N_{t_i} = \frac{(n+1) - i_{t_i - 1}}{1 + (n - \text{number of preceding items})} \tag{3.7}$$

in which n is sample size.
5 Calculate median rank for each failed item, using the approximation from (3.5):

$$r_{t_i} = \frac{i_{t_i} - 0.3}{n + 0.4}\ \% \tag{3.8}$$

For the applications of this method, please see Example 3.2 later in this chapter (Section 3.4.2).

3.4 Weibull Distribution

In reliability engineering Weibull probability data analysis is probably the most widely utilized technique of processing and interpreting life data. One of many advantages is the flexibility of the Weibull distribution, easy interpretation of the distribution parameters, and their relation to the failure rates and the bathtub curve concept shown in Figure 1.6. In this chapter the Weibull distribution will be used to illustrate the techniques of probability plotting and life data analysis. Most of the same principles apply to data analysis involving other statistical distributions, many of which were covered in Chapter 2.

3.4.1 Two Parameter Weibull

The simpler version of the Weibull distribution is the 2-parameter model. In accordance with its name, this distribution is defined by two parameters. As described in Chapter 2, Section 2.6.6 the cumulative failure distribution function $F(t)$ is:

$$F(t) = 1 - \exp\left[-\left(\frac{t}{\eta}\right)^{\beta}\right] \tag{3.9}$$

where: $t =$ time.
$\beta =$ Weibull slope (the slope of the failure line on the Weibull chart), also referred as a *shape parameter*.
$\eta =$ Characteristic life, or the time by which 63.2 % of the product population will fail, also referred to as a *scale parameter*.

Equation (3.9) can be rewritten as:

$$\frac{1}{1 - F(t)} = \exp\left(\frac{t}{\eta}\right)^{\beta} \tag{3.10}$$

Or by taking two natural logarithms Eq. (3.10) will take the form of:

$$\ln \ln \frac{1}{1 - F(t)} = \beta(\ln t) - (\beta \ln \eta) \tag{3.11}$$

It can be noticed that (3.11) has a linear form of $Y = \beta X + C$.
Where:

$$\begin{aligned} X &= \ln t \\ Y &= \ln \ln \frac{1}{1 - F(t)} \\ C &= -\beta \ \ln \eta \end{aligned} \tag{3.12}$$

Therefore (3.11) represents a straight line with a slope of β and intercept C on the Cartesian X, Y coordinates (3.12). Hence, if the data follows the 2-parameter Weibull distribution, the plot of ln $\frac{1}{1-F(t)}$ against ln(t) will be a straight line with the slope of β.

3.4.2 Weibull Parameter Estimation and Probability Plotting

This type of scale is utilized in what is called *Weibull paper*, Figure 3.7. Weibull paper is constructed based on the X- and Y-transformations mentioned above, where the Y-axis (or double log reciprocal scale) represents unreliability $F(t) = 1 - R(t)$ and the X-axis represents time or other usage parameter (miles, km, cycles, runs, switches, etc.). Then, given the x and y value for each data point, each point can easily be plotted.

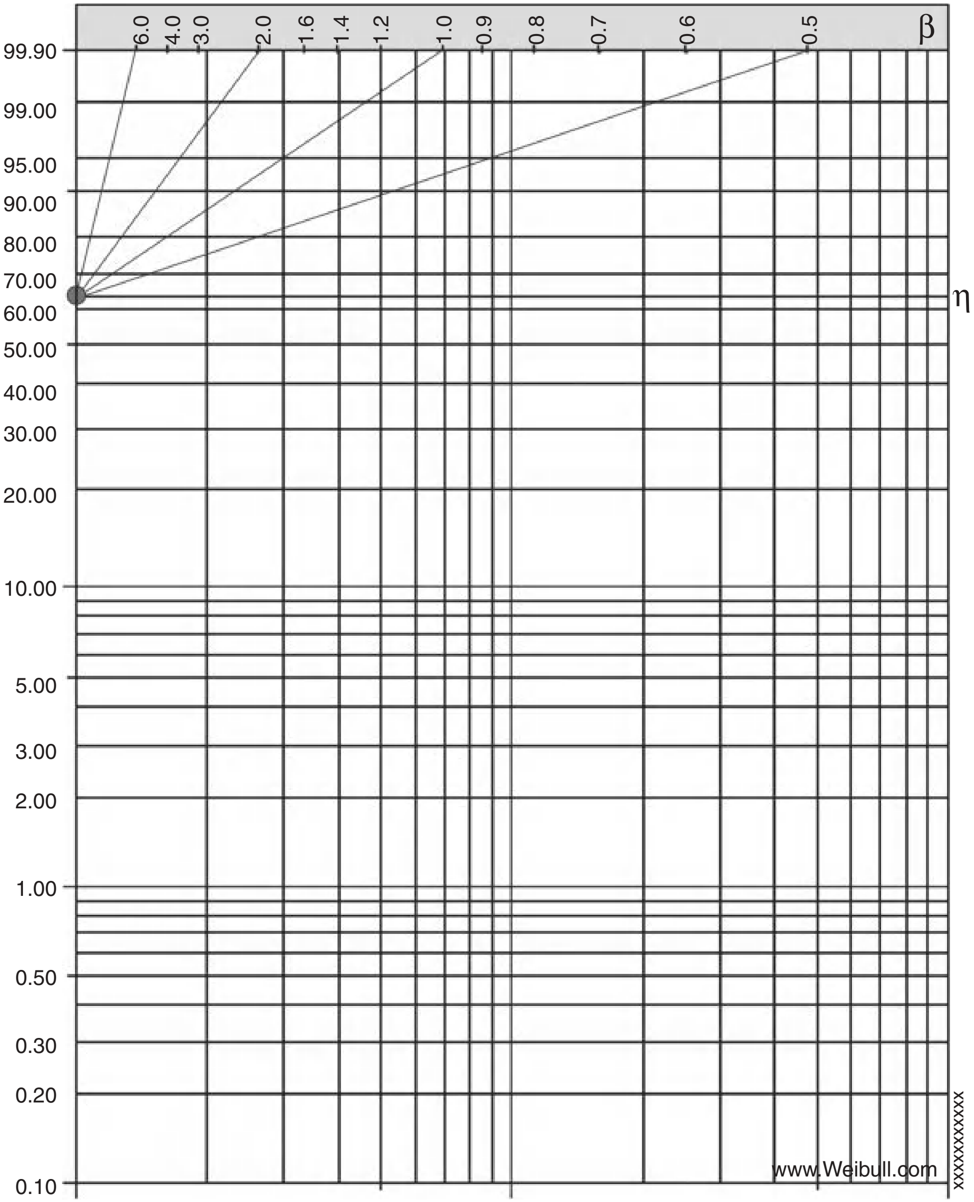


Figure 3.7 *Weibull probability paper. Abscissa - ln t, Ordinate -* $\ln\ln\frac{1}{1-F(t)}$.

The points on the plot represent our times-to- failure data. In probability plotting, we would use these times as our x values or time values. The appropriate y plotting positions, or the unreliability values would correspond to the median rank of each failure point.

After plotting each data point on the Weibull paper we draw the best fitting straight line through those points. Parameter β can be determined as the slope of that line (graphically or arithmetically) and parameter η can be determined as the time corresponding to 63.2 % unreliability on the Y-axis. To derive this number we substitute $t = \eta$ into (3.10) and calculate the cumulative failure function:

$$F(t) = 1 - \exp\left[-\left(\frac{\eta}{\eta}\right)^{\beta}\right] = 1 - \exp(-1) = 0.632 \ \ (63.2\%) \tag{3.13}$$

Even though Weibull paper is rarely used these days, understanding and ability to work with it provides a good foundation for using software tools. In addition, most commercially available Weibull analysis packages use the same graphics format as Weibull paper.

As mentioned before, one of the advantages of the Weibull distribution is its flexibility. For example, in the case of $\beta = 1$ the Weibull distribution reduces to the exponential distribution. When $\beta = 2$ the Weibull distribution resembles Rayleigh distribution (see, e.g. Hines and Montgomery, 1990). In the case of $\beta = 3.5$ the Weibull pdf will closely resemble the normal curve.

Example 3.1 Weibull Analysis using Rank Regression

Let us revisit the case where five units on test fail at 100, 200, 300, 400, 500 hours. Pairing those numbers with their median ranks (Table 3.1) would generate the following five data points: (100 hours, 12.94 %) (200 hours, 31.47 %) (300 hours, 50.0 %) (400 hours, 68.53 %) (500 hours, 87.06 %). Plotting those points on Weibull paper would produce the graph shown in Figure 3.8.

Once the line has been drawn, the slope of the line can be estimated by comparing it with the β-lines on the Weibull paper. Figure 3.8 shows the slope $\beta \approx 2.0$. According to Eq. (3.13), η is the life corresponding to 63.2 % unreliability, hence $\eta \approx 320$.

Therefore the reliability function for this product can be presented as a Weibull function:

$$R(t) = \exp\left[-\left(\frac{t}{320}\right)^{2.0}\right]$$

and can be calculated for any given time, t.

Having ranked and plotted the data (regardless of the particular statistical distribution), the question that often arises is *What is the best straight line fit to the data?* (assuming, of course, that there is a reasonable straight line fit). There can be a certain amount of subjectivity or even a temptation to adjust the line a little to fit a preconception. Normally, a line which gives a good ‘eyeball fit’ to the plotted data is satisfactory, and more refined manual methods will give results which do not differ by much. On the other hand, since the plotted data are cumulative, the points at the high cumulative proportion end of the plot are more important than the early points. However, a simple and accurate procedure to use, if rather more objectivity is desired, is to place a transparent rule on the last point and draw a line through this point such that an equal number of points lie to either side of the line.

Those are just general considerations for manual probability plotting. Clearly, computer software can process the data without subjectivity, with more precision, and with more analytical options of data processing. Computerized data analysis will be discussed in detail in Section 3.5.

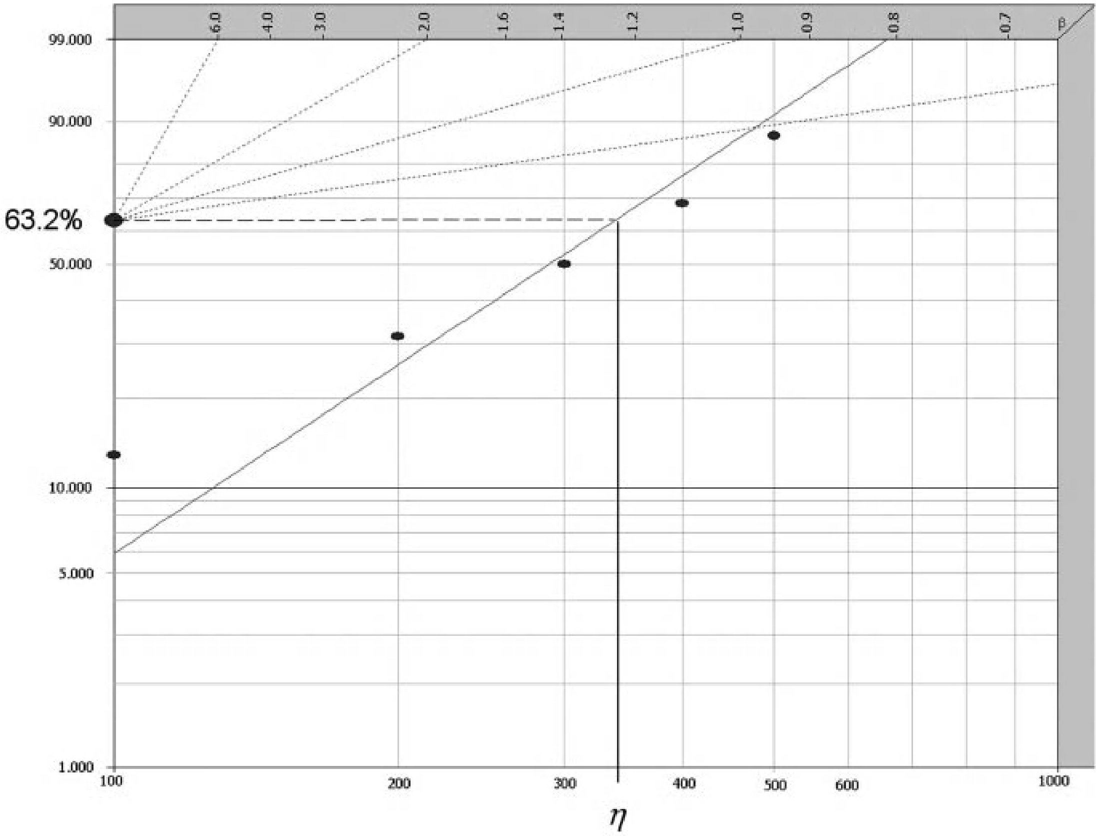


Figure 3.8 *Data plotted on Weibull paper for Example 3.1, $\beta \approx 2.0$ and $\eta \approx 320$.*

Example 3.2 Calculating Adjusted Ranks

Consider the same five items on test as in Example 3.1 only this time units 2 and 4 have not failed and were removed from test at the same times 200 and 400 hours respectively (summarized in Table 3.2). Calculate the adjusted ranks for the new data set.

From Eqs. (3.6) to (3.8)

$$N_{100} = \frac{5+1-0}{1+(5-0)} = 1.0 \qquad N_{300} = \frac{5+1-1}{1+(5-2)} = 1.25 \qquad N_{500} = \frac{5+1-2.25}{1+(5-4)} = 1.875$$

$$i_{100} = i_0 + N_{100} = 0 + 1.0 = 1.0 \qquad \begin{aligned} i_{300} &= i_{100} + N_{300} \\ &= 1.0 + 1.25 = 2.25 \end{aligned} \qquad \begin{aligned} i_{500} &= i_{300} + N_{500} \\ &= 2.25 + 1.875 = 4.125 \end{aligned}$$

$$r_{100} = \frac{1-0.3}{5+0.4} = 0.1296 \qquad r_{300} = \frac{2.25-0.3}{5+0.4} = 0.3611 \qquad r_{500} = \frac{4.125-0.3}{5+0.4} = 0.7083$$

Table 3.2 presents the summary of the steps in calculating the adjusted ranks for the three failure points. After the adjusted ranks are calculated the probability plotting follows the same procedure as in Example 3.1 only with three data points (#1, #3, #5).

Table 3.2 *Data summary and adjusted ranks calculation for Example 3.2.*

Item #	Time (hours)	Fail or Suspend	N_{t_i}	i_{t_i}	r_{t_i}
1	100	Failure 1	1.0	1.0	12.96 %
2	200	Suspended	—	—	—
3	300	Failure 2	1.25	2.25	36.11 %
4	400	Suspended	—	—	—
5	500	Failure 3	1.875	4.125	70.83 %

Even though the rank adjustment method is the most widely used method for performing suspended items analysis, it has some serious shortcomings. It can be noticed from this analysis of suspended items that only the position where the failure occurred is taken into account, and not the exact time-to-suspension. This shortfall is significant when the number of failures is small and the number of suspensions is large and not spread uniformly between failures, as with Example 3.2. That is the reason that in most cases with censored data the Maximum Likelihood method (see Section 3.5.2) is recommended to estimate the parameters instead of using least squares, presented in Section 3.5.1. The reason is that maximum likelihood does not look at ranks or plotting positions, but rather considers each unique time-to-failure or suspension.

3.4.3 Three Parameter Weibull

As mentioned in Chapter 2, the product cumulative failure distribution function $F(t)$ is presented in a form, that is slightly more complicated than (3.9) with an additional parameter γ:

$$F(t) = 1 - \exp\left[-\left(\frac{t-\gamma}{\eta}\right)^{\beta}\right] \tag{3.14}$$

Where γ = expected minimum life, also referred as *location parameter*, because it defines the starting location of the pdf graph along the X-axis of the coordinate system. (Other literature may use the characters $X_0, t_{0,}$ or ρ in place of γ for the location parameter.) Under the assumption of 3-parameter Weibull no failure of the product can possibly occur prior to the time γ, therefore it is also referred as *minimum life*.

A 3-parameter Weibull plot can no longer be represented by a straight line on a Weibull plot (see Figure 3.9) thus creating more difficulty for manual probability plotting. There is a technique for manual 3-parameter Weibull plotting, involving shifting every data point to the left (or right) by a certain value in the logarithmic scale until the data points become aligned. That shift value determines the minimum life γ, however computerized plotting would clearly provide a more accurate and certainly more expedient solution.

The inclusion of a location parameter for a distribution whose domain is normally $[0, \infty]$ will change the domain to $[\gamma, \infty]$, where γ can be either positive or negative. This can have some profound effects in terms of reliability. For a positive location parameter, this indicates that the reliability for that particular distribution is always 100 % up to that point γ. On some occasions the location parameter can be negative, which implies that failures theoretically occur before time zero. Realistically, the calculation of a negative location parameter is indicative of quiescent failures (failures that occur before a product is used for the first time) or of problems with the manufacturing, packaging or shipping processes.

Discretion must be used in interpreting data that do not plot as a straight line, since the cause of the non-linearity may be due to the existence of mixed distributions or because the data do not fit the Weibull distribution. The failure mechanisms must be studied, and engineering judgement used, to ensure that the correct interpretations are made. For example, in many cases wearout failure modes do exhibit a

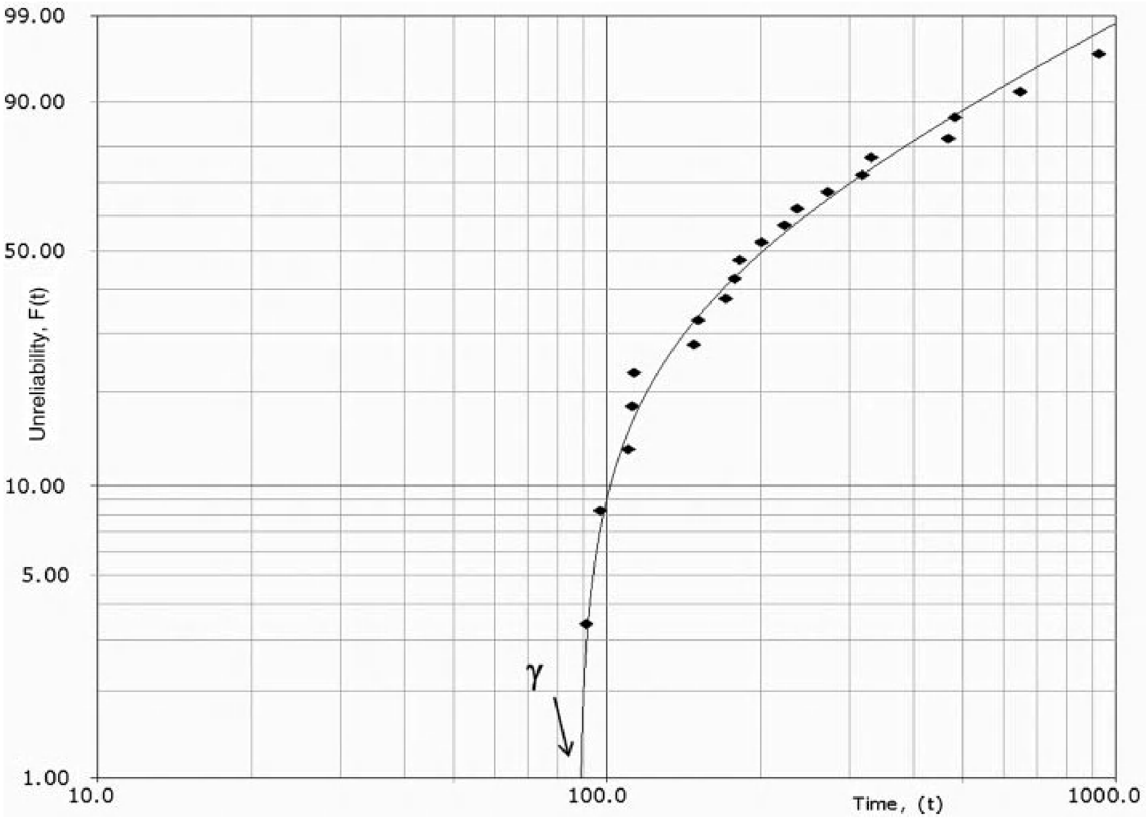


Figure 3.9 *3-parameter Weibull distribution plotted with Weibull++® (Reproduced by permission of ReliaSoft).*

finite failure-free life. Therefore a value for γ can sometimes be estimated from knowledge of the product and its application.

In quality control and reliability work we often deal with samples which have been screened in some way. For example, machined parts will have been inspected, oversize and undersize parts removed, and electronic parts may have been screened so that no zero-life ('dead-on-arrival') parts exist. Screening can show up on probability plots, as a curvature in the tails. For example, a plot of time to failure of a fatigue specimen will normally be curved since quality control will have removed items of very low strength. In other words, there will be a positive minimum life and be a good fit for 3-parameter Weibull.

Additional recommendations for preferring 3-parameter Weibull over 2-parameter Weibull include the number of data points (which should be no less than 10) and justification of the minimum life existence based on the failure mechanism. Therefore, better mathematical fit alone is not a good enough reason for choosing 3-parameter Weibull. Choice of the distribution will greatly affect reliability numbers, even between 2-parameter and 3-parameter Weibull! Both parameters β and η will be affected by that choice.

3.4.4 The Relationship of β-Parameter to Failure Rates and Bathtub Curve

As explained in Chapter 2, the value of β reflects the hazard function or the expected failure rate of the Weibull distribution and inferences can be drawn about a population's failure characteristics by considering

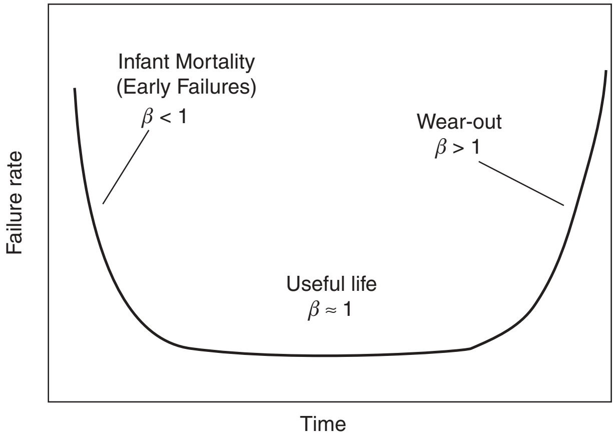


Figure 3.10 *Relationship between the bathtub curve and the Weibull slope β.*

whether the value of β is less than, equal to, or greater than one. The relationship between the value of β and the corresponding section on the bathtub curve (Figure 1.6) can be illustrated in Figure 3.10.

Since most Weibull analysis these days is done with commercially available software, the users may have a tendency to run an analysis and report the results without closely analysing their validity. Therefore the following guide would help to better evaluate the results and make an appropriate solution about the data based on the value of β:

If $\beta < 1$ indicates a decreasing failure rate and is usually associated with infant mortality, sometimes referred as *early failures*. It often corresponds to manufacturing related failures and failures recorded shortly after production. This can happen for several reasons including a proportion of the sample being defective or other signs of early failure.

If $\beta \approx 1$ a constant failure rate and is usually associated with useful life. Constant failure rate, which often corresponds to the mid-section of the life of the product and can be a result of random failures or mixed failure modes.

If $\beta > 1$ indicates an increasing failure rate and is usually associated with wearout, corresponding to the end life of the product with closer inter-arrival failure times. If recorded at the beginning of the product life cycle it can be a sign of a serious design problem or a data analysis problem.

If $\beta > 6$, it is time to become slightly suspicious. Although $\beta > 6$ is not uncommon, it reflects an accelerated rate of failures and fast wearout, which is more common for brittle parts, some forms of erosion, failures in old devices and less common for electronic systems. Some biological and chemical systems may have $\beta > 6$ value, for example human mortality, oil viscosity breakdown, and so on. Also, a large number of censored data points compared to complete data sets can result in high β. Different data analysis options can be recommended to re-evaluate the validity of the results.

If $\beta > 10$, it is time to become highly suspicious. Such a high β is not unheard of, but fairly rare in engineering practice. It reflects an extremely high rate of wearout, and not an expected value for the analysis of complete or nearly complete data set. However, it can be a result of highly censored data with a small number of failures (e.g. as an exercise, try the case with two units failing at 900 and 920 hours respectively and five units suspended at 1000 hours). Also a high β could be a result of stepped overstress testing, where environmental conditions become more and more severe with each step, therefore causing the parts to fail at the accelerated rate.

3.4.5 B$_X$-Life

Another parameter, that is used to specify reliability is the B-life, which is the time (or any other usage measure) by which a certain percent of the population can be expected to fail. It is expressed as B_X, where X is the percentage of the population failing. For example B_{10} life of 15 years would be equivalent to 90 % reliability for 15 year mission life. This relationship can be expressed by the equation:

$$R(B_X) = (100 - X)\% \tag{3.15}$$

and, as applied to Weibull distribution,

$$R(B_X) = \exp\left[-\left(\frac{B_X}{\eta}\right)^{\beta}\right] = 1 - \frac{X}{100} \tag{3.16}$$

3.5 Computerized Data Analysis and Probability Plotting

Computerized life data analysis in essence uses the same principles as manual probability plotting, except that it employs more sophisticated mathematical methodology to determine the line through the points, as opposed to just ‘eyeballing’ it. Modern data analysis software offers clear advantages by providing the capability to perform more accurate and versatile calculations and data plotting. This section will cover the two most commonly used techniques of computerized data analysis: Rank Regression and Maximum Likelihood Estimator (MLE). Probability plotting in this section will be done with the use of Weibull++® software. This program is widely utilized by reliability engineers worldwide and has enough versatility and statistical capability to handle multiple analytical tasks with various types of reliability data and a range of distribution functions. As mentioned before, even though most of the material in this chapter is applied to the Weibull distribution, the general principles remain the same regardless of the statistical distribution being modelled.

3.5.1 Rank Regression on X

One of the ways to draw the line through the set of data points is to perform a rank regression. It requires that a line mathematically be fitted to a set of data points such that the sum of the squares of the vertical or horizontal deviations from the points to the line is minimized.

Assume that a set of data pairs (x_1, y_1), (x_2, y_2), ..., (x_N, y_N) were obtained and plotted. Then, according to the least squares principle, which minimizes the horizontal distance between the data points and the straight line fitted to the data Figure 3.11, the best fitting straight line to these data is the straight line $x = \hat{a} + \hat{b}y$ such that:

$$\sum_{i=1}^{n} \left(\widehat{a} + \widehat{b}y_i - x_i\right)^2 = \min(a, b) \sum_{i=1}^{n} (a + by_i - x_i) \tag{3.17}$$

Where, $\widehat{a}$ and $\widehat{b}$ are the *least squares estimates* of a and b and N is the number of data points.

The solution of (3.17) (see ReliaSoft, 2008a) for $\widehat{a}$ and $\widehat{b}$ yields:

$$\hat{a} = \frac{\sum\limits_{i=1}^{N} x_i}{N} - \hat{b}\frac{\sum\limits_{i=1}^{N} y_i}{N} = \bar{x} - \hat{b}\bar{y} \tag{3.18}$$

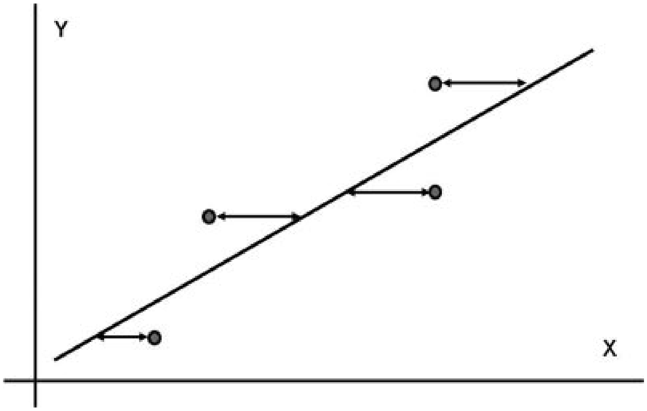


Figure 3.11 *Minimizing distance in the X-direction.*

and

$$\hat{b} = \frac{\sum_{i=1}^{N} x_i y_i - \frac{\sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N}}{\sum_{i=1}^{N} y_i^2 - \frac{\left(\sum_{i=1}^{N} y_i\right)^2}{N}} \tag{3.19}$$

One of the advantages of the rank regression method is that it can provide a good measure for the fit of the line to the data points. This measure is known as the *correlation coefficient*. In the case of life data analysis, it is a measure for the strength of the linear relation between the median ranks (Y-axis values) and the failure time data (X-axis values). The population correlation coefficient has the following form:

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{3.20}$$

where σ_{xy} is the covariance of x and y, σ_x is the standard deviation of x, and σ_y is the standard deviation of y (based on the available data sample). The estimate of the correlation coefficient for the sample of N data points can be found in ReliaSoft (2008a) or other statistical references.

The closer the correlation coefficient is to the absolute value of 1, the better the linear fit. Note that $+1$ indicates a perfect fit with a positive slope, while -1 indicates a perfect fit with a negative slope. A *perfect fit* means that all of the points fall exactly on a straight line. A correlation coefficient value of zero would indicate that the data points are randomly scattered and have no pattern or correlation in relation to the regression line model. ρ^2 is often used instead of ρ to indicate correlation, since it provides a more sensitive indication, particularly with probability plots.

As an alternative, the data can be analysed using *Rank Regression on Y*, which is very similar to Rank Regression on X with the only difference that the solution minimizes the sum of the Y-distances between the data points and the line. It is important to note that the regression on Y will not necessarily produce the same results as the regression on X, although they are usually close.

3.5.2 Maximum Likelihood Estimation (MLE)

Many computer-based methods present a probability plotting alternative to rank regression, an example is the Maximum Likelihood Estimator (MLE). The idea behind maximum likelihood parameter estimation is to determine the parameters that maximize the probability (likelihood) of the sample data fitting that distribution (see ReliaSoft, 2008a). Maximum likelihood estimation endeavours to find the most 'likely' values of distribution parameters for a set of data by maximizing the value of what is called the *likelihood function*. From a statistical point of view, the method of maximum likelihood is considered to be more robust (with some exceptions) and yields estimators with good statistical properties. In other words, MLE methods are versatile and apply to most models and to different types of data (both censored and uncensored).

If x is a continuous random variable with pdf:

$$f(x;\theta_1,\theta_2,\ldots,\theta_k),$$

where $\theta_1, \theta_2, \ldots \theta_\mathrm{k}$ are k unknown constant parameters that need to be estimated, conduct an experiment and obtain N independent observations, $x_1, x_2, \ldots, x_\mathrm{N}$ which correspond in the case of life data analysis to failure times. The likelihood function (for complete data) is given by:

$$L(x_1,x_2,\ldots,x_N\,|\,\theta_1,\theta_2,\ldots,\theta_k) = L = \prod_{i=1}^{N} f(x_i;\theta_1,\theta_2,\ldots,\theta_k) \quad i = 1,2,\ldots,N \tag{3.21}$$

The logarithmic likelihood function is:

$$\Lambda = \ln L = \sum_{i=1}^{N} \ln f(x_i;\theta_1,\theta_2,\ldots,\theta_k) \tag{3.22}$$

The maximum likelihood estimators (MLE) of $\theta_1, \theta_2, \ldots \theta_k$, are obtained by maximizing L or Λ.

By maximizing Λ, which is much easier to work with than L, the maximum likelihood estimators (MLE) of $\theta_1, \theta_2, \ldots \theta_k$ are the simultaneous solutions of k equations such that:

$$\frac{\partial(\Lambda)}{\partial\theta_\mathrm{j}} = 0, \quad j = 1,2,\ldots,k$$

Please note that many commercially available software packages plot the MLE solutions using median ranks (points are plotted according to median ranks and the line according to the MLE solutions). However as can be seen from Eq. (3.21), the MLE method is independent of any kind of ranks. For this reason, many times the MLE solution appears not to track the data on the probability plot. This is perfectly acceptable since the two methods are independent of each other, and in no way suggests that the solution is wrong.

More on Maximum Likelihood Estimator including the analysis of censored data can be found in ReliaSoft (2008a), Nelson (1982), Wasserman (2003) or Abernethy (2003).

Example 3.3 Illustrating MLE Method on Exponential distribution

This method is easily illustrated with the one-parameter exponential distribution. Since there is only one parameter, there is only one differential equation to be solved. Moreover, this equation is closed-form, owing to the nature of the exponential pdf. The likelihood function for the exponential distribution is given by:

$$L(\lambda \mid t_1, t_2, \ldots, t_n) = \prod_{i=1}^{n} f(t_i) = \prod_{i=1}^{n} \lambda e^{-\lambda t_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} t_i}$$

where λ is the parameter we are trying to estimate. For the exponential distribution, the log-likelihood function (3.22) takes the form:

$$\Lambda = \ln(L) = n \ \ln(\lambda) - \lambda \sum_{i=1}^{n} t_i$$

Taking the derivative of the equation with respect to λ and setting it equal to zero results in:

$$\frac{\partial \Lambda}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} t_i = 0$$

From this point, it is a simple matter to rearrange this equation to solve for λ:

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} t_i}$$

This gives the closed-form solution for the MLE estimate for the one-parameter exponential distribution. As can be seen, parameter λ is estimated as the inverse of MTTF (Mean Time to Failure). Obviously, this is one of the most simplistic examples available, but it does illustrate the process. The methodology is more complex for distributions with multiple parameters and often does not have closed-form solutions. Applying MLE to censored data is also fairly complex and mathematically involved process reserved for computerized solutions. More details on MLE mathematics can be found in ReliaSoft (2011), Nelson (1982) and other Bibliography at the end of this chapter.

3.5.3 Recommendation on Using Rank Regression vs. MLE

Rank regression methods often produce different distribution parameters than MLE, therefore it is a logical question to ask which method should be applied with which type of data. Based on various studies (see ReliaSoft (2008a), Wasserman (2003) and Abernethy (2003)) regression generally works best for data sets with smaller (<30) sample sizes (as sample sizes get larger, 30 or more, these differences become less important) that contain only complete data. Failure-only data is best analysed with rank regression on X, as it is preferable to regress in the direction of uncertainty. When heavy or uneven censoring is present and/or when a high proportion of interval data is present, the MLE method usually provides better results. It can also provide estimates with one or no observed failures, which rank regression cannot do.

In the case where it is not clear which method would provide more accurate results, it is advisable to run both methods and compare the results. The following scenarios are possible:

- The RR and MLE results do not differ much.
- The results differ and one method might provide unreasonable values of β- (too high or too low).
- The results differ and one method provides the values of β which do not fit the model IFR vs. DFR (see Section 3.4.4).

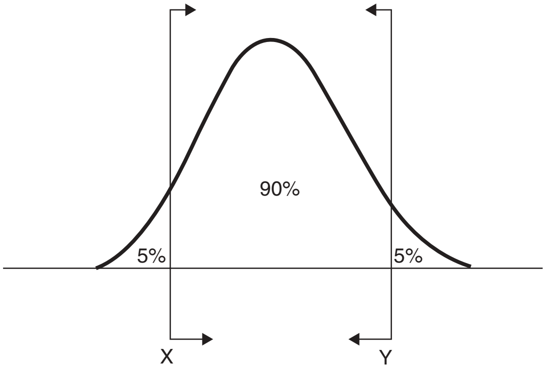


Figure 3.12 *Two-sided 90 % confidence bounds.*

Those outcomes can help to make a more intelligent choice of the analysis method. It is also advisable to try several distributions using both MLE and RR The choice of MLE vs. RR may also affect the choice of best fit distribution for a particular data set, for example 2P Weibull may show the best fit using MLE, while the normal distribution can be the best fit for the same data set using RRX method (see Section 3.7 for more details).

3.6 Confidence Bounds for Life Data Analysis

Because life data analysis results are estimates based on the observed lifetimes of a sample of units, there is uncertainty in the results due to the limited sample sizes. *Confidence bounds* (also called *confidence intervals*) briefly covered in Chapter 2 are used to quantify this uncertainty due to sampling error by expressing the confidence that a specific interval contains the quantity of interest. Whether or not a specific interval contains the quantity of interest is unknown. For continuous distributions, confidence bounds calculations involve the area under pdf curve corresponding to the percentage confidence sought for the particular solution, Figure 3.12.

When we use two-sided confidence bounds (or intervals), we are looking at a closed interval where a certain percentage of the population is likely to lie. That is, we determine the values, or bounds, between which lies a specified percentage of the population. For example, when dealing with 90 % two-sided confidence bounds of $[X, Y]$, we are saying that 90 % of the population lies between X and Y with 5 % less than X and 5 % greater than Y. Figure 3.12.

With one-sided intervals we define the target value to be greater or less than the bound value. For example, if X is a 95 % upper one-sided bound; this would imply that 95 % of the population is less than X. If X is a 95 % lower one-sided bound, this would indicate that 95 % of the population is greater than X, Figure 3.13.

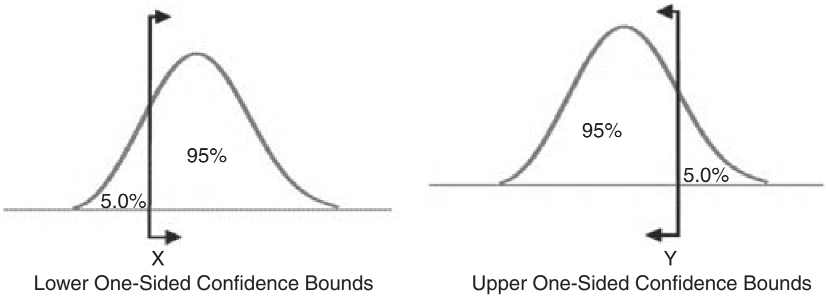

95%
5.0%
X
Lower One-Sided Confidence Bounds
95%
5.0%
Y
Upper One-Sided Confidence Bounds

Figure 3.13 *One-sided confidence bounds.*

Table 3.3 *5 and 95 % ranks for the sample size of 5.*

k	1	2	3	4	5
5 % rank, if n = 5	1.02 %	7.64 %	18.92 %	34.25 %	54.92 %
95 % rank, if n = 5	45.07 %	65.74 %	81.07 %	92.36 %	98.98 %

Care must be taken to differentiate between one- and two-sided confidence bounds, as these bounds can take on identical values at different percentage levels. For example, the analyst would use a one-sided lower bound on reliability, a one-sided upper bound for percentage failing under warranty and two-sided bounds on the parameters of the distribution. Note that one-sided and two-sided bounds are related. For example, the 90 % lower two-sided bound is the 95 % lower one-sided bound and the 90 % upper two-sided bound is the 95 % upper one-sided bound. See Figure 3.12 and Figure 3.13.

3.6.1 Confidence Intervals for Weibull Data

Weibull analysis can also be done with various degrees of confidence level. The rank calculations were made using the median rank, which corresponds to 50 % confidence level. Thus, for example, for 2-sided 90 % confidence level similar to Figure 3.12, we would need to use different ranks for the data plotting. Specifically, we would need to graph the same failure points using 5 % and 95 % ranks to provide [5 %; 95 %] confidence bounds. That can be accomplished by applying, for example, the cumulative binomial method per equation (3.2). Table 3.3 provides 5 % and 95 % ranks respectively for the sample size of 5. More 5 % and 95 % rank values for various sample sizes are provided in the tables in Appendix 4.

As applied to Example 3.1 the first failure at 100 hours has the following ranking: 1.02 % (5 % rank) and 45.07 % (95 % rank). Thus 90 % confidence interval for unreliability at 100 hours, *F(100 hrs)* would be between 1.02 % and 45.07 %. Similarly, we can plot 5 % and 95 % ranks for each of the five failure points resulting in the graph Figure 3.14.

The confidence bounds Figure 3.14 are quite wide. With 90 % confidence, the reliability at 100 hours of operation can be anywhere between 54.93 and 98.98 %. The reason for such wide intervals is the small number of data points.

As the number of samples increases the respective ranks would come in smaller increments and thus closer together. As a result, the confidence bounds become narrower and thus closer to the median rank straight line.

To illustrate the effect of a larger sample size on the confidence bounds, consider the 5 % rank of the 2nd sample out of 10. Quantitatively 2 out of 10 represents the same 20 % of the population as the 1st sample out of 5, however the 5 % rank in this case is 3.68 % as opposed to 1.02 % for the five samples (see Appendix 4). Similarly the 95 % rank of the 2nd sample out of 10 is 39.2 % as opposed to 45.07 % for the 1 out of 5 case. In both cases the 10-sample ranks would be plotted closer to the centre line, which would result in narrower confidence bounds for the same data fit.

3.6.2 Individual Parameter Bounds

It is often important to derive the confidence limits on the parameters of the distribution since decisions may be based upon those values. For example, the β-value characterizes the trend in failure rate of the product population and its place on the bathtub curve, Figure 3.10. Individual parameter bounds are used to evaluate uncertainty in terms of the expected (or mean) values of the parameters. For bounds on individual parameters, statistical software usually provides the Fisher matrix, likelihood ratio, beta binomial, Monte Carlo and Bayesian confidence bounds.

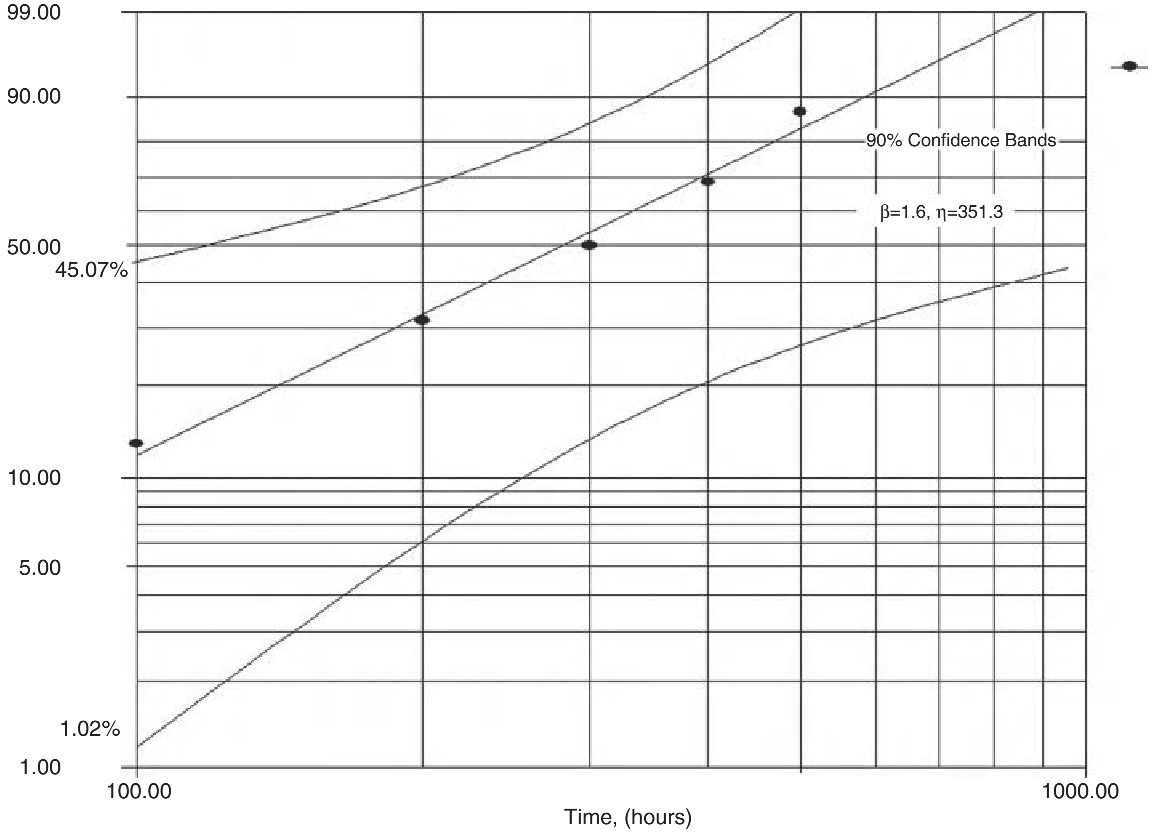


Figure 3.14 *Weibull++® two-sided 90 % confidence bounds for Weibull distribution (Reproduced by permission of ReliaSoft).*

3.6.2.1 Fisher Matrix Bounds

Fisher Matrix bounds are used widely in many statistical applications. These bounds are calculated using the Fisher information matrix. The inverse of the Fisher information matrix yields the variance-covariance matrix, which provides the variance of the parameter $Var(\hat{\theta})$, see ReliaSoft (2008a). The bounds on the parameters are then calculated using the following equations:

$$\begin{aligned}\text{Lower bound} &= \hat{\theta} - Z_{\alpha/2}\sqrt{Var(\hat{\theta})}\\ \text{Upper bound} &= \hat{\theta} + Z_{\alpha/2}\sqrt{Var(\hat{\theta})}\end{aligned} \tag{3.23}$$

where: $\hat{\theta}$ is the estimate of mean value of the parameter θ.
$Var(\hat{\theta})$ is the variance of the parameter.
$\alpha = 1 - C$, where C is the confidence level.
$z_{\alpha/2}$ is the standard normal statistic. Excel function = -NORMSINV(α/2) or see Appendix 1.

Parameters that do not take negative values are assumed to follow the lognormal distribution and the following equations are used to obtain the confidence bounds:

$$\begin{aligned}\text{Lower bound} &= \hat{\theta} \cdot \exp\left[-(z_{\alpha/2}/\hat{\theta}) \cdot \sqrt{Var(\hat{\theta})}\right]\\ \text{Upper bound} &= \hat{\theta} \cdot \exp\left[(z_{\alpha/2}/\hat{\theta}) \cdot \sqrt{Var(\hat{\theta})}\right]\end{aligned} \tag{3.24}$$

Despite its mathematical intensity, the Fisher matrix method for confidence bounds is widely utilized by most commercially available software packages. For more details see ReliaSoft (2008a).

3.6.2.2 Likelihood Ratio Bounds

For data sets with very few data points, Fisher matrix bounds are not sufficiently conservative. The likelihood ratio method produces results that are more conservative and consequently more suitable in such cases. (For data sets with larger numbers of data points, there is not a significant difference in the results of these two methods.) Likelihood ratio bounds are calculated using the likelihood function as follows:

$$-2 \cdot \ln\left(\frac{L(\theta)}{L(\hat{\theta})}\right) \geq \chi^2_{\alpha;k} \tag{3.25}$$

where: $L(\theta)$ is the likelihood function for the unknown parameter θ.
$L(\hat{\theta})$ is the likelihood function calculated at the estimated parameter value $\hat{\theta}$.
$\alpha = 1 - C$, where C is the confidence level.

$\chi^2_{(\alpha,k)}$ is the Chi-Squared statistic with k degrees of freedom, where k is the number of quantities jointly estimated. Excel function = CHIINV(α, k) or see Appendix 2.

In the calculations of the likelihood ratio bounds on individual parameters, only one degree of freedom ($k = 1$) is used in the $\chi^2_{(\alpha,k)}$ statistic. This is due to the fact that these calculations provide results for a single confidence region. For more details, refer to ReliaSoft (2008a) and Nelson (1982).

3.6.2.3 Beta Binomial Bounds

The beta-binomial method of confidence bounds calculation is a non-parametric approach to confidence bounds calculations that involves the use of rank tables or rank calculations as described in (3.2) and similar to the calculations presented in Section 3.6.1.

3.6.2.4 Monte Carlo Confidence Bounds

In this method Monte Carlo simulation is used to create many samples from a known distribution. The proportion of times that the true value of a parameter is contained in the confidence interval is estimated along with the width (or half-width) of the intervals. For more details on generating Monte Carlo confidence bounds see Wasserman (2003).

3.6.2.5 Bayesian Confidence Bounds

This method of estimating confidence bounds is based on the Bayes theorem, where prior information is combined with sample data in order to get new parameter distributions called *posterior* and make inferences about model parameters and their functions. The posterior yields estimates and Bayesian confidence limits for the parameters. Details on the calculation of these bounds are available in ReliaSoft (2008a) and ReliaSoft (2006).

Example 3.4 Manual Calculation of Confidence Bounds on the Weibull Parameter β

Manual calculations of confidence limits on the shape parameter β can be done using the Figure 3.15 graph containing factors F_β against sample size for different confidence levels (99 %, 95 %, 90 %) on β. Figure 3.15 is based on a graphical approximation of Fisher matrix bounds (3.24). The upper and lower confidence limits are then

$$\beta_{\text{Upper}} = \hat{\beta} F_\beta$$

$$\beta_{\text{Lower}} = \hat{\beta} \frac{1}{F_\beta}$$

Derive the upper and lower confidence limits if $n = 10$, $\beta = 1.6$ for $C = 90\,\%$ (double-sided).

From Figure 3.15, $F_\beta = 1.37$; therefore,

$$\beta_{\text{Upper}} = 1.6 \times 1.37 = 2.19$$

$$\beta_{\text{Lower}} = \frac{1.6}{1.37} = 1.17$$

that is we have a 90 % confidence that $2.19 \geq \beta \geq 1.17$.

3.6.3 Alternative Methods for Calculating Confidence Bounds

Obtaining confidence bounds, especially on censored data is an involved process and can be done in several different ways. Besides using the cumulative binomial method per (3.2) described in Section 3.6.1, there are other techniques described in Section 3.6.2. Those methods have a high level of mathematical complexity and are utilized mostly by commercially available statistical software packages. Based on a variety of methods, confidence bounds in life data analysis applications may differ based on the value being estimated. Several software packages including Weibull++ give the user the option to perform separate calculation of confidence bounds on time (Type 1) and on reliability (Type 2).

Confidence bounds on time (Type 1) can be estimated by first solving the Weibull reliability Eq. (3.9) for time T:

$$T = \widehat{\eta} \left(\ln \frac{1}{R} \right)^{\frac{1}{\beta}} \tag{3.26}$$

The confidence bounds on T will be defined by the upper and lower bounds on the estimates of the individual Weibull parameters $\widehat{\beta}$ and $\widehat{\eta}$ depending on which analytical method is chosen (see Section 3.6.2) and also will depend on the value of the third parameter R.

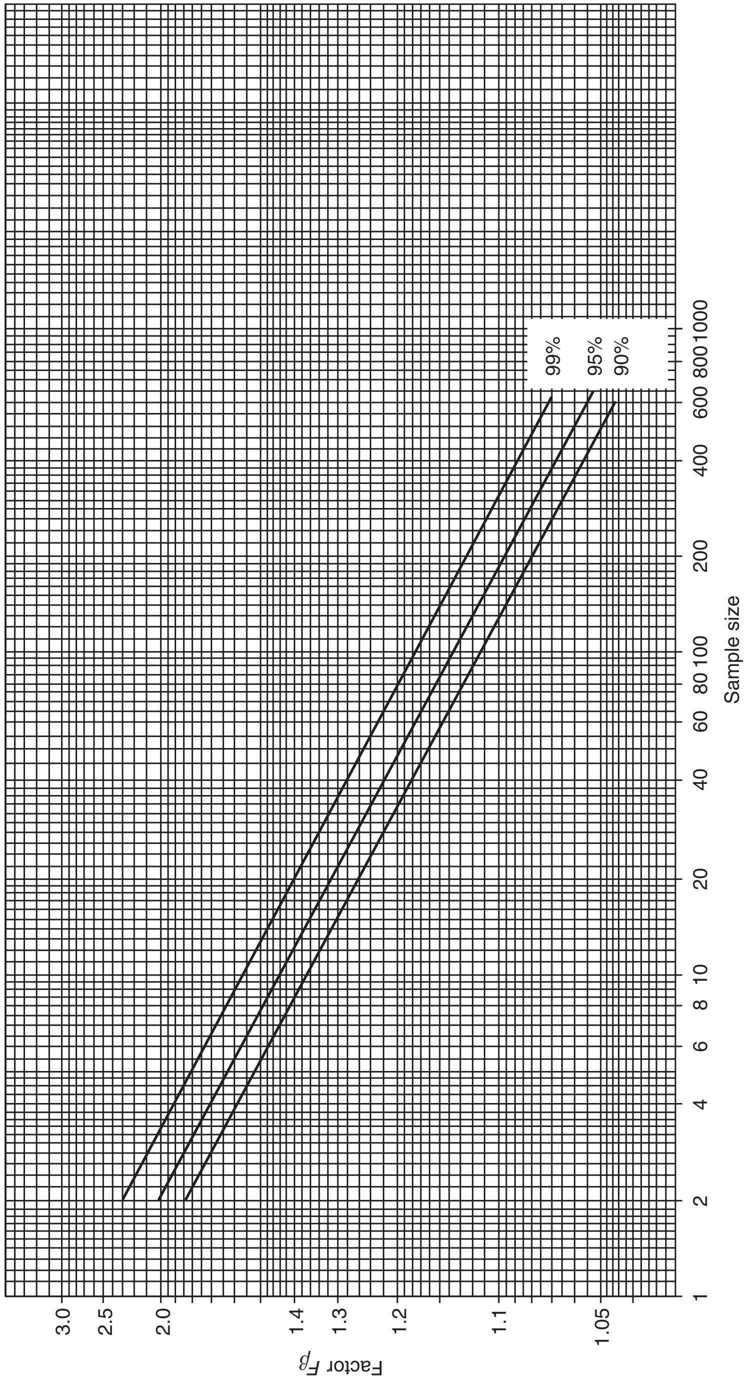


Figure 3.15 *Confidence limits for shape parameter β for different confidence values.*

For the confidence bounds on reliability (Type 2) the Weibull Eq. (3.9) can be written in its traditional form:

$$R(t) = \exp\left[-\left(\frac{t}{\widehat{\eta}}\right)^{\widehat{\beta}}\right] \tag{3.27}$$

The confidence bounds on a reliability for a given time value are calculated in the same manner as the bounds on time. The only difference is that the solution must now be considered in terms of β, η and t, where t is now considered to be a parameter instead of R, since the value of time must be specified in advance. In this approach, the confidence bounds on R can be defined by the upper and lower bounds on the estimates of Weibull parameters $\widehat{\beta}$ and $\widehat{\eta}$, which would produce results different from those for time (Type 1). The Fisher Matrix method, which is most commonly used in software packages, would generate a different set of confidence bounds depending on whether it is Type 1 or Type 2. The choice between Type 1 and Type 2 would depend on the value that we are trying to estimate. For example, to determine the B_{10}-life (time by which 10 % of the units have failed) then we would use confidence bounds on time (Type 1). To estimate reliability at a certain point of time (e.g. 200 hours of operation) then the confidence bound on reliability should be displayed. More details on those calculations can be found in the Life Data Analysis ReliaSoft (2008a).

Figure 3.16 shows the 90 % confidence bounds on B_{10}-life (Type 1) and 90 % confidence bounds on reliability at 200 hours (Type 2) based on the analysis of the data presented in Example 3.1.

Note that similar approaches can be used to calculate the confidence bounds on the variables calculated with the data analysis involving most other statistical distributions.

3.7 Choosing the Best Distribution and Assessing the Results

A range of statistical distributions is available to reliability practitioners. The most commonly used were presented in Chapter 2 and due to the available computing power all of them can be applied to probability plotting. However, that presents a question of which distribution model should a practitioner use to fit the data in a given particular data set? The distribution which is likely to provide the best fit to a set of data is not always readily apparent. The process of determining the best model can be fairly comprehensive and usually starts with evaluating the available distributions based on how well they fit the data, that is mathematical goodness of fit. However, in engineering applications having the best mathematical fit is not sufficient, since the chosen statistical distribution should also be appropriate for the physical nature of observed failures. Therefore both approaches need to be carefully considered before selecting the best mathematical model to analyse the data.

3.7.1 Goodness of a Distribution Fit

Statistical goodness-of-fit tests should be applied to test the fit to the assumed underlying distributions. There are many statistical tools that can help in deciding whether or not a distribution model is a good choice from a statistical point of view. Section 3.5 presents an overview of the tools and the approaches, which are often based on the type of data (complete vs. censored), number of data points (small vs. large number) and other criteria. The rank regression (least squares RRX and RRY) and Maximum Likelihood Estimator (MLE) methods have been covered earlier in Section 3.5. The correlation coefficient ρ, Eq. (3.20) is a measure of how well the straight line fits the plotted data for the rank regression method. For MLE, the likelihood L, Eq. (3.21) would best characterize its goodness of fit. In general, the best fit would provide ρ closest to 1.0

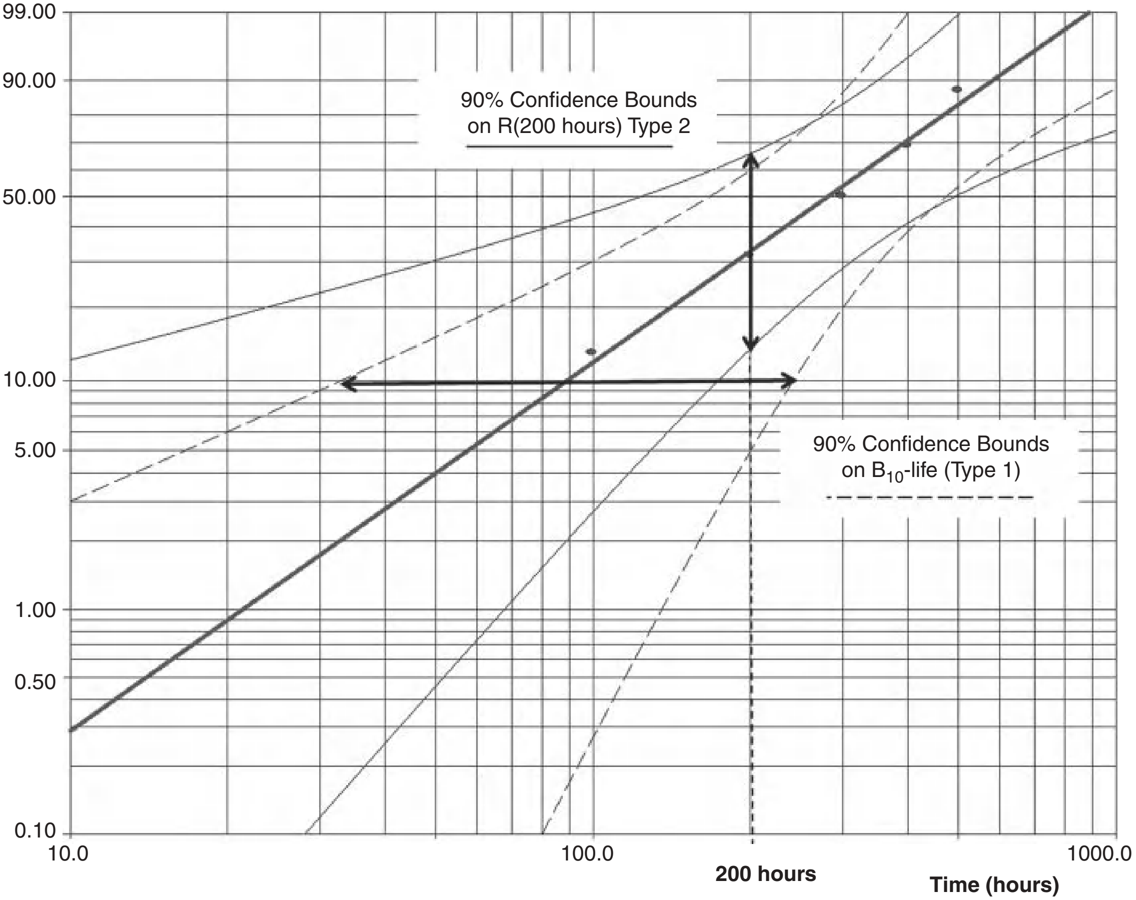
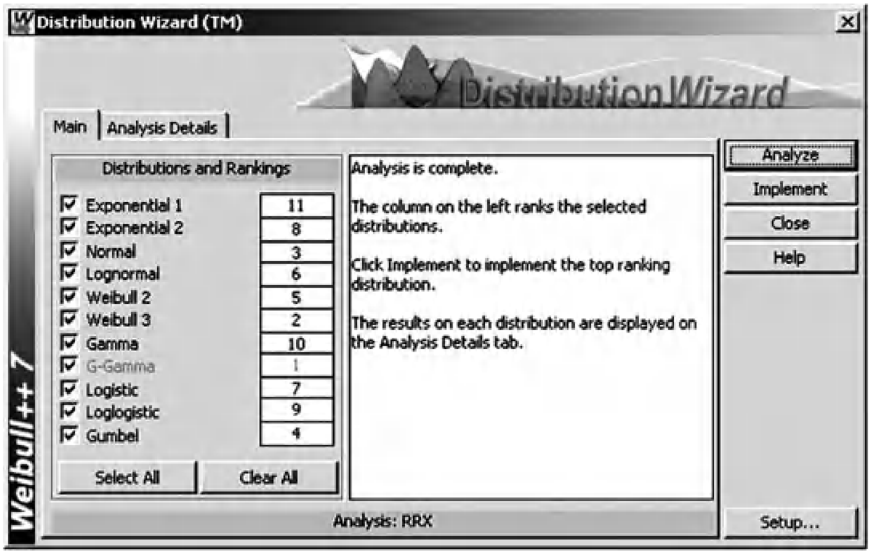


Figure 3.16 Weibull++® 90 % confidence bounds on B_{10}-life and Reliability (Example 3.1) (Reproduced by permission of ReliaSoft).

(Rank Regression) or the highest likelihood value in case of MLE. Goodness-of-fit tests including χ^2 (chi-square) and Kolmogorov–Smirnov (K–S) methods are covered in Chapter 2.

It is also important to ensure that the time axis chosen is relevant to the problem; otherwise misleading results can be generated. For example, if a number of items are tested, and the running times recorded, the failure data could show different trends depending upon whether all times to failure are taken as cumulative times from when the test on the first item is started, or if individual times to failure are analysed. If the items start their tests at different elapsed or calendar times the results can also be misleading if not carefully handled. For example, a trend might be caused by exposure to changing weather conditions, in which case an analysis based solely on running time could conceal this information. The same considerations would apply to warranty data analysis, where time count starts from the date of sales or production. In those cases the life data should not be counted as a calendar time, but rather as an individual age of the warranted part. The methods of exploratory data analysis, described in Chapter 13, can be applied when appropriate.

Statistical software can be used to rank different distributions based on the best mathematical fit depending on which statistical method is chosen. Figure 3.17 shows the ReliaSoft Weibull++ tool called



Distributions and Rankings	
☑ Exponential 1	11
☑ Exponential 2	8
☑ Normal	3
☑ Lognormal	6
☑ Weibull 2	5
☑ Weibull 3	2
☑ Gamma	10
☑ G-Gamma	1
☑ Logistic	7
☑ Loglogistic	9
☑ Gumbel	4

Analysis is complete.

The column on the left ranks the selected distributions.

Click Implement to implement the top ranking distribution.

The results on each distribution are displayed on the Analysis Details tab.

Analysis: RRX

Figure 3.17 *Weibull++® distribution ranking based on the goodness of fit (Rank Regression on X) (Reproduced by permission of ReliaSoft).*

Distribution Wizard® allowing the user to compare different distribution models based on how well they fit the analysed data.

The table in Figure 3.17 shows the ranking of the available distributions for the data in Example 3.1. Distribution Wizard uses a combination of goodness of fit (K–S method), correlation coefficient and likelihood value to determine the best fitting distribution. Note that for the MLE data analysis the ranking of the distributions will be different, since a quantitative measure of goodness of fit (combination of weight factors) will depend on the chosen data analysis method. However, the ranking based on the goodness of fit as the one shown in Figure 3.17 should only be considered as the first step in the decision-making process. The next step should be based on evaluating data groups/patterns, failure modes, amount of data and other considerations presented later in this section.

3.7.2 Mixed Distributions

A plot of failure data may relate to one failure mode or to multiple failure modes within an item. If a straight line does not fit the failure data, particularly if an obvious change of slope is apparent, the causes of failure should be investigated to estimate the possible number of failure modes. For example, after a certain length of time on test, a second failure mode may become apparent, or an item may have two superimposed failure modes. In such cases each failure mode must be isolated and analysed separately. However, such separation is appropriate only if the failure processes are independent, that is there is no interaction. In cases where it is difficult or impossible to separate the failure modes the life data can be analysed as a homogeneous population with a changing trend. In most cases it can be approached as a mixed Weibull distribution.

Many of the commercially available software packages can handle clusters of data points and to fit the data into mixed distributions. For example Figure 3.18 shows three distinct groups of data points clearly showing different trends and their slopes. One group includes the data points between 1 and 100 hours, another between 100 and 1000 hours and the third is between 1000 and 10000 h.

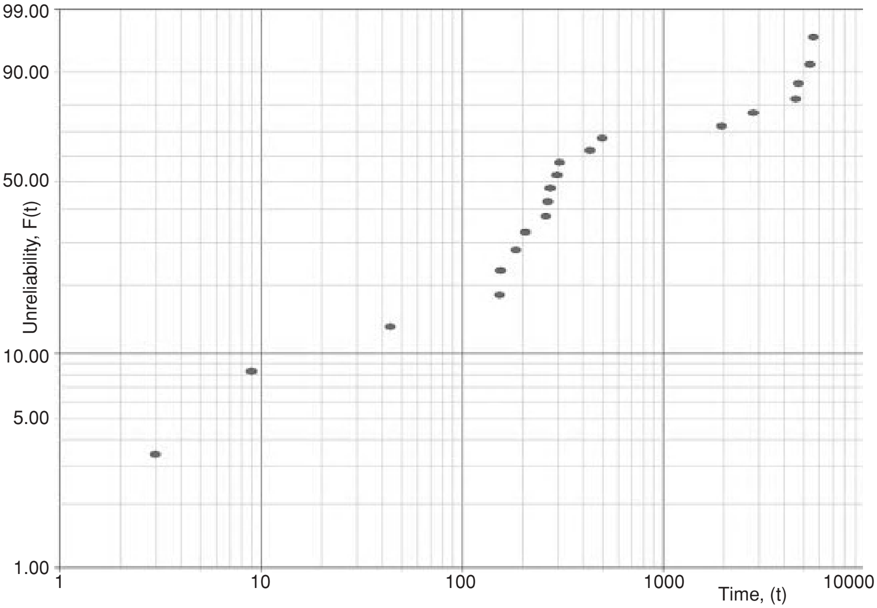


Figure 3.18 *Separate groups of data points (Reproduced by permission of ReliaSoft).*

These data points can be split into three groups and fit into the mixed Weibull distribution as shown in Figure 3.19. It is important to note that each group has different β-slope and therefore different failure rates and data pattern.

Figure 3.19 shows the plot of mixed Weibull distribution as unreliability vs. time as a fitted curve with three sections and failure rate vs. time, which follows the bathtub curve pattern due to the failure rate changing from one group to another.

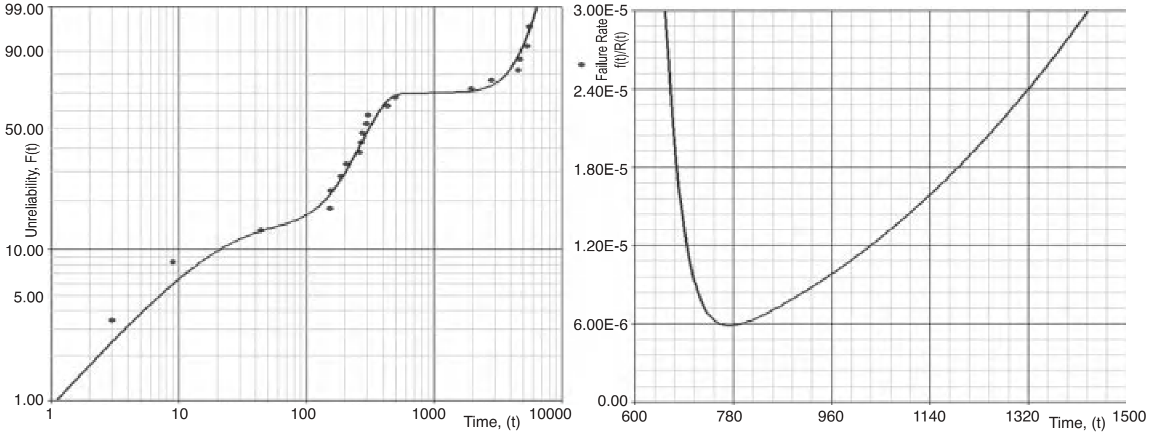


Figure 3.19 *Mixed Weibull distribution plotted with Weibull++® (Reproduced by permission of ReliaSoft): (A) Probability Plot (unreliability vs. time); (B) Hazard Rate.*

3.7.3 Engineering Approach to Finding Best Distribution

While analysing life data and finding the best statistical model for it (Section 3.7.1), it is important to remember that we are dealing with failures of a real engineering device and the knowledge and understanding of that device will also be a factor in determining the likely best distribution. The engineering considerations should include the following:

Maturity of the system and its place on the bathtub curve.
Type of failures (failure modes and physics of failure).
Sample size and size of the population it represents.

Maturity of the system will affect the trend for its failure rate and in the case of the Weibull distribution can be easily characterized by the β-value as presented by the bathtub model (see Figure 3.10).

As mentioned before, situations, where $\beta < 1$ represents the decreasing failure rate (DFR). This is usually a sign of manufacturing defects and/or immature products. Therefore, if that life data comes from the mature system, which has been in the field for a significant amount of time, some other distributions need to be considered. For example, normal and extreme value distributions exhibit a perpetually increasing failure rate. It is important to understand that the forecast based on the distribution with decreasing failure rate may significantly underestimate the percentage of the failed population because it will extrapolate that trend into the useful life section of the bathtub curve. Therefore Weibull curves with DFR trends in some cases need to be merged with different distributions (see, e.g. Kleyner and Sandborn, 2005).

If a constant hazard rate ($\beta \approx 1$) is apparent, this can be an indication that multiple failure modes exist or that the time-to-failure data are suspect. This is often the case with systems in which different parts have different ages and individual part operating times are not available. A constant hazard rate can also indicate that failures are due to external events, such as faulty use or maintenance. If none of those conditions apply, again other statistical options should be considered.

Whilst wearout failure modes are characterised by $\beta > 1$, the Weibull parameters of the wearout failure mode can be derived if it is possible to identify the defective items and analyse their life and failure data separately from those of the non-defective items.

Failure modes and physics of the observed failures should play an important part in the analysis. It is very important that the physical nature of failures be investigated as part of any evaluation of failure data, if the results are to be used as a basis for engineering action, as opposed to logistic action. For example, if an assembly has several failure modes, but the failure data are being used to determine requirements for spares and repairs, then we do not need to investigate each mode and the parameters applicable to the mixed distributions are sufficient. If, however, we wish to understand the failure modes in order to make improvements, we must investigate all failures and analyse the distributions separately, since two or more failure modes may have markedly different parameters, yet generate overall failure data which are fitted by a distribution which is different to that of any of the underlying ones. Also certain failure modes have known historical β-slopes, which can be considered as guidelines. For example, in the electronics industry low cycle solder fatigue is characterized by the β-slopes in the range of [2.0; 4.0] and metal fatigue failures often have β-slopes in the range of [3.0; 6.0]. According to Abernethy (2003) ball bearing failures have $\beta \approx 2.0$, rubber V-belts $\beta \approx 2.5$, corrosion-erosion [3.0; 4.0]. Obviously those are just generic numbers, but they can be used as an additional verification of the results, especially in the cases of analysing censored data. Also if the failures are caused by some extreme conditions, like extreme high values of electrical load or extreme low values of bond strength, then extreme value distribution may be the best way to fit the data regardless of the goodness of fit ranking.

It is also important to understand that some distributions are not as flexible as others. While the Weibull distribution can fit the data with any type of failure rate, other distributions exhibit only a certain pattern, for example, the normal and extreme value distributions would demonstrate only increasing failure rate (IFR), where lognormal displays more complex pattern of early IFR transitioning later into DFR. The physics of failure can also affect the choice between the 2-parameter and the 3-parameter Weibull distribution. Existence of the additional parameter γ provides an opportunity to better fit the data (e.g. Distribution Wizard in Figure 3.17 ranks 3P Weibull higher than 2P Weibull), however it would also mean that by choosing 3-parameter Weibull we accept the existence of the failure free time. In engineering practice only a few failure mechanisms have a true failure-free period (e.g. corrosion, fatigue), in the most cases failure modes can exhibit themselves from the very beginning of the product life. Therefore when choosing 3-parameter Weibull one should be able to justify the reason why the product is very unlikely to fail before it passes γ hours (days, miles, etc.) of operation.

Sample size and size of the population also can play a critical part in defining the mathematical model. It is important to realize that cumulative probability plots are to a large extent self-aligning, since succeeding points can only continue upwards and to the right. Goodness-of-fit tests will nearly always indicate good correlation with any straight line drawn through such points. Analysing the large amount of failure data, such as warranty claims, sometimes presents a problem where several distributions may show a high degree of goodness of fit due to the fact that the number of failed parts can still be relatively small compared to the overall size of the population. On the chart in Figure 3.1 it would show a very small shaded section of $f(t)$ on the left compared to the rest of the distribution. In the automotive warranty databases it is not uncommon to process several thousand failure data points based on the 3-year warranty claims from the population of several hundred thousand vehicles. In those cases 2-parameter Weibull often shows almost identical likelihood value with the lognormal data fit, however the extrapolation of those distribution data to 10-year life may show significant differences in the forecast number of failures (sometimes a factor of 2). In those cases understanding of failure rate trend (increasing-decreasing pattern of the lognormal distribution vs. increasing pattern of Weibull with $\beta > 1$) can help to make a correct choice. This example shows that it is doubly important for large data sets to carefully consider the engineering aspects of the failures. In addition to that, large data sets, such as warranty databases, may contain secondary failures, which would require a totally different approach to probability plotting.

On the other spectrum are small samples or small numbers of actual failures. It is not uncommon during product testing to experience one or two failures out of relatively small sample size of five to ten units. Most software packages can handle two or even one failure using the MLE technique; however the plotting produces questionable results with arbitrary values or β-slopes. In some of those cases the knowledge of the expected β-value could help to force-fit the data into the most probable distribution and obtain more practical results than in the case of straight mathematical fitting.

Other data analysis criteria may also apply, for example the earliest time(s) to failure could be more important than later times (or vice versa), so the less important data points could be considered as ‘outliers’ unless there is adequate engineering justification for not doing so. Overall, it is clear that in the majority of the cases the principle of *engineering trumps mathematics* should apply in choosing the best distribution for the life data analysis.

Example 3.5 Breaking Strength of a Wire

The breaking strength of a long wire was tested, using a sample of 15 equal lengths. Since the strength of a wire can be considered to depend upon the existence of imperfections, the extreme value distribution of minimum values might be an appropriate fit. The results are shown in Table 3.4. Since this would be a

Table 3.4 Breaking strengths of 15 samples of wire of equal length.

Rank order	Cumulative probability per cent (median ranks, c.d.f)	Breaking strength (N)
1	4.5	76
2	10.9	75
3	17.4	74
4	23.9	72.5
5	30.4	72
6	36.9	69
7	43.4	69
8	50.0	65
9	56.5	64
10	63.0	63
11	69.5	62
12	76.0	61
13	82.5	58
14	89.0	52
15	95.4	48

distribution of minimum value it will be left-skewed, and the data are therefore arranged in descending order of magnitude. Plotting the data the other way around would generate a convex curve, viewed from above. A plot of an extreme value distribution of maximum values would be made with the data in ascending order.

The mode $\hat{\mu}$ can be estimated directly from the data. In this case $\hat{\mu} = 69.0$ N.

σ, the measure of variability, can be derived from the expression in Chapter 2, Section 2.6.7.1.

$$\text{Mean} = \hat{\mu} - 0.577\hat{\sigma}$$

$$\hat{\sigma} = \frac{\hat{\mu} - \text{mean}}{0.577} = \frac{69.0 \ - \ 65.4}{0.577} = 6.30\,\text{N} \tag{3.28}$$

The value of $1/\sigma$ is sometimes referred to as the *Gumbel slope*. Now, applying the cdf equation from Chapter 2, $\text{F}(y) = 1 - \exp\left[-\exp\left(y\right)\right]$, where $y = (x - \mu)/\sigma$, we can calculate $x = 50.3$N for $F(y) = 0.05$.

In this case the probability scale represents the cumulative probability that the breaking strength will be greater than the value indicated, so that there is a 95 % probability that the strength of a wire *of this length* will be greater than 50.3N. If the wire is longer it will be likely to be weaker, since the probability of its containing extreme value imperfections will be higher.

The return period *1/F*(*y*) represents the average value of *x* (e.g. number of times, or time) between recurrences of a greater (or lesser) value than represented by the return period. Therefore, there is a 50 % chance that a length of wire of breaking strength less than 50.3N will occur in a batch of 1/0.05 = 20 lengths. The return period is used in forecasting the likelihood of extreme events such as floods and high winds, but it is not often referred to in reliability work.

The use of Weibull++$^{®}$ provides a more refined extreme value solution presented in Figure 3.20. It calculates using the rank regression the parameters $\hat{\mu} = 69.2858$ and $\hat{\sigma} = 7.2875$.

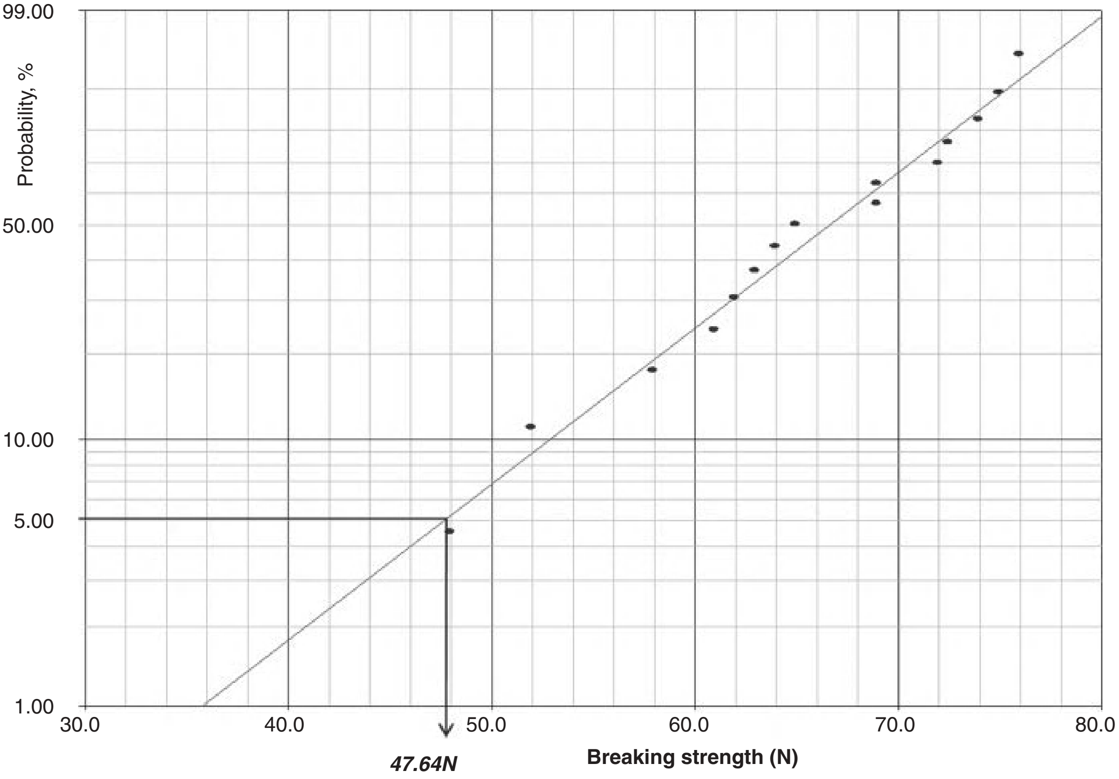


Figure 3.20 *Probability plot of the breaking strength (Weibull++®), Extreme value distribution (Reproduced by permission of ReliaSoft).*

Figure 3.20 demonstrates that there is 95 % probability of breaking strength being higher than 47.64N. This number is 5.3 % lower than the earlier value of 50.3N obtained by the manual calculations.

3.8 Conclusions

Probability plotting methods can be very useful for analysing reliability data; however the methods presented in this chapter apply to circumstances where items can only fail once. This distinction is important, since the methods, and the underlying statistical theory, assumes that the individual times to failure are independently and identically distributed (IID). That is, that failure of one item cannot affect the likelihood of, or time to, failure of any other item in the population, and that the distribution of times to failure is the same for all the failures considered. If these conditions do not hold, then the analysis can give misleading results. Techniques for analysing reliability data when these conditions do not apply (e.g. the failed units can be repaired) are described in Chapter 13.

Probability plotting and life data analysis can also be used for analysing other IID data, such as sample measurements in quality control. It is important to understand the decision making process leading to the best statistical model to analyse the data, especially with today's range of available software based tools.

The technique of choosing the best distribution should involve mathematical goodness of fit tests. This should be combined with engineering judgement, including understanding of the relevant physics or other causes of failures. It is essential that practical engineering criteria are applied to all phases of the data analysis and the interpretation of the results.

Questions

Some of the problems below require life data analysis software. If such software is not available a trial version of Weibull++ can be downloaded from: http://www.reliasoft.com/downloads.htm.

1. a Explain briefly (and in non-mathematical terms) why, in Weibull probability plotting, the *i*th ordered failure in a sample of *n* is plotted at the 'median rank' value rather than simply at *i*/*n*.
   b Planned replacement is to be applied to a roller bearing in a critical application: the bearing is to be replaced at its B_{10} life. Ten bearings were put on a test rig and subjected to realistic operating and environmental conditions. The first seven failures occurred at 370, 830, 950, 1380, 1550 and 1570 hours of operation, after which the test was discontinued. Estimate the B_{10} life from (i) the data alone; (ii) using a normal plot (normal paper can be downloaded from www.weibull.com/GPaper/index.htm). Comment on any discrepancy.
   c Process this data choosing the 2-parameter Weibull distribution. What is the difference in B_{10} value?
2. a Twenty switches were put on a rig test. The first 15 failures occurred at the following numbers of cycles of operation: 420, 890, 1090, 1120, 1400, 1810, 1815, 2150, 2500, 2510, 3030, 3290, 3330, 3710 and 4250. Plot the data on Weibull paper and give estimates of (i) the shape parameter β; (ii) the mean life μ; (iii) the B_{10} life; (iv) the upper and lower 90 % confidence limits on β; (v) the upper and lower 90 % confidence limits on the probability of failure at 2500 operations. Finally, use the Kolmogorov-Smirnov test to assess the goodness-of-fit of your data.
   b Alternatively, find the solution to the part (a) using a software. If using Weibull++ run Distribution Wizard to compare your choice against other different distributionns in the package. What is the best fit distribution for this data.
3. Six electronic controllers were tested under accelerated conditions and the following times to failure were observed: 46, 64, 83, 105, 123 and 150 hours. Do the following:
   a Determine how you would classify this data, that is individual, grouped, suspended, censored, uncensored, and so on.
   b Select rank regression (least squares) method on X as the parameter estimation method and determine the parameters for this data using the following distributions and plot the data for each distribution. From the plot, note how well you think each distribution tracks the data, that is how well does the fitted line track the plotted points?
      i 2-Parameter Weibull.
      ii 3-Parameter Weibull.
      iii Normal.
      iv Log-normal.
      v Exponential.
      vi Extreme Value (Gumbel).
      vii Gamma.
4. A pump used in large quantities in a sewage works is causing problems owing to sudden and complete failures. There are two dominant failure modes, impeller failure (I) and motor failure

(M). These modes are thought to be independent. Records were kept for 12 of these pumps, as follows:

Pump no	Age at failure (h)	Failure mode
1	1180	M
2	6320	M
3	1030	I
4	120	M
5	2800	I
6	970	I
7	2150	I
8	700	M
9	640	I
10	1600	I
11	520	M
12	1090	I

Estimate Weibull parameters for each mode of failure.

If using Weibull++ you can differentiate the failure modes by different 'Subset ID' and run the 'Batch Auto run' option.

5. A type of pump used in reactors at a chemical processing plant operates under severe conditions and experiences frequent failures. A particular site uses five reactors which, on delivery, were fitted with new pumps. There is one pump per reactor: When a pump fails, it is returned to the manufacturer in exchange for a reconditioned unit. The replacement pumps are claimed to be 'good as new'. The reactors have been operating concurrently for 2750 h since the plant was commissioned, with the following pump failure history:
   Reactor 1 – at 932, 1374 and 1997 h.
   Reactor 2 – at 1566, 2122 and 2456 h.
   Reactor 3 – at 1781 h.
   Reactor 4 – at 1309, 1652, 2337 and 2595 h.
   Reactor 5 – at 1270 and 1928 h.
   a Calculate the Laplace statistic (Eq. 2.46) describing the behaviour of the total population of pumps as a point process.
   b Estimate Weibull parameters for both new and reconditioned pumps.
   c In the light of your answers to (a) and (b), comment on the claim made by the pump manufacturer.
6. A vehicle manufacturer has decided to increase the warranty period on its products from 12 to 36 months. In an effort to predict the implications of this move, it has obtained failure data from some selected fleets of vehicles over prolonged periods of operation. The data below relate to one particular fleet of 20 vehicles, showing the calendar months in which a particular component failed in each vehicle. (There is one of these components per vehicle and when repaired, it is returned to 'as good as new' condition.) If there are no entries under 'Component failure dates', that vehicle had zero failures. The data are up-to-date to October 1995, at which data all the vehicles were still in use. The component currently gives about 3 % failures under warranty.

Vehicle	Start date	Component failure dates
1	May 93	—
2	Jun 93	Nov 93, Jul 94
3	Jun 93	—
4	Aug 93	Feb 95
5	Oct 93	Jan 95
6	Oct 93	Oct 94
7	Oct 93	Feb 95
8	Oct 93	Sep 94, Mar 95
9	Nov 93	—
10	Nov 93	Dec 94
11	Dec 93	Jan 95, Jul 95
12	Jan 94	—
13	Jan 94	—
14	Feb 94	—
15	Feb 94	—
16	Jul 94	—
17	Jul 94	Feb 95
18	Aug 94	—
19	Dec 94	Aug 95
20	Feb 95	—

Use any suitable method to estimate the scale and shape parameters of a fitted Weibull distribution, and comment on the implications for the proposed increase in warranty period.

7. The data below refer to failures of a troublesome component installed in five similar photocopiers in a large office. When photocopier fails, it is repaired and is ‘good as new’.

Machine no	Cumulative copies at which failures occurred	Current cumulative copies
1	13 600, 49 000	64 300
2	16 000, 23 800, 40 400	60 000
3	18 700, 28 900	46 700
4	22 200	40 600
5	6 500	39 000

a Estimate the parameters of a Weibull distribution describing the data.

b Calculate a Laplace trend statistic (Eq. 2.46) and, in the light of its value, discuss whether your answer to part (a) is meaningful.

8. The data below relate to failures of terminations in a sample of 20 semiconductor devices. Each failure results from breaking of either the wire (W) or the bond (B), whichever is the weaker. The specification requirement is that fewer than 1 % of terminations shall have strengths of less than 500 mg.

a Estimate Weibull parameters for (i) termination strength; (ii) wire strength; (iii) bond strength. Comment on the results.

b If using Weibull++ select rank regression on X (RRX) and run 'Distribution Wizard' for this data. Choose the best fitting statistical distribution for those results and re-evaluate the strength termination requirement.

Failure load (mg)	B or W	Failure load (mg)	B or W
550	B	1250	B
750	W	1350	W
950	B	1450	B
950	W	1450	B
1150	W	1450	W
1150	B	1550	B
1150	B	1550	W
1150	W	1550	W
1150	W	1850	W
1250	B	2050	B

9. Derive the solution for the MLE estimate of the parameters of the normal distribution $\widehat{\mu}$ and $\widehat{\sigma}$ similar way to how it is done in Example 3.3.
10. Seventeen electronic units were put through a 1000 hr temperature test. Every 200 hr the number of units that failed was counted. At 200 hr of testing, one failure was observed; at 400 hr two more failures were observed; at 600 hr four more failures were observed; at 800 hr five more failures were observed; and five more failures were observed at 1000 hr. Find the β and η of the 2-Parameter Weibull pdf representing this data using Rank Regression on X (least squares) and Maximum Likelihood Estimation. Compare RRX vs. MLE results. Explain the difference.
11. While fitting data with Weibull distribution (either using software or a Weibull paper), can you determine by the β-value if your data can be potentially fitted better with other distributions, such as normal or exponential?
12. Automotive manufacturer is testing the ignition system for the number of on/off switches. Seven ignition devices were tested to failure with the following results: 10 522, 14 232, 17 811, 21 762, 29 830, 39 411 and 43 901 switches. Using the software or Weibull probability paper determine
    a the parameters of 2-parameter Weibull distribution.
    b 2-sided 90 % confidence on reliability of the system at 10 000 switches.
13. As discussed in Section 3.4.4, run the analysis of the case where 2 units failing at 900 and 920 hours and the remaining 5 units are suspended at 1000 hours. Analyse the Weibull distribution parameters. What is the beta value? Explain the reason for such a high value.
14. Analysis of the interval data: Eight electronic units were released in a small test market. Every 200 hours of operation the units functionally tested for failures and malfunctions. As a result 2 units showed some sign of malfunctioning sometime between 400 and 600 hr, 2 units begin malfunctioning between 600 and 800 hr, 1 was considered failed between 800 and 1000 hr and 3 are still operating in the field after 1500 hours:
    a Determine the parameters of the 2-Parameter Weibull distribution using Rank Regression (least squares) on X.
    b Obtain the probability plot for this data.
    c Estimate the B_{10} life for this device.
    d Try different options with MLE vs. rank regression, Weibull vs. lognormal distribution. Compare the B_{10} life for different methods.

15. Given the Weibull parameters of $\beta = 1.86$ and $\eta = 21\ 620$ h with a population of 1600 units:
    a How many failures can be expected when each unit reaches 1500 hours?
    b 3000 hours?
16. 50 units were placed on test without continuous monitoring. After 100 hours of high temperature exposure the units were tested for functionality. Four units have failed and the remaining ones were still functional. What type of life data we are dealing with?

Bibliography

Abernethy, R. (2003) *The New Weibull Handbook*, 5th edn, Dr. Robert Abernethy.

Hines, W. and Montgomery, D. (1990) *Probability and Statistics in Engineering and Management Science*, 3rd edn, Wiley.

Kleyner A. and Sandborn P. (2005) *A warranty forecasting model based on piecewise statistical distributions and stochastic simulation*. Reliability Engineering and System Safety, **88**(3), 207–214.

Lawless, J. (2002) *Statistical Models and Methods for Lifetime Data*, Wiley.

Meeker, W. and Escobar, L. (1998) *Statistical Methods for Reliability Data*, Wiley.

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

ReliaSoft (2006) *Life Data Analysis Reference: Confidence Bounds*. Online tutorial. Available at: http://www.weibull.com/LifeDataWeb/confidence_bounds.htm.

ReliaSoft (2008a) *Life Data Analysis Reference*, ReliaSoft Publishing.

ReliaSoft (2008b) *Weibull++® User's Guide*, ReliaSoft Publishing.

ReliaSoft (2011) *Probability Plotting Papers*. Available at http://www.weibull.com/GPaper/index.htm.

Wasserman, G. (2003) *Reliability Verification Testing, and Analysis in Engineering Design*, Marcel Dekker.

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

4
Monte Carlo Simulation

4.1 Introduction

Monte Carlo (MC) simulation is a useful tool for modelling phenomena with significant uncertainty in inputs and has a multitude of applications including reliability, availability and logistics forecasting, risk analysis, load-strength interference analysis (Chapter 5), random processes simulation including repairable systems (Chapter 13), probabilistic design, uncertainty propagation, geometric dimensioning and tolerancing, and a variety of business applications.

The concept of the Monte Carlo method comes from the gaming tables at the casinos of Monte Carlo. It is a class of probabilistic computational algorithms that rely on repeated sampling of random variables of interest to compute the results.

Simplistic simulation can be done with spreadsheet software, while more sophisticated modelling can be done with the use of software packages, like Palisade @Risk®, Minitab®, Crystal Ball® and many others.

4.2 Monte Carlo Simulation Basics

Monte Carlo simulation can be defined as a method for iteratively evaluating a deterministic model using sets of random numbers as inputs. It is a fairly simple mathematical procedure, with random inputs and random outputs: $y = f(x_1, x_2, \ldots, x_n)$, where the input values are sampled and the output values are recorded and analysed as illustrated in Figure 4.1.

In order to run Monte Carlo simulation we need to generate random variables that follow an arbitrary statistical distribution. The inputs are randomly generated from probability distributions to simulate the process of sampling from an actual population, therefore we choose a distribution for each input that best represents our current state of knowledge. The data generated from a simulation can be represented in a basic statistic format, a histogram, fitted into a probability distribution function, or any other format needed for the analysis.

4.3 Additional Statistical Distributions

Before exploring the Monte Carlo simulation techniques we need to introduce here two additional statistical distributions, which are important to the Monte Carlo method, but were not covered in Chapter 2. Those

Practical Reliability Engineering, Fifth Edition. Patrick D. T. O'Connor and Andre Kleyner.

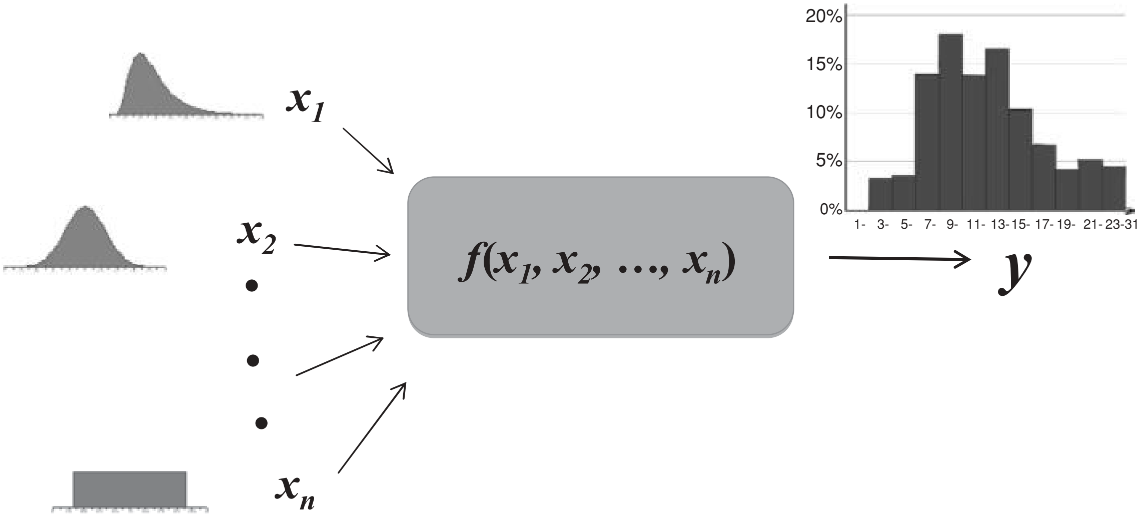


Figure 4.1 *Simplified Monte Carlo simulation procedure with $y = f(x_1, x_2, \ldots, x_n)$.*

distributions are not typically used to model failures, but are often utilised for engineering approximations and basic random number generations.

4.3.1 Uniform Distribution

The ability to generate random numbers is a key to a successful Monte Carlo simulation. The Uniform distribution holds a special place in the Monte Carlo simulation arsenal because sampling any statistical distribution typically employs the uniformly distributed random variable. The continuous uniform distribution, sometimes also known as *rectangular distribution*, is a distribution that has constant probability on the interval [*a; b*], Figure 4.2 (a) and has the pdf of

$$f(x) = \begin{cases} \dfrac{1}{(b-a)} & a \le x \le b \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

Monte Carlo software programs use various tools to generate uniformly distributed random variables. For example, Microsoft Excel has a built-in uniform distribution function =RAND(), which is the most basic form of rectangular distribution with $a = 0$ and $b = 1$. When the formula =RAND() is entered into a cell, it generates a number, that is equally likely to assume any value between 0 and 1. The ability to generate a uniformly distributed random variable on the [0; 1] interval enables the practitioner to perform a wide range of simulation tasks.

4.3.2 Triangular Distribution

The Triangular distribution is often used in engineering approximations, where a random variable is defined by the minimum, most likely and maximum values, also referred as *three-point estimator*. Values around the most likely value have higher probability of occurrence.

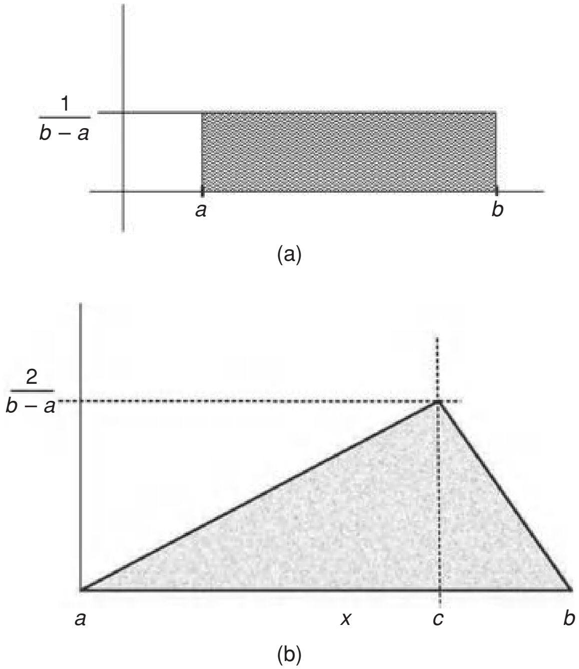


Figure 4.2 *(a) Rectangular and (b) Triangular distributions.*

The generic (asymmetric) triangular distribution has the pdf of

$$f(x) = \begin{cases} \dfrac{2(x-a)}{(b-a)(c-a)} & \text{for } a \leq x \leq c \\[2ex] \dfrac{2(b-x)}{(b-a)(b-c)} & \text{for } c \leq x \leq b \\[2ex] 0 & \text{otherwise} \end{cases} \tag{4.2}$$

and its geometric form is shown in Figure 4.2 (b).

The triangular distribution in its symmetrical form, where $c = (b - a)/2$ is listed in Table 4.1 and is often used as an engineering approximation of the normal distribution. This approximation would eliminate the effect of $x = \pm\infty$ tails of the normal pdf and would serve as a simplified form of the curtailed normal pdf (Figure 2.12, Chapter 2) in various engineering applications.

4.4 Sampling a Statistical Distribution

The Monte Carlo simulation procedure requires the capability to sample from arbitrary distributions. Once we have the ability to generate a uniformly distributed random variable on the interval [0; 1] we can extend this capability to any general form of distribution. Since the cdf of a statistical distribution $F(x)$ belongs to the same range [0; 1], for most distributions solved closed-form analytical solution for x can be found in terms of the given uniform random number. This method is called the *inverse transform sampling method*, (see Wikipedia, 2010) and is used for generating sample numbers at random from any probability distribution given its cumulative distribution function cdf (see Hazelrigg, 1996).

4.4.1 Generating Random Variables Using Excel Functions

As mentioned before, a basic spreadsheet program can be used to run a Monte Carlo simulation. In this case the generation of random variables is implemented by propagating a basic formula as many times as the number of simulation runs required by the model. For that purpose we need a capability to generate random numbers following the distributions of interest associated with the input variables. Inverse transform sampling is efficient if the cdf can be analytically or computationally inverted, which can be easily done using Excel statistical functions. Even when a distribution does not have a closed-form mathematical expression for cdf, such as the normal or lognormal distributions, it can still be resolved using the Excel inverse statistical function as shown in Table 4.1.

Table 4.1 *Statistical distributions sampling using Microsoft Excel®.*

Distribution	cdf	Excel Function
Uniform	$F(x) = \begin{cases} \dfrac{x-a}{(b-a)} & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$	=(b−a)*RAND()
Triangular (Symmetrical)	$F(x) = \begin{cases} 2\left(\dfrac{x-a}{b-a}\right)^2 & \text{for } a \le x \le \dfrac{a+b}{2} \\ 1-2\left(\dfrac{b-x}{b-a}\right)^2 & \text{for } \dfrac{a+b}{2} \le x \le b \end{cases}$	=a+(b−a)*(RAND()+RAND())/2
Normal	$F(x) = \Phi\left(\dfrac{x-\mu}{\sigma}\right)$	=NORMINV(RAND(), μ, σ)
Lognormal	$F(x) = \Phi\left(\dfrac{\ln x-\mu}{\sigma}\right)$	=LOGINV(RAND(), μ, σ)
Weibull (2 Parameter)	$F(x) = 1 - e^{-\left(\frac{x}{\eta}\right)^{\beta}}$	=(η(-LN(RAND()))^(1/β))
Weibull (3 Parameter)	$F(x) = 1 - e^{-\left(\frac{x-\gamma}{\eta}\right)^{\beta}}$	=(η(-LN(RAND()))^(1/β))+γ
Extreme Value (Minimum)	$F(x) = 1 - \exp\left\{-\exp\left[\dfrac{1}{\sigma}(x-\mu)\right]\right\}$	=μ+σ*LN(LN(1/RAND()))
Extreme Value (Maximum)	$F(x) = \exp\left\{-\exp\left[-\dfrac{1}{\sigma}(x-\mu)\right]\right\}$	=μ−σ*LN(LN(1/RAND()))

4.4.2 Number of Simulation Runs and the Accuracy of Results

There is no simple way to estimate the number of Monte Carlo simulation runs needed to achieve the required accuracy. The number of runs (also referred as trials, simulations, iterations, or sample size) depends on the complexity of the deterministic model, variance of the input and the sought accuracy of the output. High variance of the input and high complexity of the model increases the variance of the output and thus necessitates more simulation runs to achieve 'stability' of the output.

Monte Carlo simulation is a statistical measure; therefore based on the central limit theorem and the confidence bounds estimate for the normal distribution (see Chapters 2 and 3) the standard error of the distribution mean can be expressed as:

$$Er(\mu) = \frac{Z_{\alpha/2}\sigma}{\sqrt{m}} \tag{4.3}$$

where: $Er(\mu) =$ standard error of the mean.
$\alpha = 1 - C$, where C is the confidence level.
$z_{\alpha/2} =$ is the standard normal statistic (see z-value Section 2.6.1).
$\sigma =$ standard deviation of the output.
$m =$ number of Monte Carlo runs.

Eq. (4.3) can estimate the required number of runs to reach a certain level of confidence in statistics of the simulated output. This equation clearly shows that in order to reduce the error by one order of magnitude the number of runs should be two orders. However Eq. (4.3) has limited applications because the value of σ is not known and can only be assumed *a priori* or estimated after the first simulation.

Example 4.1

An electric circuit current was modelled with 1000 Monte Carlo simulation runs. The mean value of the outputs is 20 A with the standard deviation of 10 A. Estimate the number of runs required to achieve 1 % accuracy with 95 % confidence.

For this analysis we need to convert Eq. (4.3) into the percentage format by dividing both sides by μ. This turns σ into the relative standard deviation of $10/20 = 0.5$ (50 %) and the desired error of mean $Er(\mu)/\mu$ into the percentage value of 0.01 (1.0 %). For $\alpha = 1 - \mathrm{C} = 0.05$, $\mathrm{Z}_{0.025} = 1.645$ (= NORMSINV(0.95) in Excel). Therefore, based on Eq. (4.3) the number of required runs can be calculated as:

$$m = \frac{Z_{\alpha/2} \times \sigma/\mu}{Er(\mu)/\mu} = \left(\frac{1.645 \times 0.5}{0.01}\right)^2 = 6764 \; runs$$

As mentioned before, Eq. (4.3) should only be used as an approximation, therefore the number above can only be considered as a rough estimate of the required number of runs.

In order to make the simulation faster and more efficient, MC practitioners often utilize stratified sampling (as opposed to pure random sampling). One popular approach to stratified sampling is called *Latin Hypercube Sampling* (LHS). In Latin Hypercube Sampling, the range of each input variable is divided into intervals (bins) of equal probability. Then the sampling is performed according to the algorithm where each bin is sampled once before repeating. This algorithm also defines the order in which the samples from the bins are combined between the different input variables. This strategy helps to produce more evenly distributed (in probability) random values and reduce the occurrence of less likely combinations, such as those where all the input variables come from the tails of their respective distributions. Overall LHS generates a set of samples that more precisely reflect the shape of a sampled distribution and the mean of a set of simulation results more quickly approaches the 'true' value. Many commercially available Monte Carlo software packages have an option to run Latin Hypercube sampling in addition to the random sampling. Furthermore some of the commercial packages, like @Risk® can automatically determine the sufficient number of runs by tracking the convergence of the output during the simulation, see, for example Palisade (2005). For more on LHS and other methods of stratified sampling see Rubinstein and Kroese (2008) and Roberts and Casella (2004).

4.5 Basic Steps for Performing a Monte Carlo Simulation

A Monte Carlo simulation study may be divided into different steps. Those steps could vary based on the scope of the problem, but some basic steps that should be included in any analysis are outlined below:

Step 1: Define the problem and the overall objectives of the study. Evaluate the available data and outcome expectations.
Step 2: Define the system and create a parametric model, $y = f(x_1, x_2, \ldots, x_q)$.
Step 3: Design the simulation. Quantities of interest need to be collected, such as the probability distributions for each of the inputs. Define how many simulation runs should be used. The number of runs, m is affected by the complexity of the model and the sought accuracy of results (Section 4.4.2).
Step 4: Generate a set of random inputs, $x_{i1}, x_{i2}, \ldots, x_{iq}$.
Step 5: Run the deterministic system model with the set of random inputs. Evaluate the model and store the results as y_i.
Step 6: Repeat steps 4 and 5 for $i = 1$ to m.
Step 7: Analyse the results statistics, confidence intervals, histograms, best fit distribution, or any other statistical measure.

These steps are summarized and depicted in the diagram Figure 4.3.

Example 4.2 Calculating the Probability of Exceeding Yield Strength

In order to illustrate the Monte Carlo method, let us consider a simple stress analysis problem, where a random force F is applied to a rectangular area with dimensions $A \times B$. Based on the previously recorded data and the goodness of fit criteria, force F can be statistically described by the 2-parameter Weibull with $\beta = 2.5$ and $\eta = 11\,300$ N (mean value 10 026 N). Dimension A has the mean value of 2.0 cm with the tolerance of ±1.0 mm and B has the mean of 3.0 cm with the tolerance of ±1.5 mm. The structure is expected to function properly while within the elastic strain range, therefore the probability of exceeding the yield strength of 30 MPa (30×10^6 N/m^2) needs to be estimated.

Uniaxial stress can be calculated as the force divided by the area it is acting on:

$$S = \frac{F}{AB} \tag{4.4}$$

Despite the apparent simplicity of Eq. (4.4) it would be difficult to calculate analytically even the simplest statistics of the result, such as mean or standard deviation, let alone obtaining the statistical distribution of the resulting stress value. Monte Carlo simulation is perhaps the most efficient way to solve this problem.

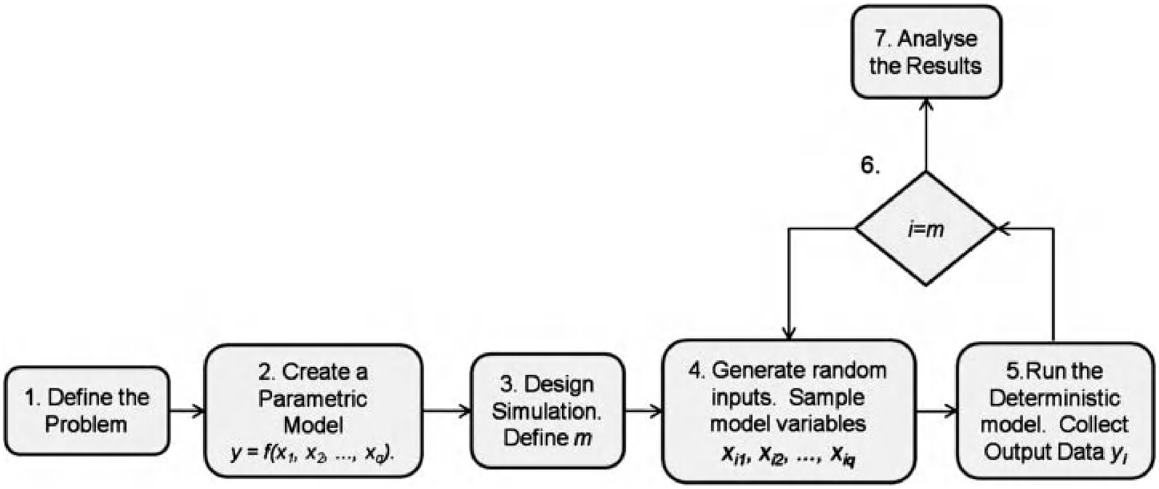

7. Analyse the Results
6.
$i=m$
1. Define the Problem
2. Create a Parametric Model $y = f(x_1, x_2, \ldots, x_q)$.
3. Design Simulation. Define m
4. Generate random inputs. Sample model variables $x_{i1}, x_{i2}, \ldots, x_{iq}$
5.Run the Deterministic model. Collect Output Data y_i

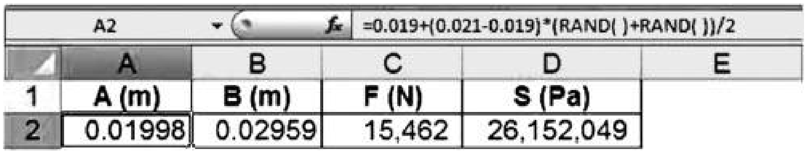
Figure 4.3 *Monte Carlo simulation process.*

Following the process described in Section 4.5, the first two steps have defined the problem and created the parametric model (see the previous paragraphs and Eq. (4.4)). Step 3 involves designs of the simulation. While the distribution for the force F is already known, A and B still need to be modelled as random variables due to the dimensional tolerances. There is a number of ways to model a tolerance, and one of them is to use a triangular distribution with the minimum and maximum values corresponding to the minus and plus tolerances. Therefore, A can be defined by the three point estimator: [0.019 (min), 0.02 (most likely), 0.021 (max)] metres and B as [0.0285, 0.03, 0.0315] metres. Considering that there are only three variables in this model, we will start with the relatively low number of iterations $m = 1000$.

Steps 4 and 5 involve using the Excel spreadsheet as a simulation tool, as shown in Figure 4.4. Let us start with entering the variables' names: A (m), B (m), F (m), and S (Pa) respectively in row 1. In row 2 we write the equations for their respective random variables.

Cell A2 : = 0.019 + (0.021 − 0.019)*(RAND() + RAND())/2 to simulate the dimension A.
Cell B2 : = 0.0285 + (0.0315 − 0.0285)*(RAND() + RAND())/2 to simulate the dimension B.
Cell C2 : = (11300*(−LN(RAND()))^(1/2.5)) to simulate the force F.
Cell D2 : = C2/(A2*B2) to simulate the resulting stress value per (4.4).

Then each of the cells A2 through D2 is copied down through row 1001. Depending on the Excel calculation settings, it may be required to hit the 'recalculate' key (often F9 in Windows applications) to generate the


A2 =0.019+(0.021-0.019)*(RAND( )+RAND( ))/2

	A	B	C	D	E
1	A (m)	B (m)	F (N)	S (Pa)	
2	0.01998	0.02959	15,462	26,152,049	

Figure 4.4 *Monte Carlo Simulation using Microsoft Excel®.*

random variables. At this point, the process of simulation (step 6) is complete and the output values are generated in column D. In order to estimate the probability of S exceeding 30 MPa we need to calculate the ratio of the cells with the stress values greater than 30 MPa ($S > 30\ 000\ 000$) to the total number of cells generated during the simulation. It can be easily done with the Excel formula:

$$= \text{COUNTIF}(\text{D2} : \text{D1001}, ^{\prime\prime} > 30, 000, 000^{\prime\prime})/\text{COUNT}(\text{D2} : \text{D1001}) \tag{4.5}$$

One of the ways to assess the sufficiency of the number of runs (number or generated rows in this example) is to repeatedly hit 'Recalculate' (typically F9) and observe the value of the ratio (4.5). In this case, this procedure produced a random sequence of numbers (4.6; 3.7; 4.1; 3.95; 4.25; 4.05; 4.3; 4.55; 4.1; 3.5; 4.1; 4.05) with the average of 4.104 %. To complete the analysis the generated output values in column D can be presented as a histogram or plotted as a cdf of the resulting distribution. The standard deviation of the data in column D can also be used to calculate the number of required simulation runs (Excel rows) per Eq. (4.3). For this purpose we would need to specify the required accuracy and the confidence level, similar to that in Example 4.1.

In order to run a more sophisticated analysis we can employ commercially available software, such as @Risk®. @Risk works off a standard Excel spreadsheet. All the random numbers are compactly generated in their respective single cells (both inputs and outputs). It provides the option of graphical representation of the input variables and the histogram generation of the output. Table 4.2 shows the input variables F, A, B graphically generated with @Risk software.

Completing this simulation with 10 000 runs took approximately 30 seconds. The output values were automatically presented as a histogram with the best fitting distribution shown in Figure 4.5. This distribution according to both Chi-square and Kolmogorov-Smirnov criteria was 3 parameter Weibull with $\beta = 2.48$, $\eta = 18.8 \times 10^6$ Pa, $\gamma = 38\ 593$ Pa). The right tail of this distribution in Figure 4.5 also shows that in 4.2 % of the cases the stress S exceeded 30 MPa.

In addition to the basic simulation, the sensitivity analysis was completed with the results shown in Figure 4.6. This analysis helps to determine how sensitive the output is to variations in each input.

Figure 4.6 shows that based on the correlation coefficients the output S is approximately eight times more sensitive to variations in force value (F) than to variation in the cross sectional dimensions. That happened due in part to a much larger variation of the random variable F than that of either A or B. Both A and B have their values restricted by the highest and lowest values, where force F can theoretically be very high due to the tail of the Weibull distribution.

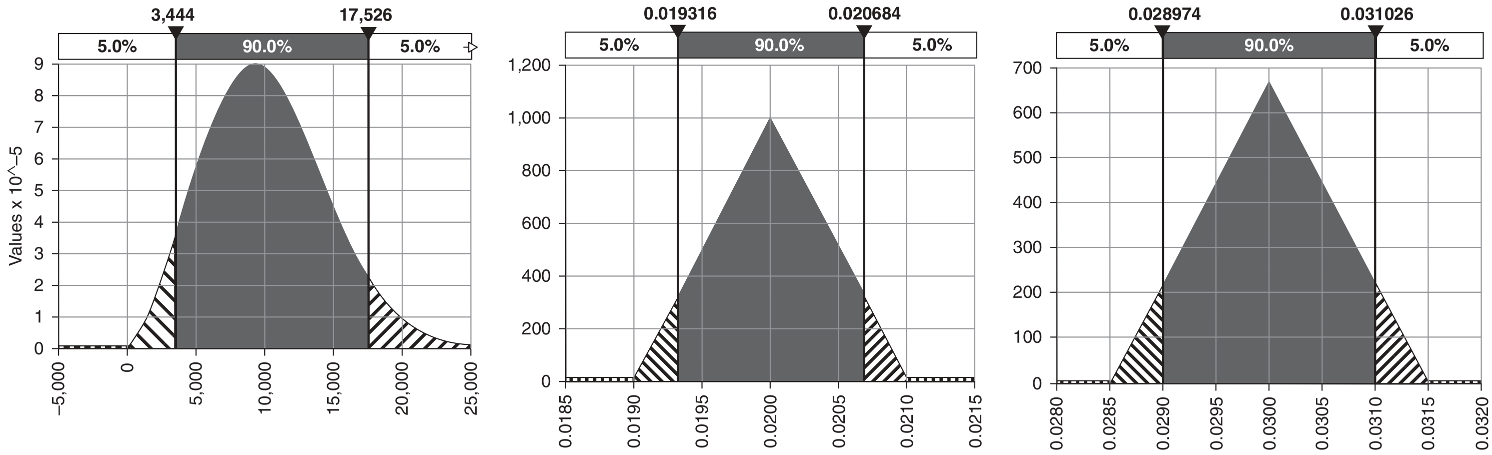
4.6 Monte Carlo Method Summary

Over the years the Monte Carlo method has proven itself as a very useful tool in a variety of applications involving uncertainty. However it is important for a practitioner to understand the advantages and disadvantages of using Monte Carlo simulation in problem solving.

The main advantage of the method is based on its low level of complexity. Compared to the other numerical methods that can solve the same problem, MC is conceptually very simple and is relatively easy to implement on a computer. It does not require specific knowledge of the form of the solution or its analytic properties. It does not constrain what form the distributions take, and the distributions need not necessarily even have a mathematical representation. The Monte Carlo method is useful for modelling phenomena with significant uncertainty in inputs and it always works regardless of the complexity of the model.

Another important advantage is the ease of comprehension by decision-makers. 'What-if' scenarios and the sensitivity of the outputs to input assumptions can be quickly analysed.

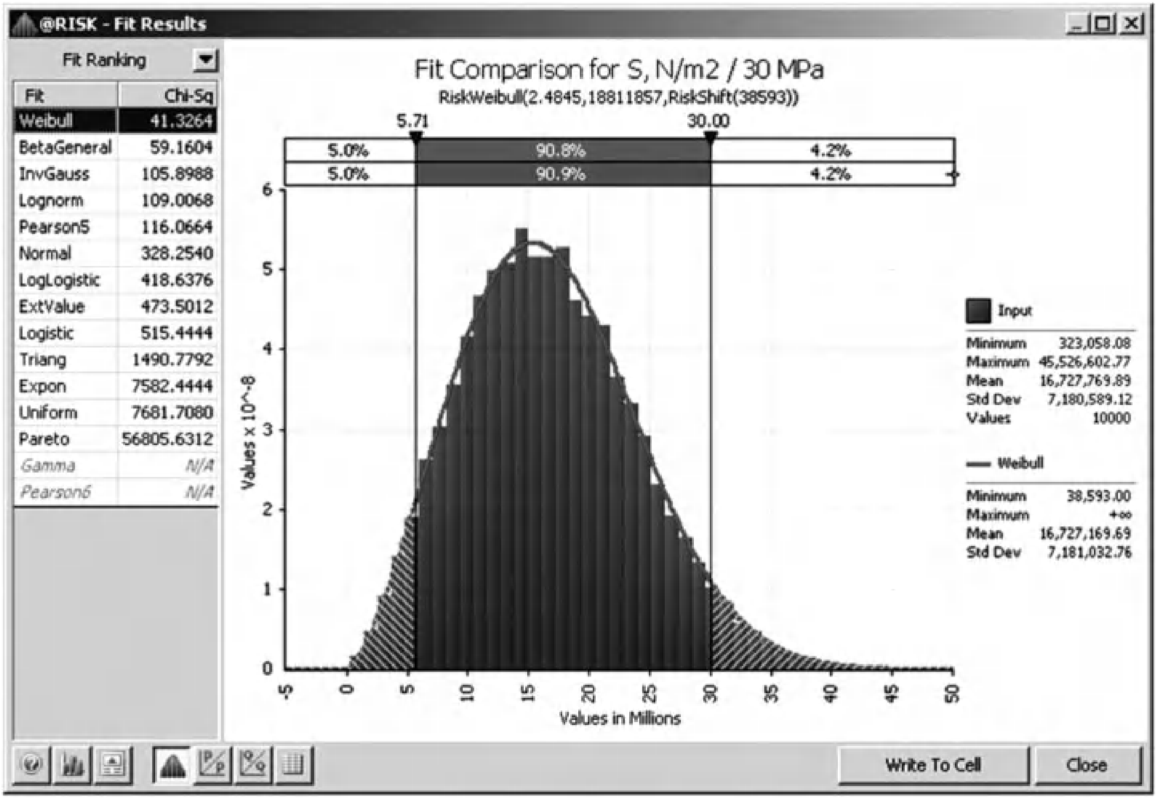
Table 4.2 *Input variables generated by @Risk® for Example 4.2 (Reproduced by permission of Palisade Corporation).*



F = RiskWeibull(2.5,11300)

A = RiskTriang(0.019,0.02,0.021)

B = RiskTriang(0.0285,0.03,0.0315)



Fit	Chi-Sq
Weibull	41.3264
BetaGeneral	59.1604
InvGauss	105.8988
Lognorm	109.0068
Pearson5	116.0664
Normal	328.2540
LogLogistic	418.6376
ExtValue	473.5012
Logistic	515.4444
Triang	1490.7792
Expon	7582.4444
Uniform	7681.7080
Pareto	56805.6312
Gamma	N/A
Pearson6	N/A

■ Input	
Minimum	323,058.08
Maximum	45,526,602.77
Mean	16,727,769.89
Std Dev	7,180,589.12
Values	10000

— Weibull	
Minimum	38,593.00
Maximum	+∞
Mean	16,727,169.69
Std Dev	7,181,032.76

Figure 4.5 *Simulation results including the histogram and the best fit distribution for Example 4.2 using @Risk v.5.7 (Reproduced by permission of Palisade Corporation).*

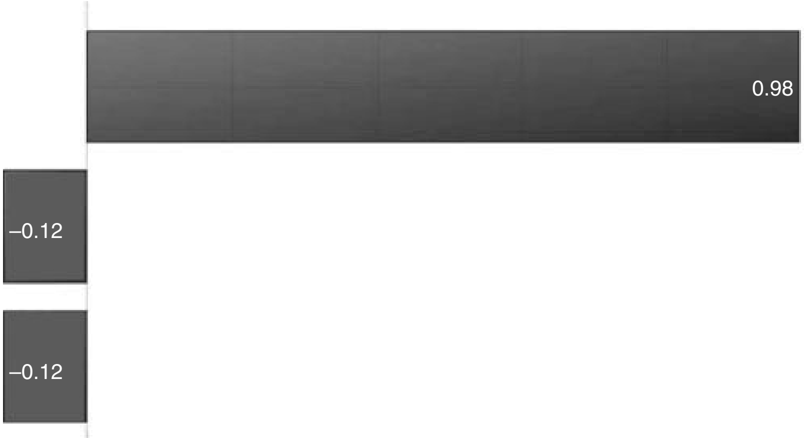


Figure 4.6 *Monte Carlo Simulation sensitivity analysis by @Risk® (Reproduced by permission of Palisade Corporation).*

The disadvantages of using Monte Carlo include computational intensity, especially with complex models requiring large numbers of simulation runs, although with growing computing power, this becomes less of a problem. The arguments against Monte Carlo also include claims that it is a 'brute force' solution heavily relying on computer power. Furthermore, it is difficult to estimate an error, since there are no hard bounds on the error of the computed result. The probabilistic error bound, which is essentially based on the variance, may not be a good measure of the error, especially for skewed distributions. Another potential drawback is that Monte Carlo implicitly assumes that all the parameters are independent, which may not be the case, especially with complex models. Correlated inputs should be identified in advance and simulated as such; otherwise the simulation may produce biased results.

Faulin *et al.* (2010) describe simulation applications to complex systems reliability and availability.

Questions

The following problems can be solved using Excel spreadsheet. A trial version of @Risk software can be downloaded from http://www.palisade.com/ or this textbook's version from http://www.palisade.com/bookdownloads/oconnorkleyner for more sophisticated analysis.

1. Program 2-paramter Weibull distribution with $\beta = 3.0$ and $\eta = 1000$ into Excel spreadsheet and generate 100 rows by copying down the equation from Table 4.1. In the next column generate 1000 rows and in the next column 10 000 rows. Calculate the mean and standard deviation for each column. By hitting 'Recalculate' (typically F9) observe the mean and standard deviation values. What can you say about the variation for each group?
2. When simulating the function $Z = X^Y$ where X and Y are random functions. If X and Y are comparable statistical distribution (e.g. both normal with $\mu = 10.0$, $\sigma = 2.0$), will the function Z be more sensitive to X or Y? Justify your answer.
3. Derive an Excel formula to simulate a non-symmetrical version of triangular distribution shown in Table 4.1.
4. Estimate top one-sided 80 % confidence on warranty claims cost of a washing machine. Warranty cost can be calculated as (Sales Volume)×C_W×(1-NFF)×[1-R(3yrs)], where C_W is the cost per warranty claim, R(3yrs) is reliability at 3 years and NFF is the percent of 'no fault found' claims. Sales volume is uniformly distributed between 800 000 and 1 million units. Cost of warranty is lognormally distributed with the parameters $\mu = 5.8$ and $\sigma = 0.5$. The washing machine has a constant failure rate, which can be between 0.001 and 0.002 failures per year (uniformly distributed). NFF can be modelled by a symmetrical triangular distribution with the minimum and maximum values of 20 and 50 %.

   If you are using @Risk or other specialized Monte Carlo simulation software, run a sensitivity analysis and determine which variable has the most impact on the total warranty cost.
5. Suppose that you have run a Monte Carlo analysis (*m* samples) and wish to cut the standard deviation in half. How many samples do you need to run?
6. Test the hypothesis that whenever several random variables are added together, the resulting sum tends to normal regardless of the distribution of the variables being added. Sample the sum of 10 random variables from different statistical distribution and test the normality of this sum by constructing the histogram or using other statistical tools.
7. An electric circuit current was modelled with 1000 experiments. The mean value of the outputs is 25 amps with the standard deviation of 8 amps. Estimate the number of runs required to achieve 1 % accuracy with 95 % confidence.

Bibliography

Faulin, J., Juan, A., Martorell, S. and Ramírez-Márquez, J.-E. (eds) (2010) *Simulation Methods for Reliability and Availability of Complex Systems*, Springer-Verlag.

Hazelrigg, G. (1996) *Systems Engineering: An Approach to Information-Based Design*, Prentice Hall.

Palisade (2005) *Guide to Using @Risk. Advanced Risk Analysis for Spreadsheets*, Palisade Corporation, Newfield, New York. Available at: http://www.palisade.com (Accessed February 2011).

Roberts, C. and Casella, G. (2004) *Monte Carlo Statistical Methods*, 2nd edn, Springer.

Rubinstein R. and Kroese D. (2008) *Simulation and the Monte Carlo Method*, 2nd edn (Wiley Series in Probability and Statistics), Wiley.

Sandborn, P. (2011) *Electronic Systems Cost Modeling – Economics of Manufacturing and Life Cycle*. *Chapter 10* (*Uncertainty Modeling*). World Scientific Publishing, Co.

Wikipedia (2010) *Inverse Transform Sampling*: Available at: http://en.wikipedia.org/wiki/Inverse_transform_sampling.

5

Load–Strength Interference

5.1 Introduction

In Chapter 1 we set out the premise that a common cause of failure results from the situation when the applied load exceeds the strength. Load and strength are considered in the widest sense. 'Load' might refer to a mechanical stress, a voltage, a cyclical load, or internally generated stresses such as temperature. 'Strength' might refer to any resisting physical property, such as hardness, strength, melting point or adhesion. Please note that the Load-Strength concept is often referred in the literature as 'Stress-Strength'.

Examples are:

1 A bearing fails when the internally generated loads (due perhaps to roughness, loss of lubricity, etc.) exceed the local strength, causing fracture, overheating or seizure.
2 A transistor gate in an integrated circuit fails when the voltage applied causes a local current density, and hence temperature rise, above the melting point of the conductor or semiconductor material.
3 A hydraulic valve fails when the seal cannot withstand the applied pressure without leaking excessively.
4 A shaft fractures when torque exceeds strength.
5 Solder joints inside a vehicle radio develop cracks before the intended service life due to temperature cycling fatigue caused by the internal heating.

Therefore, if we design so that strength exceeds load, we should not have failures. This is the normal approach to design, in which the designer considers the likely extreme values of load and strength, and ensures that an adequate safety factor is provided.

Additional factors of safety may be applied, for example as defined in pressure vessel design codes or electronic component derating rules. This approach is usually effective. Nevertheless, some failures do occur which can be represented by the load-strength model. By our definition, either the load was then too high or the strength too low. Since load and strength were considered in the design, what went wrong?

5.2 Distributed Load and Strength

For most products neither load nor strength are fixed, but are distributed statistically. This is shown in Figure 5.1 (a). Each distribution has a mean value, denoted by $\bar{L}$ or $\bar{S}$, and a standard deviation, denoted by

Practical Reliability Engineering, Fifth Edition. Patrick D. T. O'Connor and Andre Kleyner.

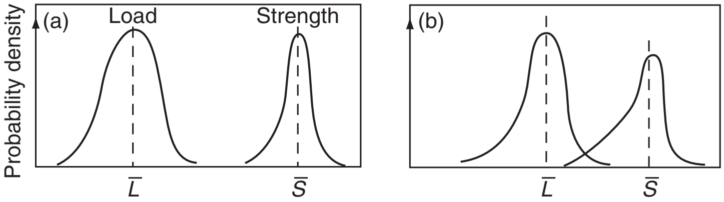


Figure 5.1 *Distributed load and strength: (a) non-overlapping distributions, (b) overlapping distributions.*

σ_L or σ_S. If an event occurs in which the two distributions overlap, that is an item at the extreme weak end of the strength distribution is subjected to a load at the extreme high end of the load distribution, such that the 'tails' of the distributions overlap, failure will occur. This situation is shown in Figure 5.1 (b).

For distributed load and strength, we define two factors, the *safety margin* (SM),

$$\text{SM} = \frac{\bar{S} - \bar{L}}{\left(\sigma_S^2 + \sigma_L^2\right)^{1/2}} \tag{5.1}$$

and the loading roughness (LR),

$$\text{LR} = \frac{\sigma_L}{\left(\sigma_S^2 + \sigma_L^2\right)^{1/2}} \tag{5.2}$$

The safety margin is the relative separation of the mean values of load and strength and the loading roughness is the standard deviation of the load; both are relative to the combined standard deviation of the load and strength distributions.

The safety margin and loading roughness allow us, in theory, to analyse the way in which load and strength distributions interfere, and so generate a probability of failure. By contrast, a traditional deterministic safety factor, based upon mean or maximum/minimum values, does not allow a reliability estimate to be made. On the other hand, good data on load and strength properties are very often not available. Other practical difficulties arise in applying the theory, and engineers must always be alert to the fact that people, materials and the environment will not necessarily be constrained to the statistical models being used. The rest of this chapter will describe the theoretical basis of load-strength interference analysis. The theory must be applied with care and with full awareness of the practical limitations. These are discussed later.

Some examples of different safety margin/loading roughness situations are shown in Figure 5.2. Figure 5.2 (a) shows a highly reliable situation: narrow distributions of load and strength, low loading roughness and a large safety margin. If we can control the spread of strength and load, and provide such a high safety margin, the design should be intrinsically failure-free. (Note that we are considering situations where the mean strength remains constant, i.e. there is no strength degradation with time. We will cover strength degradation later.) This is the concept applied in most designs, particularly of critical components such as civil engineering structures and pressure vessels. We apply a safety margin which experience shows to be adequate; we control quality, dimensions, and so on, to limit the strength variations, and the load variation is either naturally or artificially constrained.

Figure 5.2 (b) shows a situation where loading roughness is low, but due to a large standard deviation of the strength distribution the safety margin is low. Extreme load events will cause failure of weak items. However, only a small proportion of items will fail when subjected to extreme loads. This is typical of a situation where

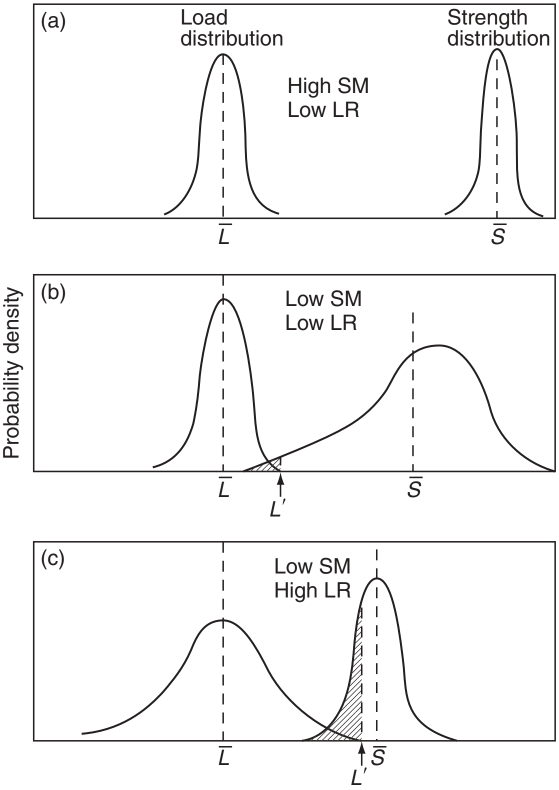


Figure 5.2 *Effect of safety margin and loading roughness. Load L′ causes failure of a proportion of items indicated by the shaded area.*

quality control methods cannot conveniently reduce the standard deviation of the strength distribution (e.g. in electronic device manufacture, where 100 % visual and mechanical inspection is seldom feasible). In this case deliberate overstress can be applied to cause weak items to fail, thus leaving a population with a strength distribution which is truncated to the left (Figure 5.3). The overlap is thus eliminated and the reliability of the surviving population is increased. This is the justification for high stress burn-in of electronic devices,

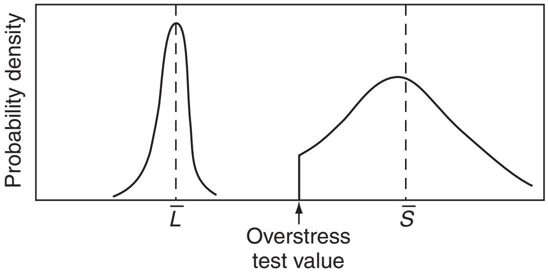


Figure 5.3 *Truncation of strength distribution by screening.*

proof-testing of pressure vessels, and so on. Note that the overstress test not only destroys weak items, it may also cause weakening (strength degradation) of good items. Therefore the burn-in test should only be applied after careful engineering and cost analysis.

Figure 5.2 (c) shows a low safety margin and high loading roughness due to a wide spread of the load distribution. This is a difficult situation from the reliability point of view, since an extreme stress event could cause a large proportion of the population to fail. Therefore, it is not economical to improve population reliability by screening out items likely to fail at these stresses. The options left are to increase the safety margin by increasing the mean strength, which might be expensive, or to devise means to curtail the load distribution. This is achieved in practice by devices such as current limiters and fuses in electronic circuits or pressure relief valves and dampers in pneumatic and hydraulic systems.

5.3 Analysis of Load–Strength Interference

The reliability of a part, for a discrete load application, is the probability that the strength exceeds the load:

$$\begin{aligned} R &= P(S > L) \\ &= \int_0^\infty \mathrm{f_L}(L) \left[\int_{\mathrm{L}}^\infty \mathrm{f_S}(S)\, \mathrm{d}S \right] \mathrm{d}L \\ &= \int_0^\infty \mathrm{f_S}(S) \left[\int_0^S \mathrm{f_L}(L)\, \mathrm{d}L \right] \mathrm{d}S \end{aligned} \tag{5.3}$$

where $\mathrm{f_S}(S)$ is the pdf of strength and $\mathrm{f_L}(L)$ is the pdf of load.

Also, if we define $y = S - L$, where y is a random variable such that

$$\begin{aligned} R &= P(y > 0) \\ &= \int_0^\infty \int_0^\infty \mathrm{f_S}(y + L)\, \mathrm{f}\,(L)\mathrm{d}L\, \mathrm{d}y \end{aligned} \tag{5.4}$$

5.3.1 Normally Distributed Strength and Load

If we consider normally distributed strength and load, so that the cdfs are

$$\mathrm{F_L}(L) = \Phi\left(\frac{L - \bar{L}}{\sigma_\mathrm{L}}\right)$$

$$\mathrm{F_S}(S) = \Phi\left(\frac{S - \bar{S}}{\sigma_\mathrm{S}}\right)$$

if $y = S - L$, then $\bar{y} = \bar{S} - \bar{L}$ and $\sigma_y = (\sigma_\mathrm{S}^2 + \sigma_\mathrm{L}^2)^{1/2}$. So

$$\begin{aligned} R &= P(y > 0) \\ &= \Phi\left(\frac{S - \bar{L}}{\sigma_y}\right) \end{aligned} \tag{5.5}$$

Therefore, the reliability can be determined by finding the value of the standard cumulative normal variate from the normal distribution tables or statistical calculators. The reliability can be expressed as

$$\begin{aligned} R &= \Phi\left[\frac{\bar{S}-\bar{L}}{\left(\sigma_S^2+\sigma_L^2\right)^{1/2}}\right] \\ &= \Phi(\mathrm{SM}) \text{ (from Eq. 5.1)} \end{aligned} \tag{5.6}$$

Example 5.1

A component has a strength which is normally distributed, with a mean value of 5000 N and a standard deviation of 400 N. The load it has to withstand is also normally distributed, with a mean value of 3500 N and a standard deviation of 400 N. What is the reliability per load application?

The safety margin is

$$\frac{5000-3500}{(400^2+400^2)^{1/2}} = 2.65$$

From Appendix 1,

$$\Phi(2.65) = 0.996 \text{ or using Microsoft Excel}^{®} \text{ (=NORMSDIST(2.65))}$$

5.3.2 Other Distributions of Load and Strength

The integrals for other distributions of load and strength can be derived in a similar way. For example, we may need to evaluate the reliability of an item whose strength is Weibull distributed, when subjected to loads that are extreme-value distributed. These integrals are somewhat complex and most of the time cannot be solved analytically. Therefore, Monte Carlo simulation, covered in Chapter 4, can be used to randomly select a sample from each distribution and compare them. After a sufficient number of runs, the probability of failure can be estimated from the results, which is demonstrated later in this chapter.

5.4 Effect of Safety Margin and Loading Roughness on Reliability (Multiple Load Applications)

For multiple load applications:

$$R = \int_0^{\infty} \mathrm{f_S}(S)\left[\int_0^{s} \mathrm{f_L}(L)\mathrm{d}L\right]^n \mathrm{d}S$$

where n is the number of load applications.

Reliability now becomes a function of safety margin and loading roughness, and not just of safety margin. This complex integral cannot be reduced to a formula as Eq. (5.6), but can be evaluated using computerized numerical methods.

Figure 5.4 shows the effects of different values of safety margin and loading roughness on the failure probability per load application for large values of n, when both load and strength are normally distributed. The dotted line shows the single load application case (from Eq. 5.6). Note that the single load case is less reliable per load application than is the multiple load case.

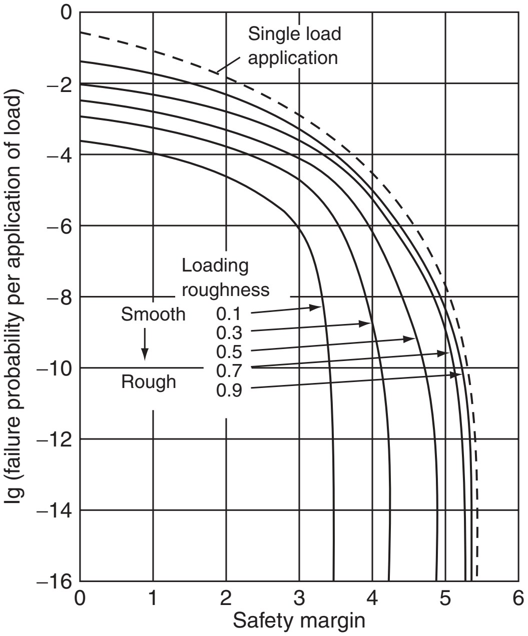


Figure 5.4 *Failure probability–safety margin curves when both load and strength are normally distributed (for large n and n = 1) (Carter, 1997).*

Since the load applications are independent, reliability over n load applications is given by

$$R = (1 - p)^n \quad \text{(from Eq. 2.2)}$$

where p is the probability of failure per load application.

For small values of p the binomial approximation allows us to simplify this to

$$R \approx 1 - np \tag{5.7}$$

The reliability for multiple load applications can then be derived, if we know the number of applications, having used Figure 5.4 to derive the value of p. Once the safety margin exceeds a value of 3 to 5, depending upon the value of loading roughness, the failure probability becomes infinitesimal. The item can then be said to be *intrinsically reliable*. There is an intermediate region in which failure probability is very sensitive to changes in loading roughness or safety margin, whilst at low safety margins the failure probability is high. Figure 5.5 shows these characteristic regions. Similar curves can be derived for other distributions of load and strength. Figure 5.6 and Figure 5.7 show the failure probability – safety margin curves for smooth and rough loading situations for Weibull-distributed load and strength. These show that if the distributions are skewed so that there is considerable interference, high safety margins are necessary for high reliability. For example, Figure 5.6 shows that, even for a low loading roughness of 0.3, a safety margin of at least 5.5 is required to ensure intrinsic reliability, when we have a right-skewed load distribution and a left-skewed strength distribution. If the loading roughness is high (Figure 5.7), the safety margin required is 8.

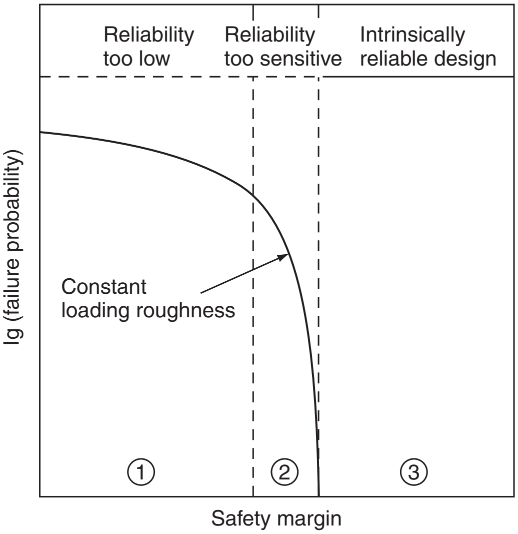


Figure 5.5 *Characteristic regions of a typical failure probability–safety margin curve (Carter, 1997).*

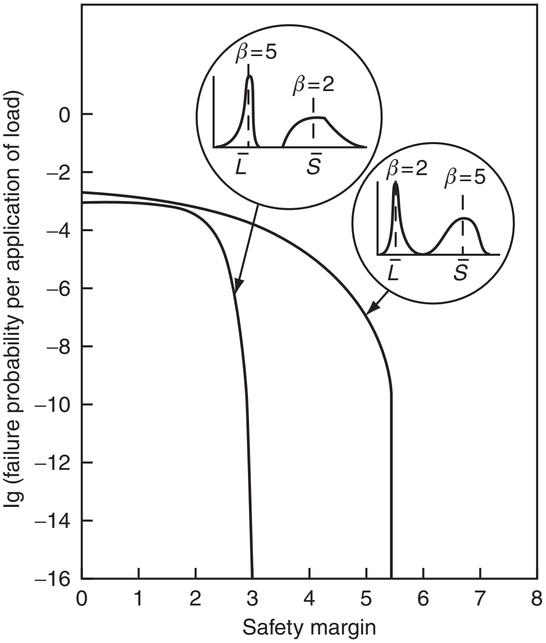


Figure 5.6 *Failure probability–safety margin curves for asymmetric distributions (loading roughness = 0.3) (Carter, 1997).*

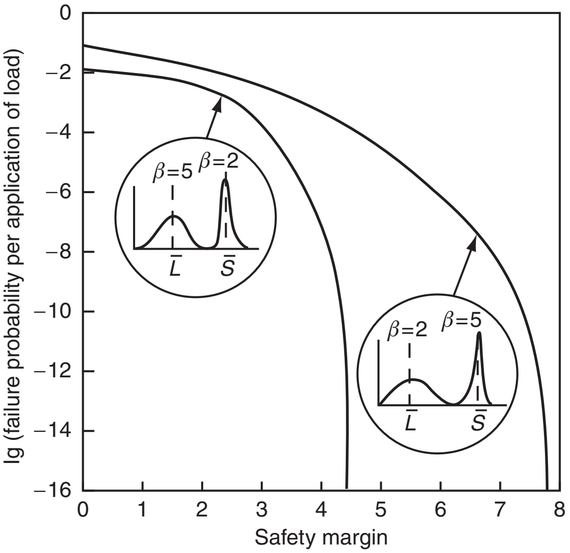


Figure 5.7 Failure probability–safety margin curves for asymmetric distributions (loading roughness = 0.9) (Carter, 1997).

These curves illustrate the sensitivity of reliability to safety margin, loading roughness and the load and strength distributions.

Two examples of Load-Strength analysis application to design are given to illustrate the application to electronic and mechanical engineering.

Example 5.2 (electronic)

A design of a power amplifier uses a single transistor in the output. It is required to provide an intrinsically reliable design, but in order to reduce the number of component types in the system the choice of transistor types is limited. The amplifier must operate reliably at 50 °C.

An analysis of the load demand on the amplifier based on customer usage gives the results in Figure 5.8. The mean ranking of the load test data is given in Table 5.1. A type 2N2904 transistor is selected. For this device the maximum rated power dissipation is 0.6 W at 25 °C.

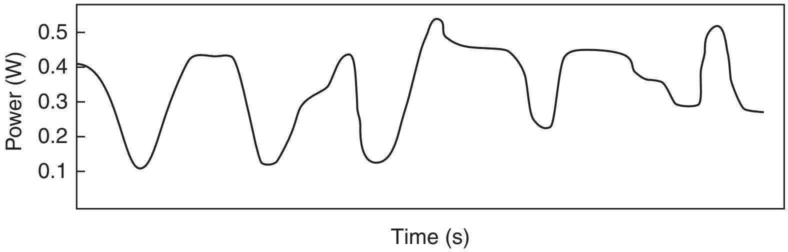


Figure 5.8 Load data (sampled at 10 s intervals).

Table 5.1 Mean ranking of load test data.

Power (W)	Cumulative percentage time (cdf)
0.1	5 %
0.2	25 %
0.3	80 %
0.4	98.5 %
0.5	99.95 %

The load test data shown in Table 5.1 can be analysed to find the best fitting distribution. There is a variety of commercial software packages with best fit capability including Weibull++® and @Risk® mentioned in the previous chapters. After applying the Weibull++ Distribution Wizard® program (see Chapter 3, Figure 3.17) we can see that the top three choices include Weibull, Normal and Gamma in that order. To simplify the calculations, let us select the normal, which has the parameters for the load distribution:

$$\bar{L} = 0.238 \text{ W}$$
$$\sigma_{\text{L}} = 0.0783 \text{ W}$$

Alternatively, this could have been obtained by plotting the data from Table 5.1 on a normal distribution paper.

However, in this case we must consider the combined effects of power dissipation and elevated temperature. The temperature derating guidelines for the 2N2904 transistor advise 3.43 mW/K linear derating. Since we require the amplifier to operate at 50 °C, the equivalent combined load distribution is normal, with the same SD, but with a mean which is $(25 \times 3.43)\,\text{mW} = 0.086\,\text{W}$ higher. The mean load $\bar{L}$ is now $0.238 + 0.086 = 0.324$ W, with an unchanged standard deviation of 0.0783 W.

To derive the strength distribution, 100 transistors were tested at 25 °C ambient, for 10 s at each power level (step stress), giving failure data as shown in Table 5.2. Similarly to the load in Table 5.1 these data

Table 5.2 Failure data for 100 transistors.

Power (W)	Number failed	Cumulative percentage failure (cdf) (mean ranking)
0.1	0	0
0.2	0	0
0.3	0	0
0.4	0	0
0.5	0	0
0.6	0	0
0.7	2	2
0.8	8	10
0.9	17	25
1.0	35	59
1.2	30	89
1.3	10	99

are processed with Weibull++, indicating a normal distribution with mean power at failure (strength) of 0.9897 W and SD of 0.142 W.

Combining the load–strength data gives

$$\mathrm{LR} = \frac{\sigma_{\mathrm{L}}}{\left(\sigma_{\mathrm{S}}^2 + \sigma_{\mathrm{L}}^2\right)^{1/2}} = \frac{0.0783}{(0.142^2 + 0.0783^2)^{1/2}} = 0.483$$

$$\mathrm{SM} = \frac{\bar{S} - \bar{L}}{\left(\sigma_{\mathrm{S}}^2 + \sigma_{\mathrm{L}}^2\right)^{1/2}} = \frac{0.989 - 0.324}{(0.142^2 + 0.0783^2)^{1/2}} = 4.10$$

Therefore

$$R = \Phi(\mathrm{SM}) = 0.9999794 = \mathrm{NORMSDIST}(4.10)$$

This is the reliability per application of load for a single load application. Figure 5.4 shows that for multiple load applications (large n), the failure probability per load application (p) is about 10^{-11} (zone 2 of Figure 5.5). Over 10^6 load applications the reliability would be about 0.999 99 (from Eq. 5.6).

In practice, in a case of this type the temperature and load derating guidelines described in Chapter 9 would normally be used. In fact, to use a transistor at very nearly its maximum temperature and load rating (in this case the measured highest load is 0.5 W at 25 °C, equivalent to nearly 0.6 W at 50 °C) is not good design practice, and derating factors of 0.5 to 0.8 are typical for transistor applications. The example illustrates the importance of adequate derating for a typical electronic component. The approach to this problem can also be criticised on the grounds that:

1 The failure (strength) data are sparse at the 'weak' end of the distribution. It is likely that batch-to-batch differences would be more important than the test data shown, and screening could be applied to eliminate weak devices from the population.
2 The extrapolation of the load distribution beyond the 0.5 W recorded peak level is dangerous, and this extrapolation would need to be tempered by engineering judgement and knowledge of the application.

Example 5.3 (mechanical fatigue)

Customer usage data often come handy in obtaining the stress distribution of a load applied to the system. A washing machine manufacturer is trying to estimate the electric motor warranty cost due to fatigue failure for the first year of operation, which on average amounts to 100 cycles for the motor. The actual washing load sizes for the motors will vary depending on the way that the user runs each machine. In order to calculate warranty cost the manufacturer wants to estimate the percentage of returns that can be expected during the first year of operation. Even though the motor was designed to operate at the stresses exceeding the maximum allowable load of 6 kg, the life of the motor is clearly dependent on the applied load. That load varies based on customer usage patterns, so the question to be answered is which load size should be used in predicting the percentage of returns during warranty. At first step, the manufacturer decides to obtain the customer usage information by conducting a survey on a representative sample of customers and recording the sizes of the loads that they placed into their washing machines.

From this data set in Table 5.3 the distribution that gives the percentage of users operating washers at different loads can be determined. The cdf values in the third column can be processed similarly to the way it

Table 5.3 Maximum loads vs. percentages of the users applying those loads.

Maximum Load (kg)	Percent of users	Cumulative percent (cdf)
2	4 %	4 %
3	42 %	46 %
4	40 %	86 %
5	12 %	98 %
6	1.7 %	99.7 %

is done in Example 5.2. Using Weibull++ Distribution Wizard® (Section 3.7.1) it was determined that this data is best fitted with the lognormal distribution with $\mu = 1.12$ and $\sigma = 0.243$.

At the next step the manufacturer needs to obtain information on the life of the motor at different loads (or stress levels). Representative samples of the motor were tested to failure at five different loads. Then Weibull analysis (Section 3.4) was performed and the percentages of a population failing at 100 cycles at each load were determined and summarised in Table 5.4.

With the help of software the three parameter Weibull distribution was fitted to this data set and the following parameters were obtained: $\beta = 1.69$, $\eta = 6.67$, $\gamma = 3.2$.

Plotting two pdf functions in the same graph produces the diagram Figure 5.9 showing the overlapping area, where failures are expected to occur.

Now the last step is to calculate the area where the two distributions overlap. Since neither of the obtained distributions is normal (thus no closed form solution) the manufacturer decided to use Monte Carlo simulation to find the percentage of the cases where values from the load distribution exceed the value from the strength distribution. As mentioned in Chapter 4 there is a variety of commercial packages designed to run Monte Carlo simulations including the 'Stress-Strength' option imbedded in Weibull++®. However, for simplicity the Excel spreadsheet solution similar to that in Chapter 4, Example 4.2 has been used. The lognormal load distribution can be simulated by the Excel® function = LOGINV(RAND(),1.12, 0.243) (see Chapter 4, Table 4.1) and the 3-parameter Weibull Strength by = (6.67*(-LN(RAND()))^(1/1.69)) + 3.2. The manufacturer has conducted 10 000 simulation runs (Excel rows) and found that the number of rows where the load exceeded the strength was on average 1.2 % of the total number of rows (10 000). Based on this analysis the manufacturer set aside the amount sufficient to cover 1.2 % motor failures of the production volume during the warranty period.

Table 5.4 Washing machine loads vs. percent of motors failing at 100 cycles.

Maximum Load (kg)	Percent failed at 100 cycles (cdf)
4	2.73 %
5	10.32 %
6	20.6 %
7	32.0 %
9	54.6 %

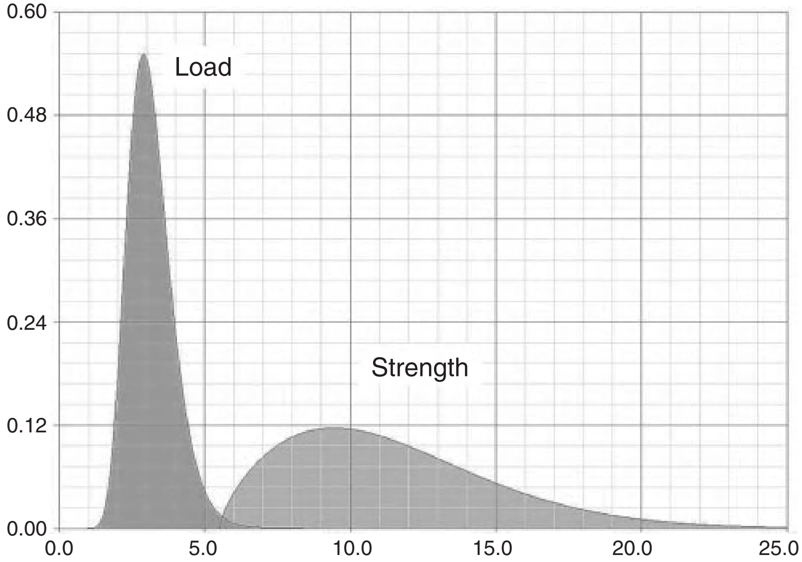


Figure 5.9 *Load-Strength distribution chart generated with Weibull++® for Example 5.3 (Reproduced by permission of ReliaSoft).*

5.5 Practical Aspects

The examples illustrate some of the advantages and limitations of the statistical engineering approach to design. The main difficulty is that, in attempting to take account of variability, we are introducing assumptions that might not be tenable, for example by extrapolating the load and strength data to the very low probability tails of the assumed population distributions. We must therefore use engineering knowledge to support the analysis, and use the statistical approach to cater for engineering uncertainty, or when we have good statistical data. For example, in many mechanical engineering applications good data exist or can be obtained on load distributions, such as wind loads on structures, gust loads on aircraft or the loads on automotive suspension components. We will call such loading situations ‘predictable’.

On the other hand, some loading situations are much more uncertain, particularly when they can vary markedly between applications. Electronic circuits subject to transient overload due to the use of faulty procedures or because of the failure of a protective system, or a motor bearing used in a hand power drill, represent cases in which the high extremes of the load distribution can be very uncertain. The distribution may be multimodal, with high loads showing peaks, for instance when there is resonance. We will call this loading situation ‘unpredictable’. Obviously it will not always be easy to make a definite classification; for example, we can make an unpredictable load distribution predictable if we can collect sufficient data. The methods described above are meaningful if applied in predictable loading situations. (Strength distributions are more often predictable, unless there is progressive strength reduction, which we will cover later.) However, if the loading is very unpredictable the probability estimates will be very uncertain. When loading is unpredictable we must revert to traditional methods. This does not mean that we cannot achieve high reliability in this way. However, evolving a reliable design is likely to be more expensive, since it is necessary either to deliberately overdesign or to improve the design in the light of experience. The traditional safety factors derived as a result of this experience ensure that a new design will be reliable, provided that the new application does not represent too far an extrapolation.

Moreover, the customer usage data utilized in both examples sometimes requires an additional layer of statistical treatment. Specifically, in Example 5.3 the manufacturer assumed the average of 100 motor cycles for the first year of operation. In reality, the number of cycles in the first year of motor operation should also be described by a statistical distribution due to differences in washing habits of the end users. This aspect of usage data will be discussed later in Chapter 7 covering variations in product usage and distribution environment.

Alternatively, instead of considering the distributions of load and strength, we can use discrete maximum/minimum values in appropriate cases. For example, we can use a simple lowest strength value if this can be assured by quality control. In the case of Example 5.2, we could have decided that in practice the transistors would not fail if the power is below 0.7 W, which would have given a safety factor of 1.4 above the 99.95 % load cdf In other cases we might also assume that for practical purposes the load is curtailed, as in situations where the load is applied by a system with an upper limit of power, such as a hydraulic ram or a human operator. If the load and strength distributions are both curtailed, the traditional safety factor approach is adequate, provided that other constraints such as weight or cost do not make a higher risk design necessary.

The statistical engineering approach can lead to overdesign if it is applied without regard to real curtailment of the distributions. Conversely, traditional deterministic safety factor approaches can result in overdesign when weight or cost reduction must take priority.

In many cases, other design requirements (such as for stiffness) provide intrinsic reliability. The techniques described above should therefore be used when it is necessary to assess the risk of failure in marginal or critical applications.

As the examples in this chapter show, the ability to fit data into a distribution is important to a successful load-strength analysis. As mentioned before there is a variety of commercial software packages (Weibull++®, @Risk®, Minitab®, Crystal Ball®, etc.) with capability of finding the best fit distribution for sampled and/or cumulative percentage data.

In this chapter we have taken no account of the possibility of strength reduction with time or with cyclic loading. The methods described above are only relevant when we can ignore strength reduction, for instance if the item is to be operated well within the safe fatigue life or if no weakening is expected to occur. Reliability and life analysis in the presence of strength degradation is covered in Chapters 8 and 14.

Finally, it is important that reliability estimates that are made using these methods are treated only as very rough, order-of-magnitude figures.

Questions

1. Describe the nature of the load and strength distributions in four practical engineering situations (use sketches to show the shapes and locations of the distributions). Comment on each situation in relation to the predictability of failures and reliability, and in relation to the methods that can be used to reduce the probabilities of failure.
2. a Give the formulae for safety margin and loading roughness in situations where the load applied to an item and the strength of the item are assumed to be normally distributed.

   b Sketch the relationship between failure probability and safety margin for different values of loading roughness, indicating approximate values for the parameters.
3. a If loads are applied randomly to randomly selected items, when both the loads and strengths are normally distributed, what is the expression for the reliability per load application?

   b Describe and comment on the factors that influence the accuracy of reliability predictions made using this approach.

4. Describe two examples each from mechanical and electronic engineering by which extreme load and strength values are curtailed, in practical engineering design and manufacture.
5. If the tests described in Example 5.2 were repeated, and
   a One transistor failed at 0.5 W, how would you re-interpret the results?
   b The first 10 failures occurred at 0.8 W, how would you re-interpret the results?
6. Calculate the reliability (Eq. (5.3)) in the case where both random load and random strength are distributed exponentially. The pdfs are given by:

$$f_S(x) = \frac{1}{\mu_S} e^{-\frac{x}{\mu_S}} \quad \text{and} \quad f_L(x) = \frac{1}{\mu_L} e^{-\frac{x}{\mu_L}}$$

   where μ_L is the mean load and μ_S is the mean strength.
7. Electrolytic capacitor leads are designed to withstand a repetitive load applied to a circuit board mounted on a moving platform. The lead's yield stress is normally distributed with the mean of 100 MPa and the standard deviation of 20 MPa. The stress generated as a result of the repetitive load during the capacitor's life time is also normally distributed with the mean of 60 MPa and the standard deviation of 15 MPa:
   a Calculate the safety margin, SM.
   b Calculate the loading roughness, LR.
   c Calculate the expected reliability of the capacitor.
8. A connecting rod must transmit a tension load which trials show to be lognormally distributed, with the following parameters: $\mu = 9.2$ and $\sigma = 1.1$. Tests on the material to be used show a lognormal strength distribution as follows: $\mu = 11.8$ and $\sigma = 1.3$. A large number of components are to be manufactured, so it is not feasible to test each one. Calculate the expected reliability of the component using Monte Carlo simulation.

Bibliography

Carter, A. (1997) *Mechanical Reliability and Design*, Wiley.

Kapur K. and Lamberson L. (1977) *Reliability in Engineering Design*, Wiley.

Palisade Corporation (2005) Guide to Using @Risk. Advanced Risk Analysis for Spreadsheets' Palisade Corporation. Newfield, New York. Available at http://www.palisade.com.

ReliaSoft (2002) Prediction Warranty Returns Based on Customer Usage Data. Reliability Edge, Quarter 1: Volume 3, Issue 1. Available at http://www.reliasoft.com/newsletter/1q2002/usage.htm (Accessed 22 March 2011).

ReliaSoft (2008) *Weibull++® User's Guide*, ReliaSoft Publishing.

Wasserman, G. (2003) *Reliability Verification, Testing and Analysis in Engineering Design*, Marcel Dekker.

6
Reliability Prediction and Modelling

6.1 Introduction

An accurate prediction of the reliability of a new product, before it is manufactured or marketed, is obviously highly desirable. Depending upon the product and its market, advance knowledge of reliability would allow accurate forecasts to be made of support costs, spares requirements, warranty costs, marketability, and so on. However, a reliability prediction can rarely be made with high accuracy or confidence. Nevertheless, even a tentative estimate can provide a basis for forecasting of dependent factors such as life cycle costs. Reliability prediction can also be valuable as part of the study and design processes, for comparing options and for highlighting critical reliability features of designs.

If a new engineered system is being planned, which will supersede an existing system, and the reliability of the existing system is known, then its reliability could reasonably be used as a starting point for predicting the likely reliability of the new system. However, the changes that will be introduced in the new system will be likely to affect its reliability: for example, more functions might be controlled through software, novel subsystems or components might be included, and so on. Some of these changes could enhance reliability, others might introduce new risks. There will also be programme and management aspects that would influence reliability, such as commitment and resources applied to achievement of quality and reliability objectives, testing strategies and constraints, in particular the time available for development. These aspects will be discussed in later chapters. This top-down approach can be applied to any new product or system. Even if there is no comparable product already in service, an estimate can and should be made, based on risks and commitment.

The prediction begins at the level of the overall system and as the system becomes more closely defined it can be extended to more detailed levels. Eventually, in principle, it is necessary to consider the reliability contributions of individual parts. However, the lower the level of analysis, the greater is the potential uncertainty inherent in predicting reliability of the whole system. It is important to remember that many system failures are not caused by failures of parts and not all part failures cause system failures.

The common approach to predicting reliability is to estimate the contributions of each part, and work upwards to the overall product or system level. The ‘parts count’ method (see later in this chapter) is widely used, but it is very dependent upon the availability of credible data. Databases providing failure rates at the part level have been developed and published, for electronic and non-electronic parts, which will be discussed later in this chapter. Since reliability is also affected strongly by factors such as knowledge and motivation of

Practical Reliability Engineering, Fifth Edition. Patrick D. T. O’Connor and Andre Kleyner.

design and test engineers, the amount and quality of testing, action on failures discovered during test, quality of production, and, when applicable, maintenance skills, these factors must be taken into account as well. In many cases they can be much more significant than past data. Therefore reliability databases must always be treated with caution as a basis for predicting the reliability of new systems. There is no intrinsic limit to the reliability that can be achieved, but the database approach to prediction can imply that there is.

6.2 Fundamental Limitations of Reliability Prediction

In engineering and science we use mathematical models for prediction. For example, the power consumption of a new electronic system can be predicted using Ohm's law and the model power = current × emf. Likewise, we can predict future planetary positions using Newton's laws and our knowledge of the present positions, velocities and masses. These laws are valid within the appropriate domain (e.g. Ohm's law does not hold at temperatures near absolute zero; Newton's laws are not valid at the subatomic level). However, for practical, everyday purposes such deterministic laws serve our purposes well, and we use them to make predictions, taking due account of such practical aspects as measurement errors in initial conditions.

Whilst most laws in physics, for practical predictive purposes, can be considered to be deterministic, the underlying mechanisms can be stochastic. For example, the pressure exerted by a gas in an enclosure is a function of the random motions of very large numbers of molecules. The statistical central limiting theorem, applied to such a vast number of separate random events and interactions, enables us to use the average effect of the molecular kinetic energy to predict the value we call pressure. Thus Boyle's law is really empirical, as are other 'deterministic' physical laws such as Ohm's law. It is only at the level of individual or very few actions and interactions, such as in nuclear physics experiments, that physicists find it necessary to take account of uncertainty due to the stochastic nature of the underlying processes. However, for practical purposes we ignore the infinitesimal variations, particularly as they are often not even measurable, in the same way as we accept the Newtonian view.

For a mathematical model to be accepted as a basis for scientific prediction, it must be based upon a theory which explains the relationship. It is also necessary for the model to be based upon unambiguous definitions of the parameters used. Finally, scientists, and therefore engineers, expect the predictions made using the models to be always repeatable. If a model used in science is found not to predict correctly an outcome under certain circumstances this is taken as evidence that the model, and the underlying theory, needs to be revised, and a new theory is postulated.

The concept of deriving mathematical models which could be used to predict reliability, in the same way as models are developed and used in other scientific and engineering fields, is intuitively appealing, and has attracted much attention. Laws of physics are taken into account in some reliability prediction models; however there are many more factors causing parts to fail (some of them are unknown), therefore predictions in reliability typically have a higher degree of uncertainty. For example, failure rate models have been derived for electronic components, based upon parameters such as operating temperature and other stresses. These are described below and in Chapters 9 and 13. Similar models have been derived for non-electronic components, and even for computer software. Sometimes these models are as simple as a single fixed value for failure rate or reliability, or a fixed value with simple modifying factors. However, some of the models derived, particularly for electronic parts, are quite complex, taking account of many factors considered likely to affect reliability.

A model such as Ohm's law is credible because there is no question as to whether or not an electric current flows when an emf is applied across a conductor. However, whilst an engineering component might have properties such as conductance, mass, and so on, all unambiguously defined and measurable, it is very unlikely to have an intrinsic reliability that meets such criteria. For example, a good transistor or hydraulic actuator, if correctly applied, should not fail in use, during the expected life of the system in which it is used. If failures do occur in a population of these components in these systems, the causes, modes of failure and distributions of times to failure could be due to a range of different physical or chemical causes, as well as factors other than those explainable in purely physical or chemical terms. Some transistors might fail because of accidental overstress, some because of processing defects, or there may be no failures at all. If the hydraulic actuators are operated for a long time in a harsh environment, some might develop leaks which some operators might classify as failures. Also, failure, or the absence of failure, is heavily dependent upon human actions and perceptions. This is never true of laws of nature. This represents a fundamental limitation of the concept of reliability prediction using mathematical models.

As Niels Bohr, the famous Danish physicist once jokingly said 'Prediction is very difficult – especially if it is about the future.' We saw in Chapter 5 how reliability can vary by orders of magnitude with small changes in load and strength distributions, and the large amount of uncertainty inherent in estimating reliability from the load-strength model.

Another serious limitation arises from the fact that reliability models are usually based upon statistical analysis of past data. Much more data is required to derive a statistical relationship than to confirm a deterministic (theory-based) one, and even then there will be uncertainty because the sample can seldom be taken to be wholly representative of the whole population. For example, the true value of a life parameter is never known, only its distribution about an expected value, so we cannot say when failure will occur. Sometimes we can say that the likelihood increases, for example, in fatigue testing or if we detect wear in a bearing, but we can very rarely predict the time of failure. A statistically-derived relationship can never by itself be proof of a causal connection or even establish a theory. It must be supported by theory based upon an understanding of the cause-and-effect relationship.

Depending on the situation, a prediction can be based on past data, so long as we are sure that the underlying conditions which can affect future behaviour will not change significantly. However, since engineering is very much concerned with deliberate change, in design, processes and applications, predictions of reliability based solely on past data ignore the fact that changes might be made with the objective of improving reliability. Alternatively, sometimes changes introduce new reliability problems. Of course there are situations in which we can assume that change will not be significant, or in which we can extrapolate taking account of the likely effectiveness of planned changes. For example, in a system containing many parts which are subject to progressive deterioration, for example, an office lighting system containing many fluorescent lighting units, we can predict the frequency and pattern of failures fairly accurately, but these are special cases.

In general, it is important to appreciate that predictions of reliability can seldom be considered as better than rough estimates, and that achieved reliability can be considerably different to the predicted value.

6.3 Standards Based Reliability Prediction

Standards based reliability prediction is a methodology based on failure rate estimates published in globally recognized standards, both military and commercial. In some cases manufacturers are obliged by their customers or by contractual clauses to perform reliability prediction based on published standards.

A typical standards based reliability prediction treats devices as serial, meaning that one component failure causes a failure of a whole system. The other key assumption is a constant failure rate, which is modelled by the exponential distribution (see Section 2.6.3). This generally represents the useful life of a component where failures are considered random events (i.e. no wearout or early failures problem). It is well recognized in the engineering community that the assumption of a constant failure rate for both electronic and non-electronic parts can be misleading.

The most common approach is to state failure rates, expressed as failures per unit of time, often million (or even billion) hours. One failure per billion hours (10^9 hours) is commonly referred as one FIT. However, since FIT stands for *Failure in Time*, this unit of failure rate measure has not been universally accepted in the reliability engineering community. For non-repairable items this may be interpreted as the failure rate contribution to the system failure rate. These data are then used to synthesize system failure rate, by summation and by taking account of system configuration, as described later. Summation of part failure rates is generally referred to as the 'parts count' method. Reliability data can be useful in specific prediction applications, such as aircraft, petrochemical plant, computers or automobiles, when the data are derived from the area of application. However, such data should not be transferred from one application area to another without careful assessment. Even within the application area they should be used with care, since even then conditions can vary widely. An electric motor used to perform the same function, under the same loading conditions as previously in a new design of photocopier, might be expected to show the same failure pattern. However, if the motor is to be used for a different function, with different operating cycles, or even if it is bought from a different supplier, the old failure data might not be appropriate.

The commonly used standards include MIL-HDBK-217, Bellcore/Telcordia (SR-332), NSWC-06/LE10, China 299B, RDF 2000 and several others discussed later in this chapter. The typical analysis methods used by these standards include *parts count* and *parts stress analysis* methods. The parts count method requires less information, typically part quantities, quality levels and application environment. It is most applicable during very early design or proposal phases of a project.

6.3.1 MIL-HDBK-217

Probably the best known source of failure rate data for electronic components is US MIL-HDBK-217 (1995). It utilizes most of the principles listed before and is based on generic failure rates for electronic components collected over the years by the US Military. This handbook uses two methods of reliability prediction – parts count and parts stress. The parts count method assumes average stress levels as a means of providing an early design estimate of the failure rates. The overall equipment failure rate (ReliaSoft, 2006) can be calculated as:

$$\lambda = \sum_{i=1}^{n} N_i \pi_{Qi} \lambda_{bi}$$

where: n = number of parts categories (e.g. electrolytic capacitor, inductor, etc.).
N_i = quantity of ith part.
π_{Qi} = quality factor of ith part.
λ_{bi} = base failure rate of ith part.

The π-factors vary for component types and categories.

The parts stress method requires the greatest degree of detailed information. It is applied in the later phases of design when actual hardware and circuits are being designed. The parts stress method takes into account more information and the failure rate equations for each part contain more π-factors reflecting product environment, electrical stress, temperature factor, application environment, and other information specific to a component type and category. For example, the predicted failure rate for microcircuits can be calculated as:

$$\lambda_p = \pi_Q \pi_L [C_1 \pi_T + C_2 \pi_E]$$

Where π_Q, π_E, π_T and π_L are quality, environmental, temperature and learning factors respectively. C_1 is a die complexity factor based upon the chip gate count, or bit count for memory devices, or transistor count for linear devices and C_2 is a complexity factor based upon packaging aspects (number of pins, package type, etc.).

The criticisms of MIL-HDBK-217, which apply to most other standards-based methods for electronics include the following:

1 Experience shows that only a proportion of failures of modern electronic systems are due to components failing owing to internal causes.
2 The temperature dependence of failure rate is not always supported by modern experience or by considerations of physics of failure.
3 Several other parameters used in the models are of doubtful validity. For example, it is not true that failure rate increases significantly with increasing complexity, as continual process improvements counteract the effects of complexity.
4 The models do not take account of many factors that do affect reliability, such as transient overstress, temperature cycling, variation, EMI/EMC and control of assembly, test and maintenance.

Despite its flaws and the fact that MIL-HDBK-217 has not been updated since 1994 it is still being used to predict reliability. Therefore, there have been some efforts led by the United States Naval Surface Warfare Center (NSWC) to release MIL-HBDK-217 Revision G which would significantly update the existing standard including the introduction of aspects of physics of failure into the prediction procedure (see McLeish, 2010).

6.3.2 Telcordia SR-332 (Formerly Bellcore)

Other data sources for electronic components have been produced and published by non-military organizations, such as telecommunications companies for commercial applications. One of the most commonly used standards (particularly in Europe) is Telcordia SR-332, which was an update of the Bellcore document TR-332, Issue 6. The Bellcore reliability prediction model was originally developed by AT&T Bell Labs and was based on equations from MIL-HDBK-217, modified to better represent telecommunications industry field experience.

Telcordia SR-332 uses three different methods. Method I allows the user to obtain only the generic failure rates that are proposed by the Bellcore/Telcordia prediction standard. Method II allows the user to combine lab test data with the generic failure rates given in the standard. Method III allows the user to combine field data with the generic failure rates given in the standard. Additionally, SR-332 stresses the early life (infant mortality) problems of electronics and the use of burn-in by manufacturers to reduce the severity of infant mortality by weeding out weak components that suffer from early life problems (see ReliaSoft (2006) for more details). SR-332 also applies a First-Year-Multiplier factor that accounts for infant mortality risks in the failure rate prediction. The standard also applies a 'credit' for the use of a burn-in period and reduces the First-Year-Multiplier accordingly (i.e. the multiplier is smaller for longer periods of burn-in).

6.3.3 IEC 62380 (Formerly RDF 2000)

IEC Standard 62 380 TR Edition 1 was developed from RDF 2000, also formerly known as UTEC 80 810 (UTEC80810, 2000). This standard was designed as a further development of MIL-HDBK-217 where the multiplicative failure rate model was replaced by additive and multiplicative combinations of π-factors and failure rates. This standard allows specification of a temperature mission profile with different phases. The phases can have different temperatures that influence the failure rate of the components. The phases can also be of different types (on/off, permanently on, dormant) with various average outside temperature swings seen by the equipment. Those phases affect the failure rate calculation in different ways, as they apply different stresses on the components.

6.3.4 NSWC-06/LE10

Several databases have also been produced for non-electronic components, such as NSWC-06, 2006. This standard was developed from its earlier NSWC-98 version and uses a series of models for various categories of mechanical components to predict failure rates, which are affected by temperature, stresses, flow rates and various other parameters. The categories of mechanical equipment include seals, gaskets, springs, solenoids, valve assemblies, bearings, gears and splines, actuators, pumps, filters, brakes, compressors, electric motors and other non-electronic parts.

Many of the categories of mechanical equipment are in fact composed of a collection of sub-components which must be modelled by the user. For example, a collection for an electric motor would include bearings, motor windings, brushes, armature shaft, housing, gears. The user should be familiar with the equipment and the Handbook so that the correct type and number of sub-components can be included in the model.

Critics of this method note that the variety of types and applications of mechanical parts covered in the NSWC standard is very diverse, which amplifies the uncertainty of this type of prediction. Also, there is no general unit of operating time for a gasket or a spring, as there might be for an electronic component.

6.3.5 PRISM and 217Plus

The PRISM® reliability prediction tool was released in early 2000 by then the Reliability Analysis Center (RAC) to overcome the limitations of MIL-HDBK-217, which was not being actively maintained or updated (for more details see Dylis and Priore, 2001 and Alion, 2011). Also, the perception was that MIL-HDBK-217 produced pessimistic results due to the combined effect of multiplied π-factors. The premise of traditional methods of reliability predictions, such as MIL-HDBK-217, is that the failure rate of a system is primarily based on the components comprising that system. RAC data identified that more than 78 % of failures stem from non-component causes, namely: design deficiencies, manufacturing defects, poor system management techniques such as inadequate requirements, wearout, software, induced and no-defect-found failures. In response to this, the RAC developed PRISM for estimating the failure rate of electronic systems. The PRISM model development was based on large amounts of field and test data from both defense and commercial applications.

Instead of using the multiplicative modelling approach as used by MIL-HDBK-217, PRISM utilizes additive and multiplicative combinations of π-factors and failure rates λ for each class of failure mechanism. This approach is somewhat similar to the method of RDF 2000. PRISM incorporates new component reliability prediction models, includes a process for assessing the reliability of systems due to non-component variables, and includes a software reliability model. PRISM also allows the user to tailor the prediction based upon all available data including field data on a similar 'predecessor' system, and component-level test data. The PRISM-predicted failure rate model takes a form of:

$$\lambda = \sum_{i=1}^{n} N_i \sum_{j=1}^{m} \pi_{ij}\lambda_{ij}$$

where: n = number of parts categories.
N_i = quantity of ith part.
m = number of failure mechanisms appropriate for the ith part category.
π_{ij} = π–factor for the ith part category and jth failure mechanism.
λ_{ij} = failure rate for the ith part category and jth failure mechanism.

For example, $\pi_e\lambda_e$-addend would represent the contribution of product failure rate and π-factor for environmental stresses, $\pi_o\lambda_o$ for operational stresses, and so on.

217Plus^{TM} is a spin-off of PRISM, which generally uses the same modelling methodology, but has increased the number of part type failure rate models. 217Plus^{TM} models also include: connectors, switches, relays, inductors, transformers and opto-electronic devices. For more information see Nicholls, (2007).

6.3.6 China 299B (GJB/z 299B)

The GJB/z 299B Reliability Calculation Model for Electronic Equipment (often referred as China 299B) is a Chinese standard translated into English in 2001. This standard is a reliability prediction program based on the internationally recognized method of calculating electronic equipment reliability and was developed for the Chinese military. The standard is very similar to MIL-HDBK-217 and includes both parts count and parts stress analysis methods. This standard uses a series of models for various categories of electronic, electrical and electro-mechanical components to predict failure rates that are affected by environmental conditions, quality levels, stress conditions and various other parameters. It provides the methodology of calculating failure rates at both component and system level.

6.3.7 Other Standards

The list of the standards mentioned in this section is non-exhaustive and does not include less commonly used reliability prediction standards, some of which have been discontinued, but still maintain a limited use.

The British Telecom Handbook of Reliability Data (see British Telecom, 1995) is quite similar in approach to MIL-HDBK-217. Other less common standards include Siemens reliability standard SN29500.1 and its updated version SN 29 500-2005-1 as well as ‘Italtel Reliability Prediction Handbook,’ published by Italtel in 1993 and Nippon’s NTT procedure (Nippon, 1985) both discontinued.

The maintained standards include FIDES guide, updated in 2009 (FIDES, 2009). The FIDES methodology is based on the physics of failures and is supported by the analysis of test data, field returns and existing modelling. It is therefore different from the traditional methods developed mainly through statistical analysis of field returns. The methodology takes account of failures derived from development or manufacturing errors and overstresses (electrical, mechanical, thermal) related to the application.

Another non-electronic components standard is NPRD-95, ‘Non-electronic Parts Reliability Data’, released by RAC in the mid 1990-s. Part categories include actuators, batteries, pumps, and so on. Under the category the user would select a certain subtype (e.g. for batteries – Carbon Zinc, Lithium, etc.) in the same way as in NSWC-06.

6.3.8 IEEE Standard 1413

IEEE Standard 1413 (2003) has been created to establish a framework around which reliability prediction should be performed for electronic systems, though it could be applied to any technology. Prediction results obtained from an IEEE 1413 compliant reliability prediction are accompanied by responses to a set of questions identified in the standard. It identifies required elements for an understandable and credible reliability prediction with information to evaluate the effective use of the prediction results. IEEE 1413 however does not provide instructions for how to perform reliability prediction and does not judge any of the methodologies. In particular, an IEEE 1413 compliant prediction provides documentation of:

- the prediction results,
- the intended use of prediction results,
- the method(s) used,
- inputs required for selected method(s),
- the extent to which each input is known,

- the source of known input data,
- assumptions for unknown input data,
- figures of merit,
- confidence in prediction,
- sources of uncertainty,
- limitations and
- repeatability.

6.3.9 Software Tools for Reliability Prediction

Calculating reliability prediction by hand, especially for a system with a large number of parts, is obviously a long and tedious procedure prone to errors. There is a variety of commercial software packages available to a practitioner to run the reliability prediction based on a bill of materials (BOM), operational environments, applications, component stress data, and other information available during a system design phase. Most software packages allow a user to choose between the reliability prediction standards or run them in parallel. A non-exhaustive list of reliability prediction packages includes Lambda Predict® by ReliaSoft, CARE® by BQR, ITEM ToolKit, Reliability Workbench by Isograph, RAM COMMANDER by Reliass, and several others. Most of the software packages have capabilities of adding user-defined failure rates databases in addition to the published standards listed in this chapter.

6.4 Other Methods for Reliability Predictions

6.4.1 Field Return Based Methods

Some manufacturers prefer reliability predictions conducted with databases containing their own proprietary data. Failure rates can be calculated based on the field return, maintenance replacement, warranty claims or any other sources containing the information about failed parts along with parts still operating in the field. The definite advantage of this method is that results of the analysis are specific to the company's products, manufacturing processes and applications. Those predictions typically produce more accurate results than those coming from the generic databases. However, the drawback is the absence of common comparison criteria for a manufacturer to conduct supplier benchmarking.

6.4.2 Fusion of Field Data and Reliability Prediction Standards

The common criticism of reliability prediction standards is the empirical nature of the failure rates, which are quite generic. As a consequence, the predictions are not specific to any particular applications (automotive, avionics, consumer, etc.) even when all the appropriate π-factors are applied. Hence, there were various attempts to improve the accuracy of the calculated failure rates while staying within the realm of the known reliability prediction standards and models. Some of these methodologies include 'calibration' of base failure rates λ_b with internal warranty claims or field failure data (see, e.g. Kleyner and Bender, 2003).

Talmor and Arueti (1997) proposed an alternative method of merging internal company data with the reliability prediction standards. The procedure suggests evaluating quality factors π_Q to 'tailor' the reliability prediction models based on the results of the environmental stress screening (ESS) obtained early in the manufacturing process. Also, Kleyner and Boyle (2003) added a statistical dimension to the deterministic failure rate models by calculating 'equivalent failure rates' based on the temperature distribution for the part application instead of one fixed temperature value.

Some commercial reliability prediction packages allow the user to merge internal data with the standard-based models with the option to tune the prediction to the specific user's needs.

6.4.3 Physics of Failure Methods

The objective of physics-of-failure (PoF) analysis is to predict when a specific end-of-life failure mechanism will occur for an individual component or interconnect in a specific application. A physics-of-failure prediction looks at each individual failure mechanism such as metal fatigue, electromigration, solder joint cracking, wirebond adhesion, and so on, to estimate the probability of component failure within the expected life of the product (RAIC, 2010). In contrast to empirical reliability prediction methods based on historical failure data, this analysis requires detailed knowledge of all material characteristics, geometries and environmental conditions. The calculations involve understanding of the stresses applied to the part, types of failure mechanisms they would be causing and the appropriate model to calculate the expected life to a failure caused by the particular failure mechanism in question. More on mechanical and electronic time to failure models will be covered in Chapters 8, 9 and 13.

The advantage of the physics-of-failure approach is that fairly accurate predictions using known failure mechanisms can be performed to determine the wearout point. PoF methods address the potential failure mechanism and the stresses on the product; therefore it is more specific to the product design, its applications and is expected to be more accurate than other types of reliability prediction. The disadvantage is that this method requires knowledge of the component manufacturer's materials, processes, design, and other data, not all of which may be available at the early design stage. In addition, the actual calculations and analysis are complicated and sometimes costly activities requiring a lot of information and a high level of analytical expertise. Additional criticism includes the difficulty to address the entire system, since most of the analysis is done on a component or sub-assembly level.

A large amount of work of studying physics of failure and developing PoF based reliability prediction models has been done at CALCE (Computer Aided Life Cycle Engineering) Center at the University of Maryland, USA (CALCE, 2011). CALCE PoF based methodology software packages for both component and assembly levels include CalcePWA®, CalceFast®, CalceEP® (see Foucher *et al.*, 2002).

6.4.4 'Top Down' Approach to Reliability Prediction

Having identified the fundamental limitations of reliability prediction models and data, we are still left with the problem that it is often necessary to predict the likely reliability of a new system. It is possible to make reasonably credible reliability predictions, without using the kinds of models described above, for systems under certain circumstances. These are:

1 The system is similar to systems developed, built and used previously, so that we can apply our experience of what happened before.
2 The new system does not involve significant technological risk (this follows from 1).
3 The system will be manufactured in large quantities, or is complex (i.e. contains many parts, or the parts are complex) or will be used over a long time, or a combination of these conditions applies, that is there is an asymptotic property.
4 There is a strong commitment to the achievement of the reliability predicted.

Thus, we can make credible reliability predictions for a new TV receiver or automobile engine. No great changes from past practice are involved, technological risks are low, they will be built in large quantities and they are quite complex, and the system must compete with established, reliable products.

Such reliability predictions (in the sense of a reasonable expectation) could be made without recourse to statistical or empirical mathematical models at the level of individual parts. Rather, they could be based upon knowledge of past performance at the system level, the possible effects of changes, and on management targets and priorities. This is a 'top down' prediction.

6.5 Practical Aspects

The reliability prediction does not ensure that the reliability values will be achieved; it is not a demonstration in the way that a mass or power consumption prediction, being based on physical laws, would be. Rather, it should be used as the basis for setting the objective, which is likely to be attained only if there is a management commitment to it. Reliability predictions must, therefore, take account of objectives and assessment of risks, in that order. This must be an iterative procedure, since objectives and risks must be balanced; the reliability engineer plays an important part in this process, since he or she must assess whether objectives are realistic in relation to the risks. This assessment should be made top down, but it can be aided by the educated use of appropriate models and data, so long as their limitations and margins of error are appreciated. Once the risks are assessed and the objective is quantified, development must be continuously monitored in relation to the reduction of risks through analysis, tests and corrective actions, and to the measured reliability during tests. This is necessary to provide assurance that the objective will be met, if need be by additional management action such as provision of extra resources to solve particular problems.

The purpose to which the prediction will be applied should also influence the methods used and the estimates derived. For example, if the prediction will be used to determine spare item stocks or repair costs, an optimistic figure might be acceptable. However, if it is to be used as part of a safety analysis, a pessimistic figure would be more appropriate. It is good practice to indicate the likely expected ranges of uncertainty, when possible.

Additionally, reliability predictions can be used as an effective tool in a comparative analysis. When choosing between different design alternatives, the inherent reliability of each design obtained through reliability prediction analysis could be used as a critical decision making factor. However, the uncertainty of the prediction values should be taken into account.

In situations in which the methods to be used are imposed, the reliability prediction report should state that the results are derived accordingly. The predictions should always take account of objectives and related management aspects, such as commitment and risk. If management does not 'drive' the reliability effort, the prediction can become a meaningless exercise. As overriding considerations, it must be remembered that there is no theoretical limit to the reliability that can be attained, and that the achievement of high reliability does not always entail higher costs.

6.6 Systems Reliability Models

6.6.1 The Basic Series Reliability Model

Consider a system composed of two independent components, each exhibiting a constant hazard (or failure) rate. If the failure of either component will result in failure of the system, the system can be represented by a *reliability block diagram* (RBD) (Figure 6.1). (A reliability block diagram does not necessarily represent the system's operational logic or functional partitioning.)

If λ_1 and λ_2 are the hazard rates of the two components, the system hazard rate will be $\lambda_1 + \lambda_2$. Because the hazard rates are constant, the component reliabilities R_1 and R_2, over a time of operation t, are

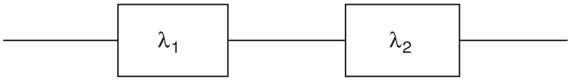


Figure 6.1 *Series System.*

exp $(-\lambda_1 t)$ and exp $(-\lambda_2 t)$. The reliability of the system is the combined probability of no failure of either component, that is $R_1 R_2 = \exp\,[-(\lambda_1 + \lambda_2)t]$. In general, for a series of n independent components:

$$R = \prod_{i=1}^{n} R_i \tag{6.1}$$

where R_i is the reliability of the ith component. This is known as the product rule or series rule (see Eqn. 2.2):

$$\lambda = \sum_{i=1}^{n} \lambda_i \quad \text{and} \quad R = \exp(-\lambda t) \tag{6.2}$$

This is the simplest basic model on which parts count reliability prediction is based.

The failure logic model of the overall system will be more complex if there are redundant subsystems or components. Also, if system failure can be caused by events other than component failures, such as interface problems, the model should specifically include these, for example as extra blocks.

6.6.2 Active Redundancy

The reliability block diagram for the simplest redundant system is shown in Figure 6.2. In this system, composed of two independent parts with reliabilities R_1 and R_2, satisfactory operation occurs if *either* one or *both* parts function. Therefore, the reliability of the system, R, is equal to the probability of part 1 *or* part 2 surviving.

From Eqn. (2.6), the probability

$$(R_1 + R_2) = R_1 + R_2 - R_1 R_2$$

This is often written

$$1 - (1 - R_1)(1 - R_2)$$

For the constant hazard rate case,

$$R = \exp(-\lambda_1 t) + \exp(-\lambda_2 t) - \exp\,[-(\lambda_1 + \lambda_2)t] \tag{6.3}$$

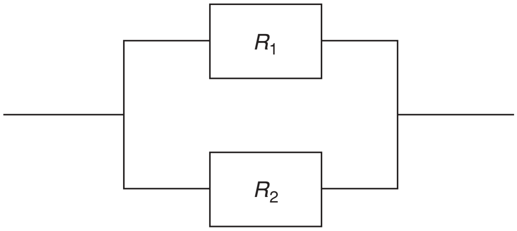


Figure 6.2 Dual redundant system.

The general expression for active parallel redundancy is

$$R = 1 - \prod_{i=1}^{n} (1 - R_i) \tag{6.4}$$

where R_i is the reliability of the ith unit and n the number of units in parallel. If in the two-unit active redundant system $\lambda_1 = \lambda_2 = 0.1$ failures per 1000 h, the system reliability over 1000 h is 0.9909. This is a significant increase over the reliability of a simple non-redundant unit, which is 0.9048. Such a large reliability gain often justifies the extra expense of designing redundancy into systems. The gain usually exceeds the range of prediction uncertainty. The example quoted is for a *non-maintained* system, that is the system is not repaired when one equipment fails. In practice, most active redundant systems include an indication of failure of one equipment, which can then be repaired. A maintained active redundant system is, of course, theoretically more reliable than a non-maintained one. Examples of non-maintained active redundancy can be found in spacecraft systems (e.g. dual thrust motors for orbital station-keeping) and maintained active redundancy is a feature of systems such as power generating systems and railway signals.

6.6.3 *m*-out-of-*n* Redundancy

In some active parallel redundant configurations, *m* out of the *n* units may be required to be working for the system to function. This is called *m*-out-of-*n* (or *m/n*) parallel redundancy. The reliability of an *m/n* system, with *n* independent components in which all the unit reliabilities are equal, is the binomial reliability function (based on Eq. 2.37):

$$R_{SYS} = 1 - \sum_{i=0}^{m-1} \binom{n}{i} R^i (1 - R)^{n-i} \tag{6.5}$$

or, for the constant hazard rate case:

$$R_{SYS} = 1 - \frac{1}{(\lambda t + 1)^n} \sum_{i=0}^{m-1} \binom{n}{i} (\lambda t)^{n-i} \tag{6.6}$$

6.6.4 Standby Redundancy

Standby redundancy sometimes referred as 'cold standby' is achieved when one unit does not operate continuously but is only switched on when the primary unit fails. A standby electrical generating system is an example. The block diagram in Figure 6.3 shows another. The standby unit and the sensing and switching system may be considered to have a 'one-shot' reliability R_s of starting and maintaining system function until the primary equipment is repaired, or R_s may be time-dependent. The switch and the redundant unit may have dormant hazard rates, particularly if they are not maintained or monitored.

Taking the case where the system is non-maintained, the units have equal constant operating hazard rates λ, there are no dormant failures and $R_s = 1$, then

$$R_{SYS} = \exp(-\lambda t) + \lambda t \exp(-\lambda t) \tag{6.7}$$

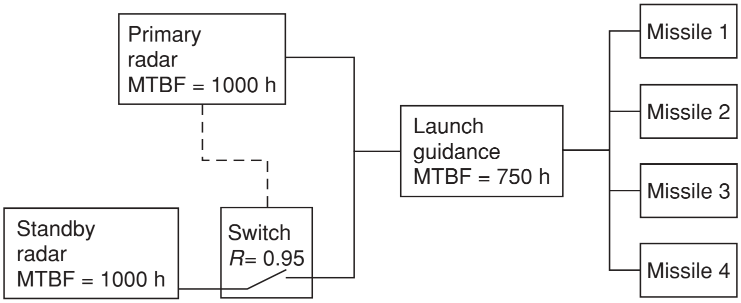

Primary radar MTBF = 1000 h
Standby radar MTBF = 1000 h
Switch R= 0.95
Launch guidance MTBF = 750 h
Missile 1
Missile 2
Missile 3
Missile 4

Figure 6.3 *Reliability block diagram for a missile system.*

The general reliability formula for n equal units in a standby redundant configuration (perfect switching) is

$$R_{SYS} = \sum_{i=0}^{n-1} \frac{(\lambda t)^i}{i!} \exp(-\lambda t) \tag{6.8}$$

If in a standby redundant system $\lambda_1 = \lambda_2 = 0.1$ failure per 1000 h, then the system reliability is 0.9953. This is higher than for the active redundant system [$R(1000) = 0.9909$], since the standby system is at risk for a shorter time. If we take into account less than perfect reliability for the sensing and switching system and possibly a dormant hazard rate for the standby equipment, then standby system reliability would be reduced.

6.6.5 Further Redundancy Considerations

The redundant systems described represent the tip of an iceberg as far as the variety and complexity of system reliability models are concerned. For systems where very high safety or reliability is required, more complex redundancy is frequently applied. Some examples of these are:

1 In aircraft, dual or triple active redundant hydraulic power systems are often used, with a further emergency (standby) back-up system in case of a failure of all the primary circuits.
2 Aircraft electronic flying controls typically utilize triple voting active redundancy. A sensing system automatically switches off one system if it transmits signals which do not match those transmitted by the other two, and there is a manual back-up system. The reliability evaluation must include the reliability of all three primary systems, the sensing system and the manual system.
3 Fire detection and suppression systems consist of detectors, which may be in parallel active redundant configuration, and a suppression system which is triggered by the detectors.

We must be careful to ensure that single-point failures which can partly eliminate the effect of redundancy are considered in assessing redundant systems. For example, if redundant electronic circuits are included within one integrated circuit package, a single failure such as a leaking hermetic seal could cause both circuits to fail. Such dependent failures are sometimes referred to as *common mode* (or *common cause*) failures, particularly in relation to systems. As far as is practicable, they must be identified and included in the analysis. Common mode failures are discussed in more detail later.

6.7 Availability of Repairable Systems

As mentioned in Section 2.15, there is a fundamental difference between the mathematical treatments of repairable and non-repairable systems. None of the commonly used statistical distributions can be applied to repairable systems due to the fact that failed units are not taken out of the population. This statement can be illustrated by a simple example where the number of repaired failures eventually exceeds the total number of parts in the field, thus making the cdf of a distribution greater than 1.0, which is mathematically impossible. Instead, repairable systems are modelled by a stochastic process. If a system can be repaired to 'as good as new' condition, then the appropriate model to describe the failure occurrence is called an *ordinary renewal process* (ORP) (see Section 2.15.2). If a system upon repair retains the same wearout characteristics as before it has a condition called 'same as old', and it is modelled by the non-homogeneous Poisson process (NHPP), see also Section 2.15.2. If the condition after repair is better than old, but worse than new (which is usually the case in real life), then it is modelled by the so-called *generalized renewal process* (GRP), see Kaminskiy and Krivtsov (2000) for more details.

Therefore, for a repairable system the 'classic' definition of reliability applies only to the time to first failure. Instead the reliability-equivalent of a repairable system is called *availability*. Availability is defined as the probability that an item will be available when required, or as the proportion of total time that the item is available for use. Therefore the availability of a repairable item is a function of its failure rate, λ, and of its repair or replacement rate μ. The difference between repairable and non-repairable systems is illustrated graphically in Figure 6.4.

The proportion of total time that the item is available (functional) is the *steady-state availability*. For a simple unit, with a constant failure rate λ and a constant mean repair rate μ, where $\mu = 1/\text{MTTR}$ (Mean Time to Repair), the steady-state availability is equal to:

$$A = \frac{\text{MTBF}}{\text{MTBF} + \text{MTTR}} = \frac{\mu}{\lambda + \mu} \tag{6.9}$$

The instantaneous availability or probability that the item will be available at time t is equal to

$$A = \frac{\mu}{\lambda + \mu} + \frac{\lambda}{\lambda + \mu} \exp[-(\lambda + \mu)t] \tag{6.10}$$

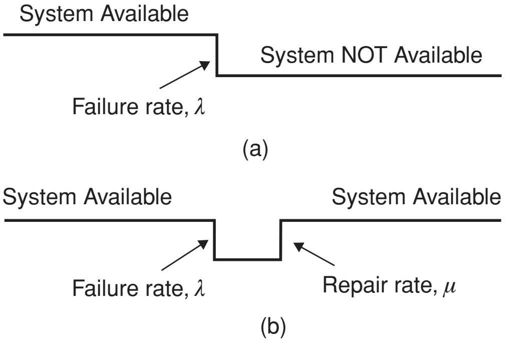
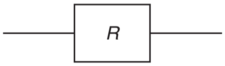
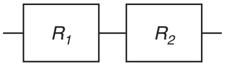
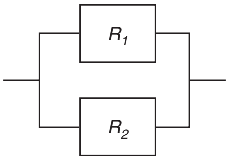
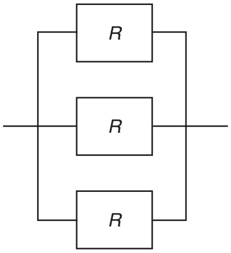


Figure 6.4 (a) Non-repairable system and (b) Repairable system.

Table 6.1 *Reliability and availability for some systems configurations. (R. H. Myers, K. L. Wong and H. M. Gordy,* Reliability Engineering for Electronic Systems*, Copyright © 1964 John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.)*

Reliability configuration		Reliability (no repair) for CFR λ	General system reliability for *n* blocks
		$\exp(-\lambda t)$	
		$\exp[-(\lambda_1+\lambda_2)t]$	$\prod_{i=1}^{n} R_i$
	Active	$\exp(-\lambda_1 t)+\exp(-\lambda_2 t)$ $-\exp[-(\lambda_1+\lambda_2)t]$	$1-\prod_{i=1}^{n}(1-R_i)$
	Standby	$\dfrac{\lambda_2\exp(-\lambda_1 t)-\lambda_1\exp(-\lambda_2 t)}{\lambda_2-\lambda_1}$[b]	$\exp(-\lambda t)\sum_{i=0}^{n-1}\dfrac{(\lambda t)^i}{i!}$[a]
	Active 1/3	$3\exp(-\lambda t)-3\exp(-2\lambda t)$ $+\exp(-3\lambda t)$	As above (active)
	Active 2/3	$3\exp(-2\lambda t)-2\exp(-3\lambda t)$	$1-\sum_{i=0}^{m-1}\binom{n}{i}R^i(1-R)^{n-i}$[c]
	Standby 1/3	$\exp(-\lambda t)+\lambda t\exp(-\lambda t)$[d] $+\frac{1}{2}\lambda^2t^2\exp(-\lambda t)$	$\exp(-\lambda t)\sum_{i=0}^{n-1}\dfrac{(\lambda t)^i}{i!}$

[a] $\lambda_1=\lambda_2=\lambda$. Assumes series repair, that is single repair team.

[b] When $\lambda_1=\lambda_2$ the reliability formula becomes indeterminate. When $\lambda_1=\lambda_2$ use $R(t)=\exp(-\lambda t)+\lambda t\exp(-\lambda t)$. If $\lambda_1\approx\lambda_2$ use $\lambda=(\lambda_1+\lambda_2)/2$. Assumes perfect switching.

[c] For *m*-out-of-*n* redundancy.

[d] Assumes perfect switching.

which approaches the steady-state availability as t becomes large. It is often more revealing, particularly when comparing design options, to consider system *unavailability*:

$$\begin{aligned}\text{Steady-state unavailability} &= 1-A\\ &= \frac{\lambda}{\lambda+\mu}\end{aligned} \tag{6.11}$$

and

$$\text{Instantaneous unavailability} = \frac{\lambda}{\lambda+\mu}-\frac{\lambda}{\lambda+\mu}\exp[-(\lambda+\mu)t] \tag{6.12}$$

If scheduled maintenance is necessary and involves taking the system out of action, this must be included in the availability formula. The availability of spare units for repair by replacement is often a further consideration, dependent upon the previous spares usage and the repair rate of replacement units.

Availability is an important consideration in relatively complex systems, such as telecommunications and power networks, chemical plant and radar stations. In such systems, high reliability by itself is not sufficient to ensure that the system will be available when needed. It is also necessary to ensure that it can be repaired

Table 6.1 *(Continued)*

R for $\lambda_1 = \lambda_2 = 0.01$, $t = 100$	Steady-state availability, A, repair rate, μ, CFR, λ	General steady-state availability, A, for n blocks	A for $\lambda = 0.01$, $\mu = 0.2$
0.37	$\dfrac{\mu}{\lambda + \mu}$	—	0.95
0.14	$\dfrac{\mu_1\mu_2}{\mu_1\mu_2 + \mu_1\lambda_2 + \mu_2\lambda_1 + \lambda_1\lambda_2}$	$\prod_{i=1}^{n} \dfrac{\mu_i}{\lambda_i + \mu_i}$	0.907
0.60	$\dfrac{\mu^2 + 2\mu\lambda^{\text{a}}}{\mu^2 + 2\mu\lambda + 2\lambda^2}$	$1 - \prod_{i=1}^{n} \dfrac{\lambda_i^{\text{a}}}{\lambda_i + \mu_i}$	0.996
0.74	$\dfrac{\mu^2 + \mu\lambda}{\mu^2 + \mu\lambda + \lambda^2}$	—	0.998
0.75	$\dfrac{\mu^3 + 3\mu^2\lambda + 6\mu\lambda^2}{\mu^3 + 3\mu^2\lambda + 6\mu\lambda^2 + 6\lambda^3}$	As above (active)	0.9999
0.31	$\dfrac{\mu^3 + 3\mu^2\lambda}{\mu^3 + 3\mu^2\lambda + 6\mu\lambda^2 + 6\lambda^3}$	$1 - \dfrac{1}{(\lambda + \mu)^n} \sum_{i=0}^{m-1} \dbinom{n}{i} \mu^i \lambda^{n-i}$	0.987
0.92	$\dfrac{\mu^3 + \mu^2\lambda + \mu\lambda^2}{\mu^3 + \mu^2\lambda + \mu\lambda^2 + \lambda^3}$	—	0.9999

quickly and that essential scheduled maintenance tasks can be performed quickly, if possible without shutting down the system. Therefore maintainability is an important aspect of design for maximum availability, and trade-offs are often necessary between reliability and maintainability features. For example, built-in test equipment (BITE) is incorporated into many electronic systems. This added complexity can degrade reliability and can also result in spurious failure indications. However, BITE can greatly reduce maintenance times, by providing an instantaneous indication of fault location, and therefore availability can be increased. (This is not the only reason for the use of BITE. It can also reduce the need for external test equipment and for training requirements for trouble-shooting, etc.)

Availability is also affected by redundancy. If standby systems can be repaired or overhauled while the primary system provides the required service, overall availability can be greatly increased.

Table 6.1 shows the reliability and steady-state availability functions for some system configurations. It shows clearly the large gains in reliability and steady-state availability which can be provided by redundancy. However, these are relatively simple situations, particularly as a constant failure rate is assumed. Also, for the standby redundant case, it is assumed that:

1 The reliability of the changeover system is unity.
2 No common-cause failures occur.
3 Failures are detected and repaired as soon as they occur.

Of course, these conditions do not necessarily apply, particularly in the case of standby equipment, which must be tested at intervals to determine whether it is serviceable. The availability then depends upon the test interval. Monitoring systems are sometimes employed, for example, built-in test equipment (BITE) for electronic equipment, but this does not necessarily have a 100 % chance of detecting all failures. In real-life situations it is necessary to consider these aspects, and the analysis can become very complex. Methods for dealing with more complex systems are given at the end of this chapter, and maintenance and maintainability are covered in more detail in Chapter 16.

Example 6.1

A shipboard missile system is composed of two warning radars, a control system, a launch and guidance system, and the missiles. The radars are arranged so that either can give warning if the other fails, in a standby redundant configuration. Four missiles are available for firing and the system is considered to be reliable if three out of four missiles can be fired and guided. Figure 6.3 shows the system in reliability block diagram form, with the MTBFs of the subsystems. The reliability of each missile is 0.9. Assuming that: (1) the launch and guidance system is constantly activated, (2) the missile flight time is negligible and (3) all elements are independent, evaluate: (a) the reliability of the system over 24 h, (b) the steady-state availability of the system, excluding the missiles, if the mean repair time for all units is 2 h and the changeover switch reliability is 0.95.

The reliabilities of the units over a 24 h period are:

Primary radar 0.9762 (failure rate $\lambda_P = 0.001$).
Standby radar 0.9762 (failure rate $\lambda_S = 0.001$).
Launch and guidance 0.9685 (failure rate $\lambda_{LG} = 0.0013$).

a The overall radar reliability is given by (from Table 6.1)

$$\begin{aligned} R_{\text{radar}} &= \exp(-\lambda t) + \lambda t \exp(-\lambda t) \\ &= 0.9762 + (0.001 \times 24 \times 0.9762) = 0.9996 \end{aligned}$$

The probability of the primary radar failing is (1 − R_P). The probability of this radar failing *and* the switch failing is the product of the two failure probabilities:

$$(1 - 0.9762)(1 - 0.95) = 0.0012$$

Therefore the switch reliability effect can be considered equivalent to a series unit with reliability

$$R_{SW} = (1 - 0.0012) = 0.9988$$

The system reliability, up to the point of missile launch, is therefore

$$\begin{aligned} R_S &= R_{\text{radar}} \times R_{SW} \times R_{LG} \\ &= 0.9996 \times 0.9988 \times 0.9685 = 0.9670 \end{aligned}$$

The reliability of any three out of four missiles is given by the cumulative binomial distribution (Eq. 6.5):

$$R_{\mathrm{M}} = 1 - \left[\binom{4}{0} 0.9^{0} \times 0.1^{4} + 410.9^{1} \times 0.1^{3} + 420.9^{2} \times 0.1^{2} \right]$$
$$= 0.9477$$

The total system reliability is therefore

$$R'_{\mathrm{S}} = R_{\mathrm{S}} \times R_{\mathrm{M}}$$
$$= 0.9760 \times 0.9477 = 0.9250$$

b The availability of the redundant radar configuration is (see Table 6.1)

$$A_{\mathrm{radar}} = \frac{\mu^{2} + \mu\lambda}{\mu^{2} + \mu\lambda + 2\lambda^{2}}$$
$$= \frac{0.5^{2} + (0.5 \times 0.001)}{(0.5)^{2} + (0.5 \times 0.001) + 2(0.001)^{2}}$$
$$= 0.999997 \qquad (\text{unavailability} = 3 \times 10^{6})$$

The availability of the launch and guidance system

$$A_{\mathrm{LG}} = \frac{\mu}{\mu + \lambda}$$
$$= \frac{0.5}{0.5 + 0.0013} = 0.9974 \qquad (\text{unavailability} = 2.6 \times 10^{-3})$$

The system availability is therefore

$$A_{\mathrm{radar}} \times A_{\mathrm{LG}} = 0.9974 \qquad (\text{unavailability} = 2.6 \times 10^{-3})$$

The previous example can be used to illustrate how such an analysis can be used for performing sensitivity studies to compare system design options. For example, a 20 % reduction in the MTBF of the launch and guidance system would have a far greater impact on system reliability than would a similar reduction in the MTBF of the two radars.

In maintained systems which utilize redundancy for reliability or safety reasons, separate analyses should be performed to assess system reliability in terms of the required output and failure rate in terms of maintenance arisings. In the latter case all elements can be considered as being in series, since all failures, whether of primary or standby elements, lead to repair action.

Table 6.2 *MTBR and replacement costs for the four modules.*

	MTBR (h)	Replacement costs ($)
Module 1	2500	3000
Module 2	4000	2000
Module 3	4000	2500
Module 4	10 000	10 000

Table 6.3 *Cost per year of replacing the modules.*

	Replacements per year	Cost per year ($)
Module 1	12	36 000
Module 2	7.5	15 000
Module 3	7.5	18 750
Module 4	3	3000
Total	30	$ 72 750

6.8 Modular Design

Availability and the cost of maintaining a system can also be influenced by the way in which the design is partitioned. ‘Modular’ design is used in many complex products, such as electronic systems and aero engines, to ensure that a failure can be corrected by a relatively easy replacement of the defective module, rather than by replacement of the complete unit.

Example 6.2

An aircraft gas turbine engine has a mean time between replacements (MTBR) – scheduled and unscheduled – of 1000 flight hours. With a total annual flying rate of 30 000 h and an average cost of replacement of $ 10 000, the annual repair bill amounted to $ 300 000. The manufacturer redesigned the engine so that it could be separated into four modules, with MTBR and replacement costs as shown in Table 6.2. What would be the new annual cost?

With the same total number of replacements, the annual repair cost is greatly reduced, from $ 300 000 to $ 72 750 (see Table 6.3).

Note that Example 6.2 does not take into account the different spares holding that would be required for the modular design (i.e. the operator would keep spare modules, instead of spare engines, thus making a further saving). In fact other factors would complicate such an analysis in practice. For example the different scheduled overhaul periods of the modules compared with the whole engine, the effect of wearout failure modes giving non-constant replacement rates with time since overhaul, and so on. Monte Carlo simulation is often used for planning and decision-making in this sort of situation (see Chapter 4).

6.9 Block Diagram Analysis

The failure logic of a system can be shown as a reliability block diagram (RBD), which shows the logical connections between components of the system. The RBD is not necessarily the same as a block schematic diagram of the system’s functional layout. We have already shown examples of RBDs for simple series and parallel systems. For systems involving complex interactions construction of the RBD can

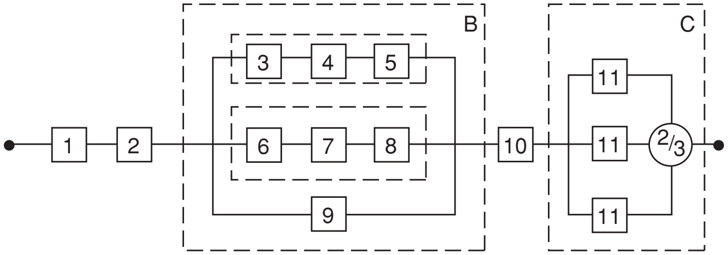


Figure 6.5 Block diagram decomposition.

be quite difficult, and a different RBD will be necessary for different definitions of what constitutes a system failure.

Block diagram analysis consists of reducing the overall RBD to a simple system which can then be analysed using the formulae for series and parallel arrangements. It is necessary to assume independence of block reliabilities.

The technique is also called block diagram *decomposition*. It is illustrated in Example 6.3.

Example 6.3

The system shown in Figure 6.5 can be reduced as follows (assuming independent reliabilities):

$$R_\text{S} = R_1 \times R_2 \times R_\text{B} \times R_{10} R_\text{C} \qquad \text{(from Eqn. 6.1)}$$

$$R_\text{B} = 1 - [1 - (R_3 \times R_4 \times R_5)][1 - (R_6 \times R_7 \times R_8)](1 - R_9) \qquad \text{(from Eqn. 6.4)}$$

$$\begin{aligned} R_\text{C} &= 1 - \frac{3 \times 2}{3 \times 2} R_{11}^0 (1 - R_{11})^3 + \frac{3 \times 2}{2} R_{11}(1 - R_{11})^2 \\ &= 1 - (1 - R_{11})^3 + 3R_{11}(1 - R_{11})^2 \end{aligned} \qquad \text{(from Eqn. 6.5)}$$

6.9.1 *Cut* and *Tie Sets*

Complex RBDs can be analysed using *cut set* or *tie set* methods. A cut set is produced by drawing a line through blocks in the system to show the minimum number of failed blocks which would lead to system failure. Tie sets (or *path sets*) are produced by drawing lines through blocks which, if all were working, would allow the system to work. Figure 6.6 illustrates the way that cut and tie sets are produced. In this system there are three cut sets and two tie sets.

Approximate bounds on system reliability as derived from cut sets and tie sets, respectively, are given by

$$R_\text{s} > 1 - \sum_j^N \prod_i^{n_j} (1 - R_i) \qquad (6.13)$$

$$R_\text{s} < \sum_j^T \prod_i^{n_j} R_i \qquad (6.14)$$

where N is the number of cut sets, T is the number of tie sets and n_j is the number of blocks in the jth cut set or tie set.

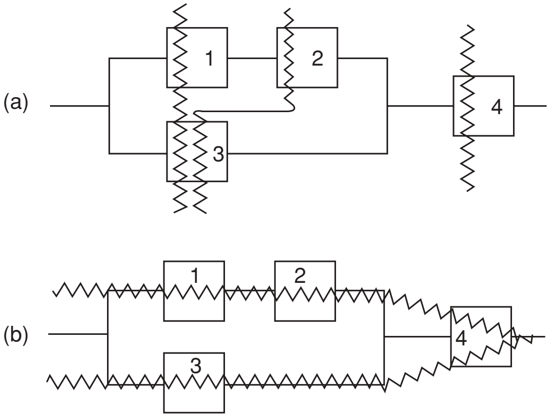


Figure 6.6 (a) Cut sets and (b) tie sets.

Example 6.4

Determine the reliability bounds of the system in Figure 6.6 for

$$R_1 = R_2 = R_3 = R_4 = 0.9.$$

Cut set:

$$\begin{aligned}R_s &> 1 - [(1 - R_1)(1 - R_3) + (1 - R_2)(1 - R_3) + (1 - R_4)]\\ &> 1 - [3 - R_1 - R_2 - 2R_3 - R_4 + R_1R_3 + R_2R_3]\\ &> 1 - 0.12 = 0.88\end{aligned}$$

Tie set:

$$\begin{aligned}R_s &< R_1R_2R_4 + R_3R_4\\ &< 1.54\,(\text{i.e.} < 1.0)\end{aligned}$$

For comparison, the exact reliability is

$$\begin{aligned}R_s &= [1 - (1 - R_1R_2)(1 - R_3)]R_4\\ &= R_3R_4 + R_1R_2R_4 - R_1R_2R_3R_4\\ &= 0.883\end{aligned}$$

The cut and tie set approaches are not used for systems as simple as in Example 6.4, since the decomposition approach is easy and gives an exact result. However, since the derivation of exact reliability using the decomposition approach can become an intractable problem for complex systems, the cut and tie set approach has its uses in such applications. The approximations converge to the exact system reliability as the system complexity increases, and the convergence is more rapid when the block reliabilities are high. Tie sets are not usually identified or evaluated in system analysis, however.

Cut and tie set methods are suitable for computer application. Their use is appropriate for the analysis of large systems in which various configurations are possible, such as aircraft controls, power generation, or control and instrumentation systems for large plant installations. The technique is subject to the constraint (as is the decomposition method) that all block reliabilities must be independent.

6.9.2 Common Mode Failures

A common mode (or common cause) failure is one which can lead to the failure of all paths in a redundant configuration. Identification and evaluation of common mode failures is very important, since they might have a higher probability of occurrence than the failure probability of the redundant system when only individual path failures are considered. In the design of redundant systems it is very important to identify and eliminate sources of common mode failures, or to reduce their probability of occurrence to levels an order or more below that of other failure modes.

For example, consider a system in which each path has a reliability $R = 0.99$ and a common mode failure which has a probability of non-occurrence $R_{CM} = 0.98$. The system can be designed either with a single unit or in a dual redundant configuration (Figure 6.7 (a) and (b)). Ignoring the common mode failure, the reliability of the dual redundant system would be 0.9999. However, the common mode failure practically eliminates the advantage of the redundant configuration.

Examples of sources of common mode failures are:

1 Changeover systems to activate standby redundant units.
2 Sensor systems to detect failure of a path.
3 Indicator systems to alert personnel to failure of a path.
4 Power or fuel supplies which are common to different paths.
5 Maintenance actions which are common to different paths, for example, an aircraft engine oil check after which a maintenance technician omits to replace the oil seal on *all* engines. (This has actually happened twice, very nearly causing a major disaster each time.)
6 Operating actions which are common to different paths, so that the same human error will lead to loss of both.

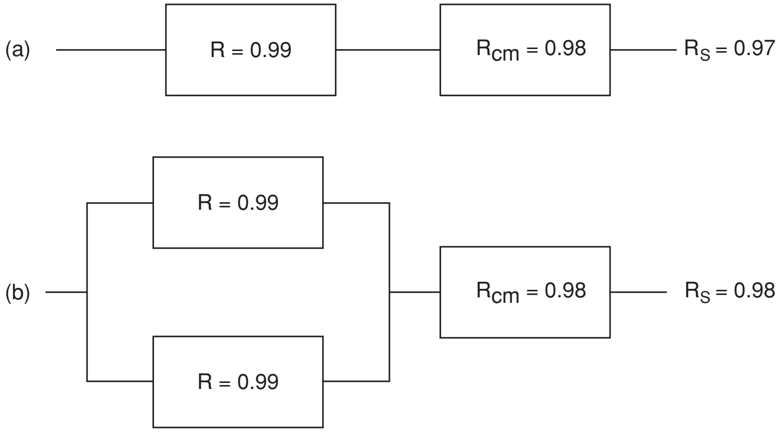


Figure 6.7 *Effect of common mode failure.*

7 Software which is common to all paths, or software timing problems between parallel processors.
8 'Next weakest link' failures. Failure of one item puts an increased load on the next item in series, or on a redundant unit, which fails as a result.

Common mode failures can be very difficult to foresee, and great care must be taken when analysing safety aspects of systems to ensure that possible sources are identified.

6.9.3 Enabling Events

An enabling event is one which, whilst not necessarily a failure or a direct cause of failure, will cause a higher level failure event when accompanied by a failure. Like common mode failures, enabling events can be important and difficult to anticipate, but they must be considered. Examples of enabling events are:

1 Warning systems disabled for maintenance, or because they create spurious warnings.
2 Controls incorrectly set.
3 Operating or maintenance personnel following procedures incorrectly, or not following procedures.
4 Standby elements being out of action due to maintenance.

6.9.4 Practical Aspects

It is essential that practical engineering considerations are applied to system reliability analyses. The reliability block diagram implies that the 'blocks' are either 'failed' or operating, and that the logic is correct. Examples of situations in which practical and logical errors can occur are:

1 Two diodes (or two check valves) connected in series. If either fails open circuit (stuck closed), there will be no current (fluid) flow, so they will be in series from a reliability point of view. On the other hand, if either fails short circuit (or stuck open), the other will provide the required system function (current/fluid will flow in one direction), so they will be in parallel from a reliability point of view.
2 Two thermally activated switches (thermostats) wired in parallel to provide over-temperature protection for a system by starting a cooling fan, with backup in case of failure of one of the switches. This configuration might be modelled as two switches in parallel. However, practical engineering considerations that could make this simple model misleading or invalid are:
   - If one fails to close, the other will perform the protection function. However, there is no way of knowing that one has failed (without additional circuits or checks).
   - If one fails permanently closed, the fan would run continuously.
   - Since the two switches will operate at slightly different temperatures, one will probably do all the switching, so the duty will not be shared equally. If the one that switches first fails open, the contacts on the other might have degraded due to inactivity, so that it fails to switch.
   - If they do both switch about the same number of times, they might both deteriorate (contact wear) at the same rate, so that when one fails the other might fail soon after.
3 Data on the in-flight failure probabilities of aircraft engines are used to determine whether new types of commercial aircraft meet safety criteria (e.g. 'extended twin overwater operations' (ETOPS)). Engines might fail so that they do not provide power, but also in ways that cause consequential damage and possible loss of the aircraft. If the analyses consider only loss of power, they would be biased against twin-engine aircraft.

4 Common mode failures are often difficult to predict, but can dominate the real reliability or safety of systems. Maintenance work or other human actions are prime contributors. For example
   – The Chernobyl nuclear reactor accident was caused by the operators conducting an unauthorized test.
5 Unexpected combinations of events can occur. For example:
   – The Concorde crash was caused by debris on the runway causing tyre failure, and fragments of the tyre then pierced the fuel tank.
   – The explosion of the Boeing 747 (TWA Flight 800) over the Atlantic in July 1996 was probably caused by damage to an electrical cable, resulting in a short circuit which allowed high voltage to enter the fuel quantity indication system in the centre fuel tank, which caused arcing which in turn caused a fuel vapour explosion.
6 System failures can be caused by events other than failure of components or sub-systems, such as electromagnetic interference (Chapter 9), operator error, and so on.

These examples illustrate the need for reliability and safety analyses to be performed by engineers with practical knowledge and experience of the system design, manufacture, operation and maintenance.

6.10 Fault Tree Analysis (FTA)

Fault tree analysis (FTA) is a reliability/safety design analysis technique which starts from consideration of system failure effects, referred to as ‘top events’. The analysis proceeds by determining how these can be caused by individual or combined lower level failures or events.

Standard symbols are used in constructing an FTA to describe events and logical connections. These are shown in Figure 6.8. Figure 6.9 shows a simple BDA for a type of aircraft internal combustion engine. There are two ignition systems in an active parallel redundant configuration. The FTA (Figure 6.10) shows that the top event failure to start can be caused by either fuel flow failure, injector failure or ignition failure (three-input OR gate). At a lower level total ignition failure is caused by failure of ignition systems 1 and 2 (two-input AND gate).

In addition to showing the logical connections between failure events in relation to defined top events, FTA can be used to quantify the top event probabilities, in the same way as in block diagram analysis. Failure probabilities derived from the reliability prediction values can be assigned to the failure events, and cut set and tie set methods can be applied to evaluate system failure probability.

Note that a different FTA will have to be constructed for each defined top event which can be caused by different failure modes or different logical connections between failure events. In the engine example, if the top event is ‘unsafe for flight’ then it would be necessary for both ignition systems to be available before take-off, and gate A1 would have to be changed to an OR gate.

The FTA shown is very simple; a representative FTA for a system such as this, showing all component failure modes, or for a large system, such as a flight control system or a chemical process plant, can be very complex and impracticable to draw out and evaluate manually. Computer programs are used for generating and evaluating FTAs. These perform cutset analysis and create the fault-tree graphics. The use of computer programs for FTA provides the same advantages of effectiveness, economy and ease of iterative analysis as described for FMECA. The practical aspects of system reliability modelling described earlier apply equally to FTA.

Since FTA considers multiple, as well as single failure events, the method is an important part of most safety analyses.

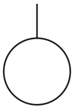
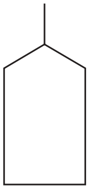
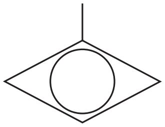
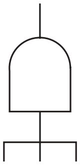
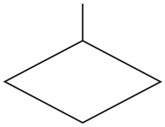
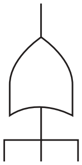
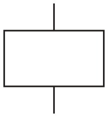
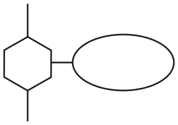
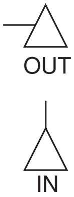
	BASIC EVENT A basic fault event that requires no further development. Is independent of other events.		SWITCH Used to include or exclude parts of the fault tree, which may or may not apply to certain situations.
	BASIC EVENT Is dependent on lower events developed as a separate fault tree.		AND GATE Failure (next higher event) will occur if all inputs fail (parallel redundancy).
	BASIC EVENT Is dependent upon lower events, but not developed downwards.		OR GATE Failure (next higher event) will occur if any input fails (series reliability).
	COMBINATION EVENT An event that results from the combination of basic events through the input logic gates.		INHIBIT GATE INHIBIT gates describe a causal relationship between one fault and another. The input event directly produces the output event if the indicated condition is satisfied.
	TRANSFERRED EVENT A line from the apex of the triangle indicates a transfer in; a line from the side denotes a transfer out.		

Figure 6.8 Standard symbols used in fault tree analysis.

6.11 State-Space Analysis (Markov Analysis)

A system or component can be in one of two states (e.g. failed, non-failed), and we can define the probabilities associated with these states on a discrete or continuous basis, the probability of being in one or other at a future time can be evaluated using *state-space* (or *state-time*) analysis. In reliability and availability analysis, failure probability and the probability of being returned to an available state, failure rate and repair rate, are the variables of interest.

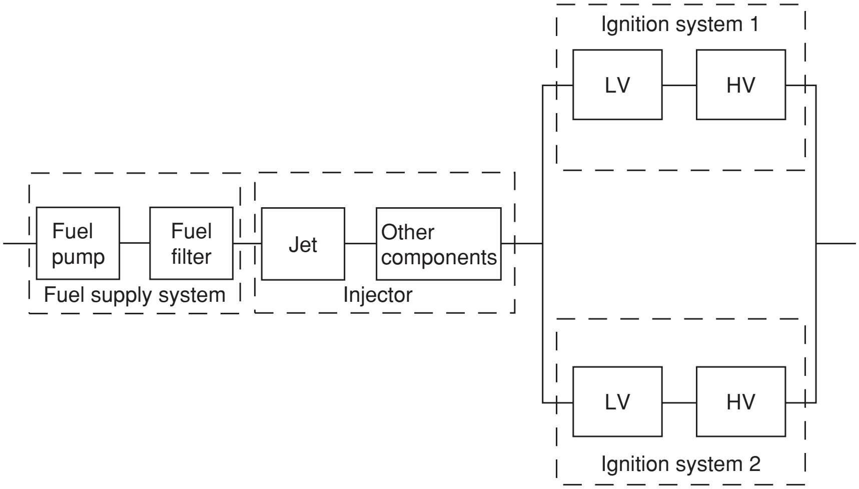


Figure 6.9 *Reliability block diagram of engine.*

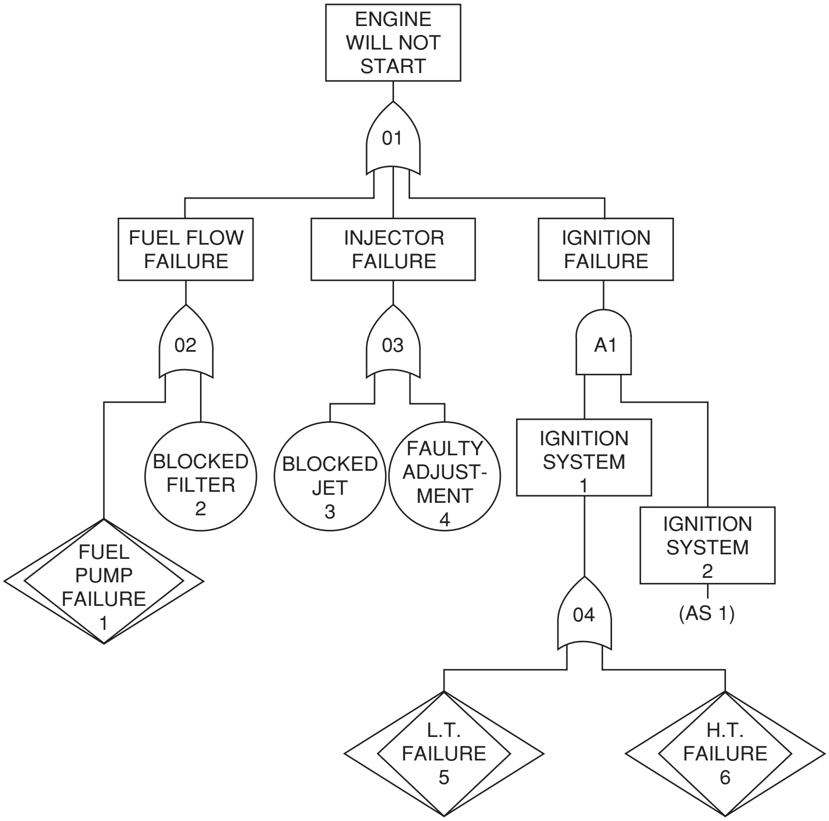


Figure 6.10 *FTA for engine (incomplete).*

The best-known state-space analysis technique is Markov analysis. The Markov method can be applied under the following major constraints:

1 The probabilities of changing from one state to another must remain constant, that is, the process must be homogenous. Thus the method can only be used when a constant hazard or failure rate assumption can be justified.
2 Future states of the system are independent of all past states except the immediately preceding one. This is an important constraint in the analysis of repairable systems, since it implies that repair returns the system to an 'as new' condition.

Nevertheless, the Markov analysis can be usefully applied to system reliability, safety and availability studies, particularly to maintained systems for which BDA is not directly applicable, provided that the constraints described above are not too severe. The method is used for analysing complex systems such as power generation and communications. Computer programs are available for Markov analysis.

The Markov method can be illustrated by considering a single component, which can be in one of two states: failed (F) and available (A). The probability of transition from A to F is $P_{A \to F}$ and from F to A is $P_{F \to A}$. Figure 6.11 shows the situation diagrammatically. This is called a *state transition* or a *state-space diagram*. All states, all transition probabilities and probabilities of remaining in the existing state ($= 1 -$ transition probability) are shown. This is a *discrete* Markov chain, since we can use it to describe the situation from increment to increment of time. Example 6.5 illustrates this.

Example 6.5

The component in Figure 6.11 has transitional probabilities in equal time intervals as follows:

$$P_{\mathrm{A \to F}} = 0.1$$

$$P_{\mathrm{F \to A}} = 0.6$$

What is the probability of being available after four time intervals, assuming that the system is initially available?

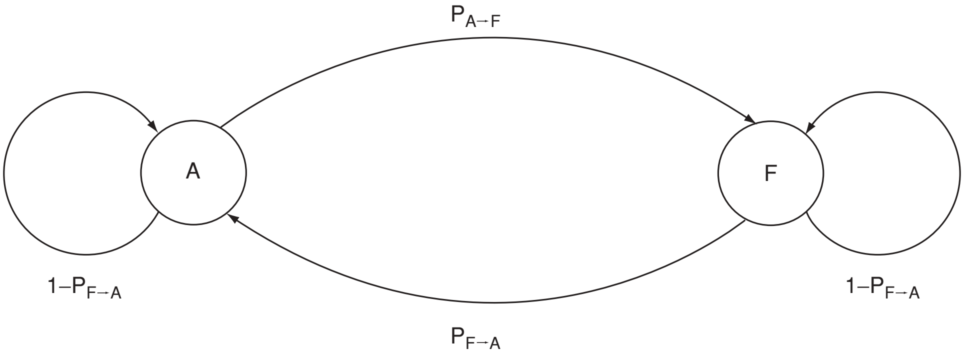


Figure 6.11 *Two-state Markov state transition diagram.*

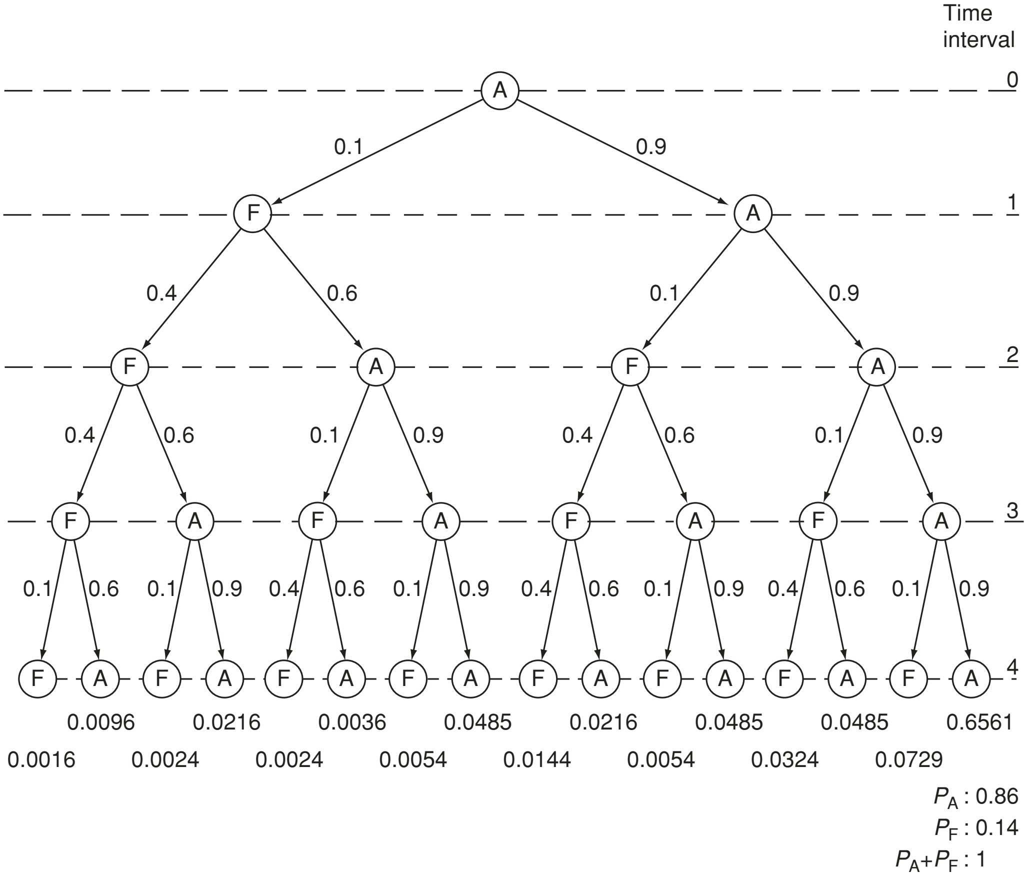


Figure 6.12 *Tree diagram for Example 6.5.*

This problem can be solved by using a *tree diagram* (Figure 6.12).

The availability of the system is shown plotted by time interval in Figure 6.13. Note how availability approaches a steady state after a number of time intervals. This is a necessary conclusion of the underlying assumptions of constant failure and repair rates and of independence of events.

Whilst the transient states will be dependent upon the initial conditions (available or failed), the steady state condition is independent of the initial condition. However, the rate at which the steady state is approached is dependent upon the initial condition and on the transition probabilities.

6.11.1 Complex Systems

The tree diagram approach used above obviously becomes quickly intractable if the system is much more complex than the one-component system described, and analysed over just a few increments. For more complex systems, matrix methods can be used, particularly as these can be readily solved by computer programs. For example, for a single repairable component the probability of being available at the end of any

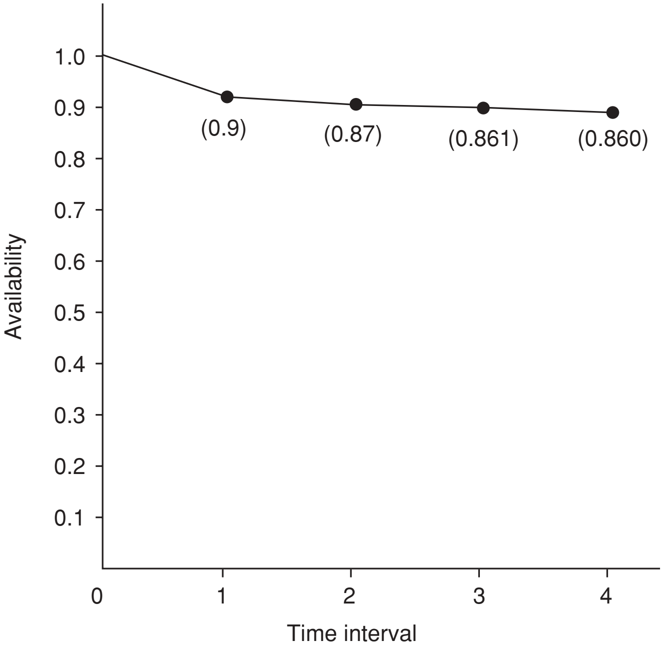


Figure 6.13 *Transient availability of repaired system.*

time interval can be derived using the *stochastic transitional probability matrix*:

$$P = \begin{vmatrix} P_{\mathrm{A}\to\mathrm{A}}\, P_{\mathrm{A}\to\mathrm{F}} \\ P_{\mathrm{F}\to\mathrm{A}}\, P_{\mathrm{F}\to\mathrm{F}} \end{vmatrix} \tag{6.15}$$

The stochastic transitional probability matrix for Example 6.5 is

$$P = \begin{vmatrix} 0.9 & 0.1 \\ 0.6 & 0.4 \end{vmatrix}$$

The probability of being available after the first time increment is given by the first term in the first row (0.9), and the probability of being unavailable by the second term in the first row (0.1). To derive availability after the second time increment, we square the matrix:

$$P^2 = \begin{vmatrix} 0.9 & 0.1 \\ 0.6 & 0.4 \end{vmatrix}^2 = \begin{vmatrix} 0.87 & 0.13 \\ 0.78 & 0.22 \end{vmatrix}$$

The probability of being available at the end of the second increment is given by the first term in the top row of the matrix (0.87). The unavailability = 1 – 0.87 = 0.13 (the second term in the top row).

For the third time increment, we evaluate the third power of the probability matrix, and so on.

Note that the bottom row of the probability matrix raised to the power 1, 2, 3, and so on, gives the probability of being available (first term) or failed (second term) if the system started from the failed state. The reader is invited to repeat the tree diagram (Figure 6.12), starting from the failed state, to corroborate this. Note also that the rows always summate to 1; that is, the total probability of all states. (Revision notes on simple matrix algebra are given in Appendix 7.)

If a system has more than two states (multi-component or redundant systems), then the stochastic transitional probability matrix will have more than 2 × 2 elements. For example, for a two-component system, the states could be:

	Component		
State	1	2	
1	A	A	
2	A	Ā	
3	Ā	A	A: available
4	Ā	Ā	Ā: unavailable

The probabilities of moving from any one state to any other can be shown on a 4 × 4 matrix. If the transition probabilities are the same as for the previous example, then:

$$P_{1\to 1} = 0.9 \times 0.9 = 0.81$$

$$P_{1\to 2} = 0.9 \times 0.1 = 0.09$$

$$P_{1\to 3} = 0.1 \times 0.9 = 0.09$$

$$P_{1\to 4} = 0.1 \times 0.1 = 0.01$$

$$P_{2\to 1} = 0.9 \times 0.6 = 0.54$$

$$P_{2\to 2} = 0.9 \times 0.4 = 0.36$$

$$P_{2\to 3} = 0.1 \times 0.6 = 0.06$$

$$P_{2\to 4} = 0.1 \times 0.4 = 0.04$$

$$P_{3\to 1} = 0.6 \times 0.9 = 0.54$$

$$P_{3\to 1} = 0.6 \times 0.9 = 0.54$$

$$P_{3\to 2} = 0.6 \times 0.1 = 0.06$$

$$P_{3\to 3} = 0.4 \times 0.9 = 0.36$$

$$P_{3\to 4} = 0.4 \times 0.1 = 0.04$$

$$P_{4\to 1} = 0.6 \times 0.6 = 0.36$$

$$P_{4\to 2} = 0.6 \times 0.4 = 0.24$$

$$P_{4\to 3} = 0.4 \times 0.6 = 0.24$$

$$P_{4\to 4} = 0.4 \times 0.4 = 0.16$$

and the probability matrix is

$$P = \begin{vmatrix} P_{1\to 1} & P_{1\to 2} & P_{1\to 3} & P_{1\to 4} \\ P_{2\to 1} & P_{2\to 2} & \cdots & \\ P_{3\to 1} & \vdots & \vdots & \vdots \\ P_{4\to 1} & \cdots & \cdots & P_{4\to 4} \end{vmatrix}$$

$$= \begin{vmatrix} 0.81 & 0.09 & 0.09 & 0.01 \\ 0.54 & 0.36 & 0.06 & 0.04 \\ 0.54 & 0.06 & 0.36 & 0.04 \\ 0.36 & 0.24 & 0.24 & 0.16 \end{vmatrix}$$

The first two terms in the first row give the probability of being available and unavailable after the first time increment, given that the system was available at the start. The availability after 2, 3, and so on, intervals can be derived from p^2, p^3, . . . , as above.

It is easy to see how, even for quite simple systems, the matrix algebra quickly diverges in complexity. However, computer programs can easily handle the evaluation of large matrices, so this type of analysis is feasible in the appropriate circumstances.

6.11.2 Continuous Markov Processes

So far we have considered discrete Markov processes. We can also use the Markov method to evaluate the availability of systems in which the failure rate and the repair rate (λ, μ) are assumed to be constant in a time continuum. The state transition diagram for a single repairable item is shown in Figure 6.14.

In the steady state, the stochastic transitional probability matrix is:

$$P = \begin{vmatrix} 1-\lambda & \lambda \\ \mu & 1-\mu \end{vmatrix} \tag{6.16}$$

The instantaneous availability, before the steady state has been reached, can be derived using Eq. (6.11).

The methods described in the previous section can be applied for evaluating more complex, continuous Markov chains. The Markov analysis can also be used for availability analysis, taking account of the holdings and repair rate of spares. The reader should refer to Singh and Billinton (1977) and Pukite and Pukite (1998) for details of the Markov method as applied to more complex systems.

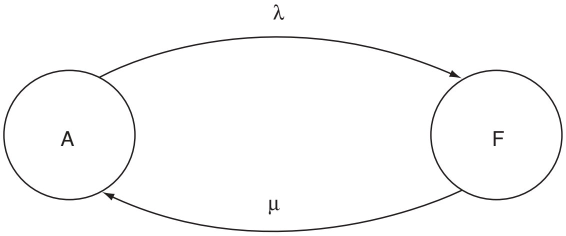


Figure 6.14 State–space diagram for a single-component repairable system.

6.11.3 Limitations, Advantages and Applications of Markov Analysis

The Markov analysis method suffers one major disadvantage. As mentioned earlier, it is necessary to assume constant probabilities or rates for all occurrences (failures and repairs). It is also necessary to assume that events are statistically independent. These assumptions are hardly ever valid in real life, as explained in Chapter 2 and earlier in this chapter. The extent to which they might affect the situation should be carefully considered when evaluating the results of a Markov analysis of a system.

The Markov analysis requires knowledge of matrix operations. This can result in difficulties in communicating the methods and the results to people other than reliability specialists. The severe simplifying assumptions can also affect the credibility of the results.

The Markov analysis is fast when run on computers and is therefore economical once the inputs have been prepared. The method is used for analysing systems such as power distribution networks and logistic systems.

6.12 Petri Nets

A further expansion of state-space analysis techniques came with development of Petri nets, which were introduced by Carl Adam Petri in 1962. A Petri net is a general-purpose graphical and mathematical tool for describing relations existing between conditions and events. The original Petri net did not include the concept of time, so that an enabled transition fires immediately (see Section 6.12.2). An extension called *stochastic Petri nets* (SPN) or *time Petri nets* was introduced in the late 1970s. Stochastic Petri nets address most of the shortcoming of Markov chains by focusing on modelling the states of components that comprise the system, so that the state of the system can be inferred from the states of its components rather than defined explicitly as required by the Markov approach.

SPN is often used as a modelling preprocessor, so the model is internally converted to Markov state space and solved using Markov methods. However, as mentioned before, the major disadvantage of the Markov method is its reliance on constant rates of all occurrences (Poisson process). In order to overcome that disadvantage and solve SPN directly, the Monte-Carlo method is often used to simulate the transition process. The original Petri nets are still used in software design, while reliability engineering uses mostly time Petri nets.

The basic symbols of Petri nets include:

○ : *Place*, drawn as a circle, denotes event.

— : *Immediate transition*, drawn as a thin bar, denotes event transfer with no delay time.

▬ : *Timed transition*, drawn as a thick bar, denotes event transfer with a period of delay time.

↷ : *Arc*, drawn as an arrow, between places and transitions.

● : *Token*, drawn as a dot, contained in places, denotes the data and also serves as an indicator of system's state.

⊸ : *Inhibitor arc*, drawn as a line with a circle end, between places and transitions.

The transition is said to *fire* if input places satisfy an enabled condition. Transition firing will remove one token from each of its input places and put one token into all of its output places. Basic structures of logic relations for Petri nets are listed in Figure 6.15, where there are two types of input places for the transition; namely, specified type and conditional type. The former one has single output arc whereas the

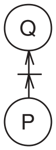
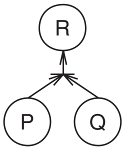
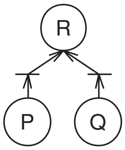
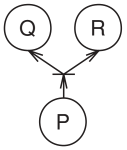
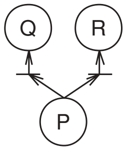
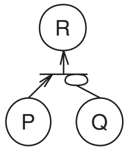
Logic relation	TRANSFER	AND	OR	TRANSFER AND	TRANSFER OR	INHIBITION
Description	If P then Q	If P AND Q then R	If P OR Q then R	If P then Q AND R	If P then Q OR R	If P AND Q' then R
Boolean function	Q=P	R=P*Q	R=P+Q	Q=R=P	Q+R=P	R=P*Q'
Petri nets						

Figure 6.15 *Basic structures of logic relations for Petri nets (Courtesy S. Yang).*

latter has multiples. Tokens in the specified type place have only one outgoing destination, that is if the input place(s) holds a token then the transition fires and gives the output place(s) a token. However, tokens in the conditional type place have more than one outgoing paths that may lead the system to different situations. For the 'TRANSFER OR' Petri nets in Figure 6.15, whether Q or R takes over a token from P depends on conditions, such as probability, extra action, or self-condition of the place.

There are three types of transitions that are classified based on time. Transitions with no time delay due to transition are called immediate transitions, while those needing a certain constant period of time for transition are called timed transitions. The third type is called a stochastic transition: it is used for modelling a process with random time. Owing to the variety of logical relations that can be represented with Petri nets, it is a powerful tool for modelling systems. Petri nets can be used not only for simulation, reliability analysis and failure monitoring, but also for dynamic behaviour observation. This greatly helps fault tracing and failure state analysis. Moreover, the use of Petri nets can improve the dialogue between analysts and designers of a system.

6.12.1 Transformation between Fault Trees and Petri Nets

Figure 6.16 is a fault tree example in which events A, B, C, D and E are basic causes of event 0. The logic relations between the events are described as well. The correlations between the fault tree and Petri net are shown in Figure 6.17.

Figure 6.18 is the Petri net transformation of Figure 6.16.

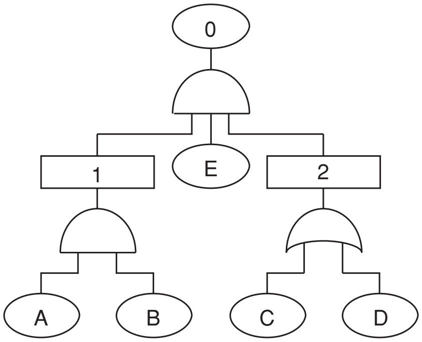


Figure 6.16 *A fault tree (Courtesy S. Yang).*

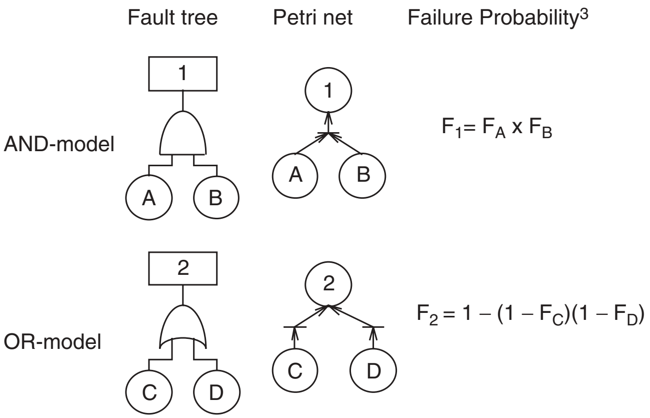


Figure 6.17 Correlations between fault tree and Petri net (Courtesy S. Yang).

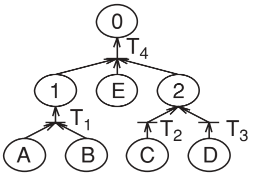


Figure 6.18 The Petri net transformation of Figure 6.16 (Courtesy S. Yang).

6.12.2 Minimum Cut Sets

To identify the minimum cut sets in a Petri net a matrix method is used, as follows:

1 Put down the numbers of the input places in a row if the output place is connected by multi-arcs from transitions. This accounts for OR-models.
2 If the output place is connected by one arc from a transition then the numbers of the input places should be put down in a column. This accounts for AND-models.
3 The common entry located in rows is the entry shared by each row.
4 Starting from the top event down to the basic events until all places are replaced by basic events, the matrix is thus formed, called the *basic event matrix*. The column vectors of the matrix constitute cut sets.
5 Remove the supersets from the basic event matrix and the remaining column vectors become minimum cut sets.

For example, the basic event matrix for Figure 6.18 is shown in Figure 6.19.

	1	E	2	2
	A		C	D
	B			
Minimum cut sets:	↑	↑	↑	↑

Figure 6.19 Minimum cut sets of Figure 6.18.

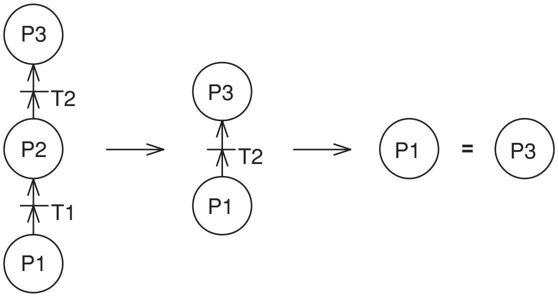


Figure 6.20 *The absorption principle of equivalent Petri nets (Courtesy S. Yang).*

Minimum cut sets can be derived in an opposite, bottom-up, direction, that is from basic places to the top place. Transitions with T = 0 are called *immediate transitions*. In a Petri net with immediate transitions, that is, the token transfers between places do not take time, then it can be absorbed to a simplified form called the *equivalent Petri net*.

Figure 6.20 shows the principle of absorption. After absorption, all the remaining places are basic events. The equivalent Petri net exactly constitutes the minimum cut sets, that is the input of each transition represents a minimum cut set.

6.12.3 Marking Transfer

A *marking* of a Petri net is defined as the total number of tokens at each place, denoted by a column vector M. Thus vector $M_k = (n_1, n_2, \ldots, n_m)$ represents that token numbers of places $P_1, P_2, \ldots, P_m$ at state k are $n_1, n_2, \ldots, n_m$, respectively. Consequently, Petri nets can be expressed in state space form which gives the next state M_{k+1} from its previous state M_k:

$$M_{k+1} = \mathbf{A}M_k + \mathbf{B}U_k, \quad k = 1, 2, \ldots \tag{6.17}$$

where M_k is the marking at state k, $a\ m \times 1$ column vector, U_k is an input vector at state k, and **A** and **B** are matrix coefficients.

Since vector M_k represents the marking in a Petri net at state k, the failure state of a system may vary with time. Hence, the markings of a Petri net depend on time dynamically. The dynamic behaviour of system failure is defined as the system failure state with time varied, and is determined by the movement of tokens in the Petri net model. Thus, the dynamic behaviour of a system failure can be investigated by Petri nets whereas it cannot be done by fault trees. The method can be extended to include failure detection and correction logic, and repair and delay times. Since Petri net analysis is an emerging technical field, most of the software packages used for creating and analysing Petri nets are of a proprietary nature. Yakovlev *et al.* (2000) provides an introduction to Petri nets, and some applications have been described, for example in Yang and Liu (1997), Yang and Liu (1998) and Kleyner and Volovoi (2008).

Example 6.6

Let us consider an automotive airbag with a fault detection capability. When an airbag controller fails during regular driving conditions (no accident involved), a fault detection system detects that failure and turns the warning light on to inform the driver.

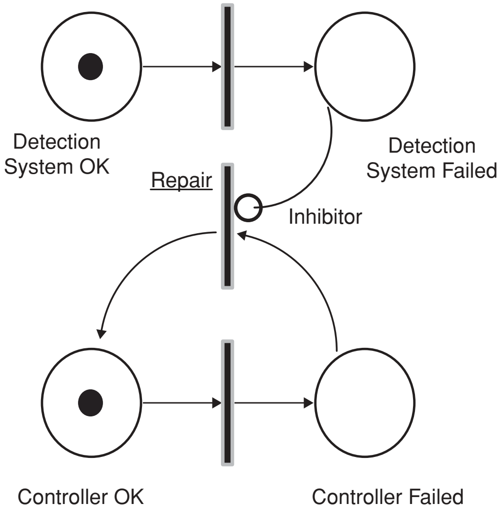


Figure 6.21 *Petri net diagram for an airbag controller with a detection system (Kleyner and Volovoi, 2008).*

Figure 6.21 depicts an airbag controller (bottom two places) and its detection system (top two places). When the controller fails, the bottom token moves to the right from the 'Controller OK' to 'Controller Failed' place. If the detection system functions properly, failure is detected by a driver, then repaired and the system is returned to the 'Controller OK' place via the 'Repair' transition in the middle. However if at some point the detection system fails, the top token moves from the 'Detection System OK' to 'Detection System Failed' place, which initiates an inhibitor disabling 'Repair' transition. Therefore if the controller fails now, the inhibitor will prevent the system from informing the driver about the failure and the controller will not be repaired. If the rates of each transition occurrence are constant, this system can be modelled as a Markov process. However if the rates are not constant (e.g. controller primary failures follow Weibull distribution) then the system should be modelled using Monte Carlo or other type of stochastic simulation.

6.13 Reliability Apportionment

Sometimes it is necessary to break an overall system reliability requirement down to individual subsystem reliabilities. This is common in large systems, particularly when different design teams or subcontractors are involved. The main contractor or system design team leader requires early assurance that subsystems will have reliabilities which will match the system requirement, and therefore the appropriate values have to be included in the subsystem specifications. This activity can be considered as a reliability 'budgeting'.

The starting point for reliability apportionment is a reliability block diagram for the system drawn to show the appropriate system structure. The system requirement is then broken down in proportions which take account of the complexity, risk and existing experience related to each block. It is important to take account of the uncertainty inherent in such an early prediction, and therefore the block reliabilities need not aggregate to the system requirement, but to some higher reliability. The apportionment and specifications derived from it should take account of different operating conditions of subsystems. For example, a radar system might have a subsystem which operates for only half of the total system operating time, and therefore this should be shown on the RBD, and the failure rate apportioned to it should be related clearly either to the operating time of the system or of the subsystem.

6.14 Conclusions

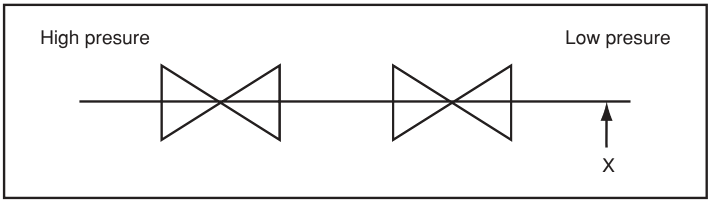
System reliability (and safety and availability) prediction and modelling can be a frustrating exercise, since even quite simple systems can lead to complex reliability logic when redundancy, repair times, testing and monitoring are taken into account. On the other hand, the parameters used, particularly reliability values, are usually very uncertain, and can be highly variable between similar systems. This chapter has described some approaches, but it must be realized that the results can be very sensitive to parameter changes. For example, a single common-cause failure, overlooked in the analysis, might have a probability of occurrence which would completely invalidate the reliability calculations for a high reliability redundant system. In real life, availability is often determined more by spares holdings, administrative times (transport, documentation, delays, etc.) than by ‘predictable’ factors such as mean repair time. Therefore predictions and models of system reliability and availability should be used as a form of design review, to provide a disciplined framework for considering factors which will affect reliability and availability, and the sensitivity to changes in assumptions. Critical aspects can then be highlighted for further attention, and alternative system approaches can be compared.

Prediction and modelling are concepts which have generated much attention, literature and controversy in the reliability field. Considerable effort has been expended on the development and updating of databases and models, and a large proportion of the journal articles and conference contributions is devoted to the topics. Much of this work is of mainly academic interest, with limited practical applicability. The mathematical techniques described in this chapter are useful only in so far as the values inserted into the formulae are known within reasonably close limits. The use of complicated formulae to analyse the effects of parameters which are highly uncertain is inefficient and potentially misleading.

Finally, the predictions must be based on the commitment by the project management, and on a realistic appreciation of the technologies and risks.

Questions

1. Assume that you are responsible for the reliability aspects of a system containing both electronic and mechanical elements. The customer for the system requires that a numerical reliability prediction be provided:
   a Describe what is meant by a ‘reliability prediction’ in this context.
   b Identify some sources of data that can be used to assist in quantifying the prediction, and discuss the dangers that have to be guarded against in the use of such data.
   c Do you expect the prediction to overestimate or underestimate the reliability that could be achieved by the system? Give your reasons.
2. Two resistors are connected in parallel in an electrical circuit. They can fail either open-circuit (resistance $= \infty$) or short-circuit (resistance $= 0$). Draw reliability block diagrams for the pair of resistors (i) for the circuit failing open-circuit; (ii) for the circuit failing-short circuit.
3. The system sketched below is used to regulate the downstream pressure of gas in a chemical plant. There are two regulators whose function is to keep the downstream pressure at a constant value. The upstream pressure fluctuates, but is always much higher than the required downstream pressure. The regulators function independently, but both sense the downstream pressure at the same point (X).



A regulator can fail in one of two ways: ‘open’, in which case it becomes ‘straight through’ so that the flow of gas is unrestricted, and reliability in this mode is R_0; or ‘closed’, in which it totally blocks the flow of gas and reliability in this mode is R_c.

a On the assumption that the two types of failure are independent, produce expressions for system reliability (i) for the system failing due to total loss of flow; (ii) for the system failing due to overpressure downstream.

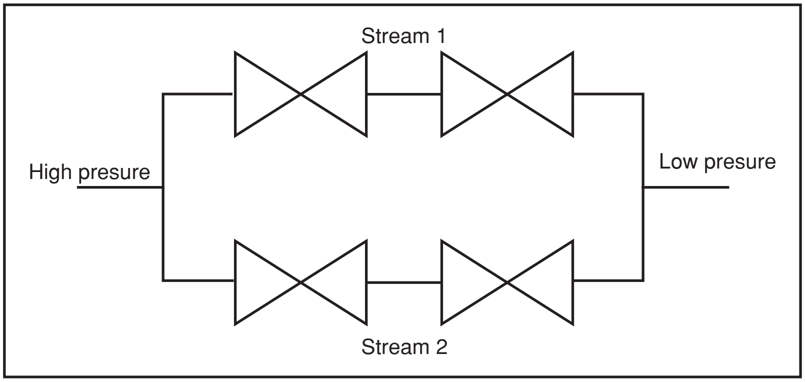
b The times to failure of the regulators are described by the following distributions: Closed – exponential with mean life 2 years.

Open – Weibull with $\beta = 1.8$ and $\eta = 1.6$ years.

What are the probabilities of the system giving one year of failure-free operation in each mode?

c What is the probability of obtaining one year of failure-free operation irrespective of the mode of failure?

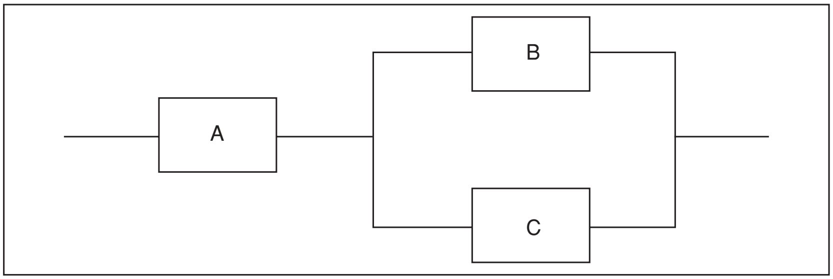
4. It has been suggested that the reliability of the system in question 3 can be improved by adopting a twin-stream system as in the diagram below.



Calculate the two system reliabilities for (loss of flow and overpressure) for this configuration.

5. Draw block diagrams for both modes of failure of the ‘twin stream’ in question 4. Draw the cut sets and tie sets in each case.

6. Calculate the reliability over a 150 h mission for the system whose reliability block diagram is shown below.



The probability distributions of lives to failure of the elements are:

A constant hazard, mean life = 837 h.

B Weibull, with shape parameter = 2.0, characteristic life 315 h.

C normal, with mean = 420 h and standard deviation 133 h.

7. An element in a manufacturing system has a mean time between failures of 25 h. When it fails it takes, on average, 2 h to restore it to an operating condition. It has been suggested that the problems caused by the unreliability of this element could be avoided by installing a second identical element as a standby:
   a Calculate the probability of completing an 8 h shift without a system failure both for the single-element and for the redundant-element system.
   b If a repair team is made available such that repairs can be started on a failed element immediately after failure, calculate the long-term system availability (assuming continuous operation) for both the single-element and the redundant-element system. State all assumptions made and comment on whether you consider them reasonable.
   c If the additional element costs £25 000 and downtime costs £100 per hour, what is the payback period for the additional element? (Ignore discounting of cash flows.)
8. In the context of fault-tree analysis, explain the meaning of each of the following:

an AND gate	'top' event
an OR gate	a basic event
a priority AND gate	an undeveloped event.

   In each case, sketch the conventional symbol used and give a practical example.
9. A control system consists of an electrical power supply, a standby battery supply which is activated by a sensor and switch if the main supply fails, a hydraulic power pack, a controller, and two actuators acting in parallel (i.e. control exists if either or both actuators are functioning).
   a Draw the system reliability block diagram.
   b Draw the fault tree appropriate to the top event 'total loss of actuator control'.
10. In question 9, if the reliabilities of the separate components are as follows:

Main electrical supply:	0.99
Standby battery supply:	0.995
Sensor and switch:	0.995
Hydraulic supply:	0.95
Controller:	0.98
Actuator (each):	0.99

   What is the system reliability for the top event 'total loss of actuator control'?
11. This is a very simple example of simulation to explore the sort of principles involved in much more complex (and realistic) situations. It describes the common 'queuing' problem, but by using simulation we are able to circumvent the convenient, but often implausible assumptions that are usually incorporated in conventional 'queuing theory'. In particular, we do not need to assume exponentially distributed arrival and service times.

   A radio taxi service operates between 9 a.m. and 7 p.m. Calls arrive at random at a rate of three per hour. The time to reach a customer is a Weibull distribution with $\beta = 2$, $\eta = 0.25$ h. The journey time is normally distributed with mean 30 minutes and standard deviation 6 minutes. A customer will only be offered a service if there is a taxi available, that is not currently on a journey.

   You are invited to do one 'manual' simulation of a day's operation with a fleet of two taxis, and evaluate the reliability (i.e. proportion of customers picked up). The calculations are simple but tedious, so computer methods are used when we want many thousands of simulations of much more complex situations. You will need to generate random sample values for:

   t_1 – the time of a customer call since the previous one.
   t_2 – the time to reach the customer.
   t_3 – the journey time with the customer.

These will be developed from random numbers between 0 and 1. Such numbers can be generated in various ways; e.g:

– doing Monte Carlo simulations (see Chapter 4).
– using the ran# function on a calculator.
– rolling a 20-sided die.
– using tables of random numbers.
– using the random number generator built into common application software (e.g. most spreadsheets).

Simulating t_1:

The reliability function (probability that the time to reach a customer exceeds $t = t_1$) is from the exponential distribution, that is $= \exp(-\lambda t)$, from which $t = 1/\lambda \ln [1/R(t)]$. Simply generate a random number between 0 and 1 for $R(t)$ and calculate the value of t using $\lambda = 1/3$. For example, for a random number of 0.439, the random time is $3 \ln (1/0.439) = 0.823$ h.

If running Monte Carlo simulation using Excel®, use the functions in Chapter 4, Table 4.1

Simulating t_2:

This time will be simulated from a Weibull distribution, where

$$R(t) = \exp[-(t/\eta)^{\beta}]$$

from which

$$t = \eta\{\ln[1/R(t)]\}^{1/\beta}$$

We have $\beta = 2$ and $\eta = 0.25$; so, for example, with a random number of 0.772, the simulated value of t_2 is

$$0.25\{\ln[1/0.772]\}^{0.5} = 0.127\,\text{h}$$

If running Monte Carlo simulation using Excel®, use the functions in Chapter 4, Table 4.1

Simulating t_3:

This comes from a normal distribution. There is no closed form for the reliability function of this distribution, so some alternative approach is necessary. The easiest is to use tables of standardized random normal deviates (as included in most statistics books) or an implementation in a computer spreadsheet. Suppose we obtained a value of -0.194. For our particular normal distribution the mean is 0.5 h and the standard deviation 0.1 h, so the simulated random time is

$$0.5 + (-0.194 \times 0.1) = 0.481\,\text{h}.$$

If running Monte Carlo simulation using Excel®, use the functions in Chapter 4, Table 4.1

Using these ideas, ‘walk through’ a random sample of one day of operation of two taxis. See how many requests were made, and how many were delivered to their destinations.

12. A device used in a ground radar system has age to failure that is described approximately by a Weibull distribution with mean life 83 h, shape parameter 1.5, and location parameter zero. When it fails it takes on average 3.5 h to repair:
    a Calculate the reliability over a 25 h period, and the ‘steady state’ availability of the device.
    b Calculate the reliability over 25 h, and the ‘steady state’ availability of a subsystem that consists of two of these devices in active parallel redundancy.
    c Identify all assumptions made in these calculations, and discuss their validity.

d Explain the meaning of the ‘steady state availability’ in (a) and (b) above, and consider whether it gives the most suitable measure of availability in this example.

13. Describe the practical limitations of using methods such as BDA, FTA, Markov chains, and so on for assessing the reliability, in qualitative and in quantitative terms, of the following (Hint: in each case, first define the failure you are considering):
    a A parachute (consider the canopy, the lines and the deployment mechanism).
    b A microprocessor (consider the power input, the data input and output connections, and every transistor and capacitor).
    c An electric motor.
    d A mechanical assembly that consists of static parts bolted together.
    e A train.
14. Calculate the failure rate for a generic ceramic capacitor 0.2 μF (circuit resistance 0.5 /Ω/V) operating at the ambient temperature T = 60°C in the ground mobile environment. The capacitor has two pins, hermetic package and is rated to operate at the maximum temperature of 100°C. Compare the results obtained using MIL-HDBK-217 with and other method such as Telcordia or IEC 62380.
15. Draw a Markov state-space diagram and a Petri net for a standby system covered in Section 6.6.4. Consider both perfect and imperfect switching.
16. At the early design stage you need to evaluate the future reliability of the product. Which reliability prediction method would you choose if you have:
    a A bill of materials for an electronic system.
    b Mechanical drawings of the system.
    c Bill of materials and the knowledge of the stress factors for the electronic components, such as temperature, vibration, voltage, and so on.
    d Electronic schematics.
    e List of mechanical parts and the detailed drawings.
    f Detailed information about device geometry, material properties and the applied stresses.
    g Field return and warranty data for the previous model of the system with clear description of the differences between the models.
    h Field return and warranty data for the previous model of the system with clear description of the differences in the environmental stresses between the two models.
17. Develop a Fault Tree Diagram a coffee maker, a hand calculator, a digital camera, a water heater, a lawnmower, a microwave oven, an electronic thermostat, TV remote control, a vacuum cleaner or other every day product. Provide at least three levels of the FTA tree depth counting the top event.
18. The system XYZ can be divided into subsystems and be represented by the block diagram in Figure 6.22.

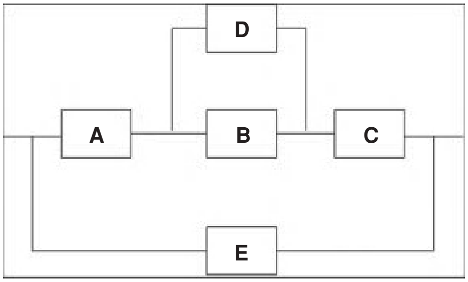


Figure 6.22 System XYZ block diagram.

The failures of each subsystem are distributed according to the following statistical distributions:

A – Exponential $\lambda = 12.0$ failures per million hours.
B – Weibull, $\beta = 2.0$, $\eta = 20\,000$ hours.
C – Normal, $\mu = 1600$ hours, $\sigma = 100$ hours.
D – Lognormal, $\mu = 24$, $\sigma = 12$.
E – Weibull, $\beta = 0.8$, $\eta = 10\,000$ hours.

Calculate the system reliability at 1200 hours.

Calculate the system reliability at 1600 hours.

For the Blocks A, B and E determine which portion of the bathtub curve best describes each subsystem? Explain why.

19. If a system consists of four major serial subsystems and the required reliability is 0.9, what is the allocated reliability of each subsystem, assuming equal apportionment?
20. Compare the reliability of three parts at 10 years mark:
    Part A (R_A): Failures are exponentially distributed with failure rate of 0.58 failures per million hours. Consider 24 hrs per day operation.
    Part B (R_B): B_{10} life equals 10 years.
    Part C (R_C): Failures are distributed according to Weibull distribution with $\beta = 2.0$ and $\eta = 20.0$ years.
21. Two identical and independent components form an active redundant (parallel) system. Each component's time to failure follows Weibull distribution with the parameters β and η. Derive the analytical expression for this system's hazard rate h(t).
22. Develop a fault tree and analyse a car's turn signal system whose failure could be caused by a battery failure ($P_B = 0.11$), broken connector ($P_C = 0.05$) or burnt out light bulb ($P_L = 0.21$). What is the probability of failure for the whole system?
23. A system with standby redundancy and perfect switching has a failure rate of 0.006 failures per hour for each subsystem: primary and standby. What is the reliability of the whole system at 60 hours?

Bibliography

Alion System Reliability Center (2011) *PRISM*. Available at: http://src.alionscience.com/prism/ (Accessed 22 March 2011)

Andrews, J. and Moss, T. (2002) *Reliability and Risk Assessment*, ASME Press.

British Standard, BS 5760 (1991) *Reliability of Systems, Equipments and Components*, British Standards Institution, London.

British Telecom (1995) HRD5, *Handbook of Reliability Data for Electronic Components used in Telecommunication Systems*.

CALCE (2011) Center for Advanced Life Cycle Engineering. Available at http://www.calce.umd.edu/ (Accessed 22 March 2011).

Dylis, D (2001) PRISM: A new approach to reliability prediction. Amer. Soc. Quality (ASQ) Reliability Review, **21**(1), March issue.

Dylis, D. and Priore, M. (2001) *A Comprehensive Reliability Assessment Tool for Electronic Systems*, IIT Research/Reliability Analysis Center, Rome NY, Available at: http://www.theriac.org/productsandservices/products/217plus/rams01.pdf

FIDES (2009) *FIDES Guide 2009 Issue A*. Available at http://fides-reliability.org/ (Accessed 22 March 2011)

Foucher, B., Boullié, J.B., Meslet, B. and Das, D. (2002) A review of reliability prediction methods for electronic devices. *Microelectronics Reliability*, **42**(8), 1155–1162.

Høyland, A., Rausand, M. (2004) *System Reliability Theory, Models and Statistical Methods*, 2nd edn, Wiley.

IEEE Standard 1413.1-2002 (2003) *IEEE Guide for Selecting and Using Reliability Predictions Based on IEEE 1413*, IEEE.

Kaminskiy, M. and Krivtsov, V. (2000) G-Renewal Process as a Model for Statistical Warranty Claim Prediction. Proceedings of the Annual Reliability and Maintainability Symposium (RAMS).

Kleyner, A. and Volovoi, V. (2008) Reliability Prediction using Petri Nets for On-Demand Safety Systems with Fault Detection, in *Safety and Reliability and Risk Analysis* (eds Martorell, Soares and Barnett), (European Safety and Reliability conference), v.3, pp. 1961–1968.

Kleyner, A. and Bender, M. (2003) Reliability Prediction Method Based on Merging Military Standards Approach with Manufacturer's Warranty Data. Proceedings of Annual Reliability and Maintainability Symposium (RAMS), Tampa, Florida, pp. 202–206.

Kleyner, A. and Boyle, J. (2003) Reliability Prediction of Substitute Parts Based on Component Temperature Rating and Limited Accelerated Test Data. Proceedings of Annual Reliability and Maintainability Symposium, Tampa, Florida, pp. 518–522.

McLeish, J. (2010) Enhancing MIL-HDBK-217 Reliability Predictions with Physics of Failure Methods. Proceedings of the Annual Reliability and Maintainability Symposium (RAMS).

MIL-HDBK-217F (1995) Military Handbook, Reliability Prediction of Electronic Equipment, Notice 2, Department of Defense, Washington, DC.

Nicholls, D. (2007) What is 217Plus$^{\text{TM}}$ and Where Did it Come From? Proceedings of the Annual Reliability and Maintainability Symposium (RAMS).

Nippon Telegraph and Telephone Corporation (1985) *Standard Reliability Table for Semiconductor Devices*.

NSWC-06/LE10 (2006) *Handbook of Reliability Prediction Procedures for Mechanical Equipment*, Edition 2006. US Naval Surface Warfare Center, Carderock Division, West Bethesda, Maryland 20 817-5700. Available at: http://www.everyspec.com/USN/NSWC/NSWC-06_RELIAB_HDBK_2006_15 051/.

Pukite J. and Pukite P. (1998) *Markov Modelling for Reliability Analysis*, IEEE Press.

RAIC (2010) *Electronic Reliability Prediction*. Available at http://www.theriac.org/DeskReference/viewDocument.php?id = 211 (Accessed 22 March 2011).

RAIC (2011) *217Plus: RAIC's Reliability Prediction Methodology*. Available at http://www.theriac.org/productsandservices/products/217plus/ (Accessed 22 March 2011).

ReliaSoft (2006) *Standards Based Reliability Prediction in a Nutshell*. Reliability Hotwire Issue 70, http://www.weibull.com/hotwire/issue70/relbasics70.htm (Accessed 22 March 2011).

Singh, C. and Billinton, R. (1977) *System Reliability Modeling and Evaluation*, Hutchinson.

Talmor, M. and Arueti, S. (1997) Reliability Prediction: The Turn-Over Point. Proceedings of Reliability and Maintainability Symposium (RAMS).

Telcordia Special Report SR-332 (2001) *Reliability Prediction Procedure for Electronic Equipment (RPP)*. Issue 1, Telcordia Customer Service, Piscataway, NJ.

UTEC80810 (2000) *Modele universel pour le calcul de la fiabileite previsionnelle des composants, cartes et equipememts electroniques* – RDF2000.

Yakovlev, A., Gomes, L. and Lavagno, L. (2007) *Hardware Design and Petri Nets*, Kluwer Academic Publishers.

Yang, S. and Liu, T. (1997) Failure Analysis for an Airbag Inflator by Petri Nets. *Quality and Reliability Engineering International*, **13**, 139–151.

Yang, S. and Liu, T. (1998) A Petri Net Approach to Early Failure Detection and Isolation for Preventive Maintenance. *Quality and Reliability Engineering International*, **14**, 319–330.

7
Design for Reliability

7.1 Introduction

The reliability of a product is strongly influenced by decisions made during the design process. Deficiencies in design affect all items produced and are progressively more expensive to correct as development proceeds. The cost of errors and design changes increases drastically over the course of the product development cycle, Figure 7.1. It can be as high as a 10-fold increase in cost from one phase to another. It is therefore essential that design disciplines are used which minimize the possibility of failure and which allow design deficiencies to be detected and corrected as early as possible. In Chapter 5 the basic requirements for failure-free design were laid down, that is adequate safety margins, protection against extreme load events and protection against strength degradation. The design must also take account of all other factors that can affect reliability, such as production methods, use and maintenance, and failures not caused by load.

The design process must therefore be organized to ensure that failure-free design principles are used and that any deviations from the principles are detected and corrected. The designers must aim to create designs which will not fail if manufactured and used as specified.

The old design concept of 'test-analyze-and-fix' (TAAF), in which reliance is placed on the test programme to show up reliability problems, no longer has a place in modern design and manufacturing due to shorter design cycles, relentless cost reduction, warranty cost concerns, and many other considerations. Therefore the reliability should be 'designed-in' to the product using the best available science-based methods. This process, called Design for Reliability (DfR) begins from the first stages of product development and should be well integrated through all its phases.

The DfR process also changes the engineers' roles in the design process. The role of the reliability engineer is changing into the mentor (Silverman, 2010), who is now responsible for finding the best design methods and techniques for reliability and then for training the designers on how to use them. As a result of this process the designers take ownership of the reliability of the product, while the reliability organization serves as a steering committee. Needless to say in order to achieve this, the reliability engineer should be integrated with the design team from the very first step of the DfR process. Therefore, the DfR process ensures that pursuit of reliability is an enterprise-wide activity.

Practical Reliability Engineering, Fifth Edition. Patrick D. T. O'Connor and Andre Kleyner.

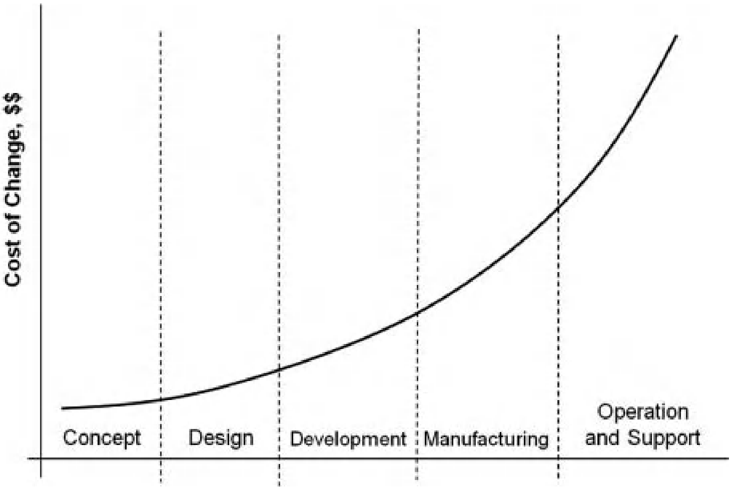


Figure 7.1 *Cost of design change.*

There have been tendencies to parallel DfR with *Design for Six Sigma* (DFSS) (for more on DFSS see Creveling *et al*. 2003). The two processes have similarities and share some of the methods and techniques. However, there are clear distinctions between the two. DFSS aims at reducing variations throughout the design process largely in order to avoid manufacturing problems, while DfR focuses on designing reliability into the product. For more on the distinctions between DFSS and DfR see ReliaSoft (2008) and Mettas (2010).

7.2 Design for Reliability Process

The DfR process encompasses various tools and practices and describes the order of their deployment that an organization needs to drive reliability into the products. Even though most DfR tools have been in use for many years, DfR as a technical discipline and an engineering process is still in the stage of development. There has been some work done on defining this process (see Crowe and Feinberg (2001), ReliaSoft (2007), Mettas (2010), Silverman (2010) and other references at the end of this chapter); however researchers and practitioners are still in the process of reaching a consensus on the specific steps and activities which comprise the DfR process. A structured approach to DfR was suggested later in ReliaSoft (2008) and Mettas (2010), where the process was outlined as an iterative progression of the key design activities along with the appropriate reliability analysis tools.

Depending on the industry, type of product, development cycle and other product specific factors, the DfR process can certainly vary, however in a generic form, it can be depicted as a flow diagram (Figure 7.2).

This flow shows the sequence of engineering activities necessary to achieve a failure-free design. This flow also aligns well with the general product development process of Concept-Design-Development-Manufacturing-Operation/Support. Each stage of the DfR process in Figure 7.2 employs the analysis methods and tools best suited for the required tasks.

Implementing DfR practices and tools is sometimes considered tedious and expensive. In most cases the analysis will show that nearly all aspects of the design are satisfactory, and much more effort will have been expended in showing this than in highlighting a few deficiencies. However, the discovery of even a few deficiencies at an appropriately early stage can save far more than the costs that might be incurred by having to

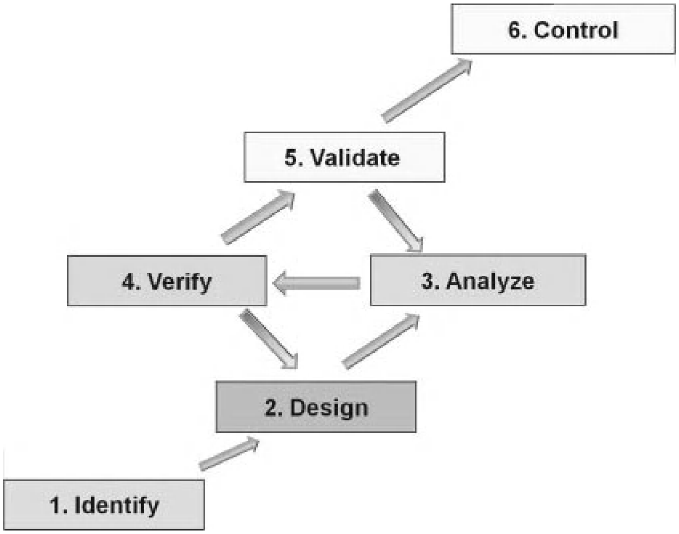


Figure 7.2 *Design for reliability (DfR) activities flow.*

modify the design at a later stage, or by having to live with the consequences of the defect. Therefore, the DfR process if implemented correctly is extremely cost-effective. The tedium and expense can be greatly reduced by good planning and preparation and by the use of appropriate software. In this chapter, we will outline DfR as a step-by-step process and describe the DfR methods and techniques. Some have been discussed in the previous chapters and some will be explained later in the book.

7.3 Identify

The important groundwork for the project design is done at this stage, which begins with understanding of the system requirements. For example, if the specification is 10 years of operation and/or 200 000 km how does that translate into the reliability terms? System reliability requirements can be set by the customer, by the design team, or by following the accepted industry practices. During system design, the top-level reliability requirements are then allocated to subsystems by design engineers and reliability engineers working together (see Chapter 6 for more on reliability apportionment and Chapter 14 on reliability metrics).

This is the time to start collecting data and/or gaining better understanding about the product's usage environments. For example, what kind of temperature profile will be experienced by an electronic controller mounted under the hood of a vehicle? The analysis techniques and design activities appropriate at this stage also include Quality Function Deployment (QFD), benchmarking against the competition, product usage analysis, reliability cost estimate, environments assessment, risk assessment and reliability apportionment. It is also important that reliability engineers become integrated with design teams at the very early stages of this process, so potential reliability issues are addressed from the beginning. If the design involves new technology, this is the appropriate time to begin its reliability and risk assessment activities in addition to determining the technology limitations. For example, the design team can specify that batteries for hybrid vehicles cannot operate below certain temperatures, CD player optics devices cannot withstand high vibration levels, and so on. Activities at this stage may also help to determine the future cost of testing and validating the product.

7.3.1 Benchmarking

Benchmarking is an important activity in process improvement, which goes beyond product reliability. According to the Automotive Industry Action Group, AIAG, 2004 'Benchmarking is the process of improving product and process performance by continuously identifying, understanding and adapting outstanding practices, processes, features and performance levels of world-class products and processes.' Therefore reliability requirements should be benchmarked against the competition in order to achieve or exceed 'best in class' characteristics. Benchmarking should begin with identifying the parameters and measures to benchmark and the companies and organizations which are considered to be 'best in class'. These identified best practices then should be studied through internet or library research, questionnaires, interviews, measurements and other legal means. The concluding step in the benchmarking process should include publishing recommendations and ultimately adapting the best tools and practices to achieve the established goals.

7.3.2 Environments

Determining the usage and environmental conditions is a very important step in the design process. Designers need to know the types of stresses their products will experience in the field. The environments in which the product will be expected to be stored, operated and maintained must be carefully assessed, as well as the expected severity and durations. The assessment must include all aspects that could affect the product's operation, safety and reliability. Physical factors include temperature, vibration, shock, humidity, pressure, and so on. Extreme values and, where relevant, rates of change must be considered. Other environmental conditions, such as corrosive atmospheres, electrical interference, power supply variation, and so on, must also be taken into account. Where appropriate, combined environmental conditions, such as temperature/corrosive atmosphere and vibration/contamination, should be assessed. An aspect of the environment often neglected is the treatment of the product by people, in storage, handling, operation and maintenance.

Environmental aspects should be reviewed systematically, and the review should be properly documented. The protective measures to be taken must be identified, as appropriate to storage, transport, handling, operation and maintenance. Protective measures include packaging, provision of warning labels and instructions, protective treatment of surfaces, and design features. Detailed design aspects are covered in Chapters 8 and 9.

7.3.3 Environment Distribution

In many cases the designed product will be used differently by different users and often in different regions of the world. For example, some people listen continuously to their CD players or iPods, while others may turn them on only once a week. The concept of loads and stresses distributed statistically was introduced in Chapter 5. Statistical representation of usage or environmental data is often referred as *distribution environment, usage distribution*, or *user severity*.

Example 7.1

An automotive supplier needs to determine the driving distances for passenger cars in Europe. The analysis of vehicle repair data from several European car dealerships produced the data sample containing number of kilometres driven by each vehicle from the sales date to repair. This data was converted into a data sample of the yearly driving distances for each vehicle. Best fit analysis for this data sample showed that the yearly

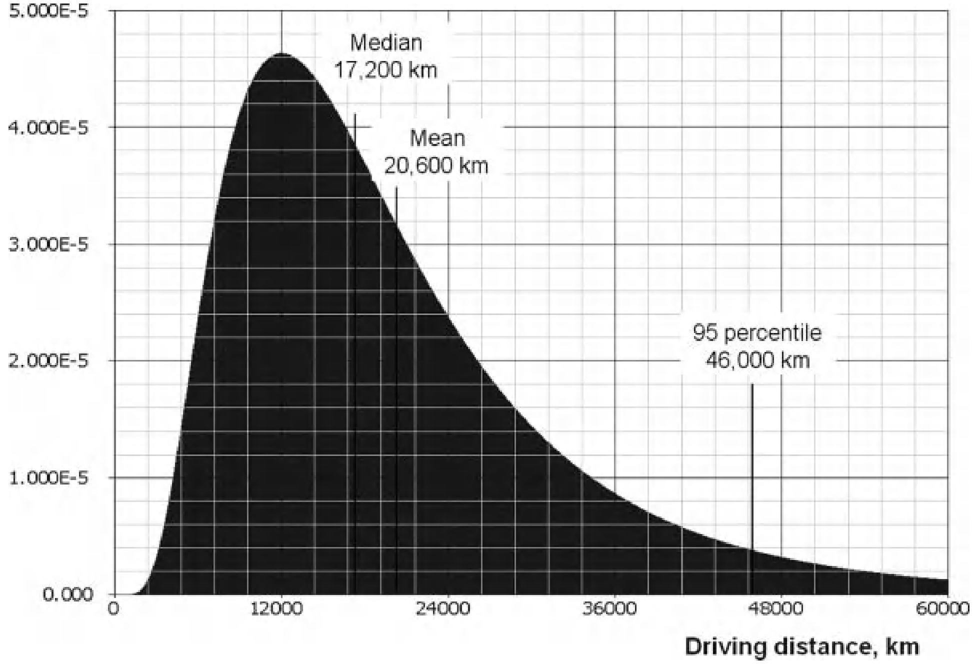


Figure 7.3 *Statistical distribution of the annual driving distances (per passenger car in Europe).*

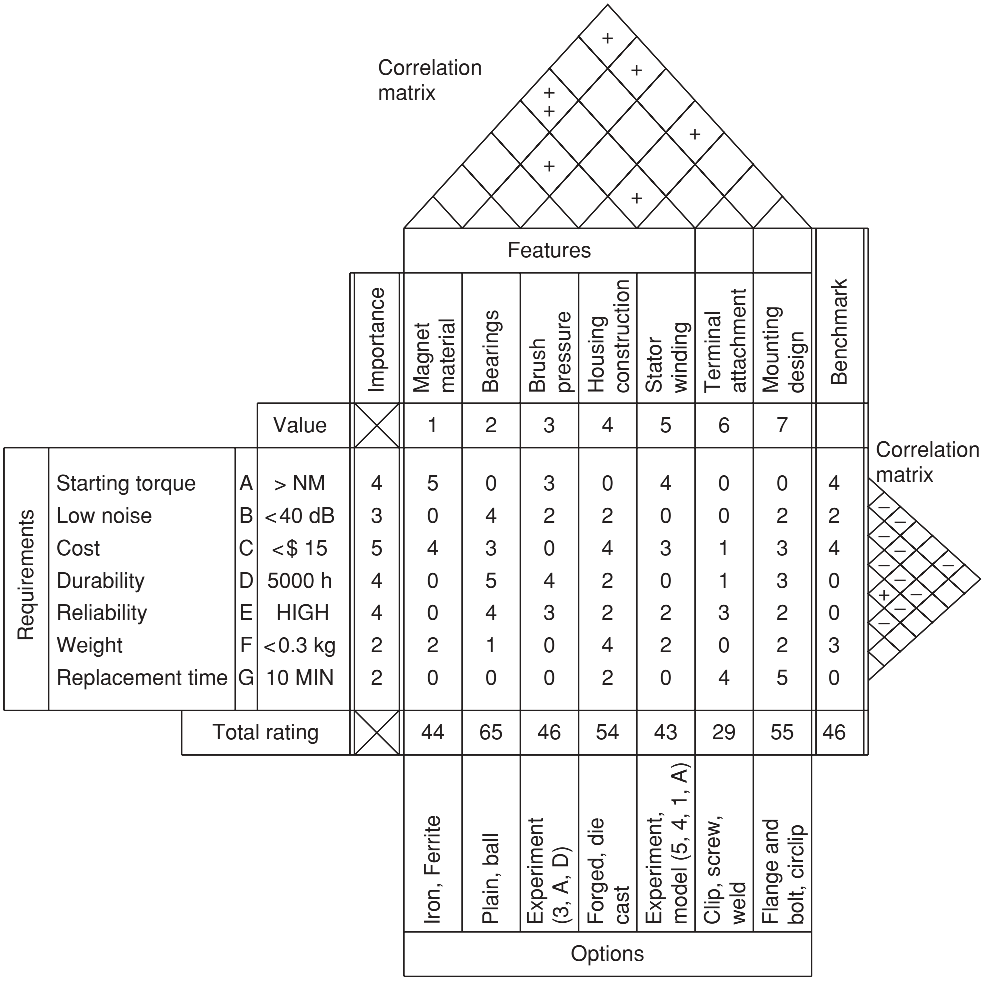
distances driven by passenger cars can be modelled by a lognormal distribution with $\mu = 3.867$ and $\sigma = 0.603$, Figure 7.3. From this distribution the manufacturer learned that the average driving distance is 20 600 km per year, the median (50 percentile) vehicle would be driven 17 200 km and the 95 percentile 46 000 km. Based on this analysis the manufacturer developed the engine reliability specifications for the Tier 2 supplier based on 46 000 km/year.

Depending on the industry or type of the product, the design may be intended for the 50, 90, 95 or even the 99.9 percentile user. Similar types of user profiles in statistical or non-statistical formats can be obtained for various applications, such as mobile phone usage, on/off machine cycles, times at extreme temperatures, laptop power-on times, expected vibration levels, and so on. This type of information can be critical in developing the realistic requirements to manufacture products which can provide reliable operation in the field.

7.3.4 Quality Function Deployment (QFD)

Quality Function Deployment (QFD) is a technique to identify all of the factors which might affect the ability of a design or product to satisfy the customer, and the methods and responsibilities necessary to ensure control. QFD goes beyond reliability, as it covers aspects such as customer preferences for feel, appearance, and so on, but it is a useful and systematic way to highlight design and process activities and controls necessary to ensure reliability.

QFD begins by a team consisting of the key marketing, design, production, reliability and quality staff working their way through the project plan or specification, and identifying the features that will require to be controlled, the control methods applicable, and the responsible people. Constraints and risks are also



			Importance	Magnet material	Bearings	Brush pressure	Housing construction	Stator winding	Terminal attachment	Mounting design	Benchmark
		Value		1	2	3	4	5	6	7	
Starting torque	A	> NM	4	5	0	3	0	4	0	0	4
Low noise	B	<40 dB	3	0	4	2	2	0	0	2	2
Cost	C	<$ 15	5	4	3	0	4	3	1	3	4
Durability	D	5000 h	4	0	5	4	2	0	1	3	0
Reliability	E	HIGH	4	0	4	3	2	2	3	2	0
Weight	F	<0.3 kg	2	2	1	0	4	2	0	2	3
Replacement time	G	10 MIN	2	0	0	0	2	0	4	5	0
		Total rating		44	65	46	54	43	29	55	46
				Iron, Ferrite	Plain, ball	Experiment (3, A, D)	Forged, die cast	Experiment, model (5, 4, 1, A)	Clip, screw, weld	Flange and bolt, circlip	

Figure 7.4 *Quality function deployment for electric motor design.*

identified, as well as resources necessary. At this stage no analysis or detailed planning is performed, but the methods likely to be applied are identified. These methods are described later in this chapter and in others.

QFD makes use of charts which enable the requirements to be listed, and controls responsibilities, constraints, and so on, to be tabulated, as they relate to design, analysis, test, production, and so on. An example is shown in Figure 7.4.

This shows requirements rated on an importance scale (1–5), and the design features that can affect them. Each feature is in turn rated against its contribution to each requirement, and a total rating of each feature is derived by multiplying each rating by the importance value, and adding these values. Thus the bearing selection, housing construction, and mounting design come out as the most critical design features.

The ‘benchmark’ column is used to rate each requirement, as perceived by potential customers, against those of competitive products. Only one benchmark is shown here; more can be added for other competitors. Benchmarking is a useful method for putting requirements into a sound marketing perspective.

The correlation matrices indicate the extent to which requirements and features interact: plus sign(s) indicate positive correlation, and minus negative correlation. For example, magnet material and stator winding design might interact strongly. The minus signs in the requirements matrix indicate conflicting requirements.

The options available are shown. In some cases further modelling or experiments are required, and this part of the chart can be used to indicate the variables that need to be included in such work.

The shape of the QFD chart has led to its being called the 'house of quality'. Of course quality here is used in the widest sense, to include all aspects of the product that will affect its reputation and cost. Figure 7.4 is a top level chart: lower level charts are used to analyse more detailed aspects, for example, more detailed design and component characteristics, and production processes and tolerances, always against the same set of requirements. Thus every aspect of design and production, including analysis, test, production process control, final inspection, packaging, maintenance, and so on, is systematically evaluated and planned for, always in relation to the most important product requirements. Requirements and features that are not important are shown up as such, and this can be a very important contribution to cost reduction and reliability improvement.

Like other conceptual analysis methods, there is no single standard approach, and users should be encouraged to develop formats and methods appropriate to their products and problems. QFD software is available.

The method is described in more detail in Akao (1990) and many other printed and web sources.

7.3.5 Programme Risk Assessment

Product development has an evolutionary nature; therefore in most cases a new product is a further step in development of the existing programme. Thus at the IDENTIFY stage it is important to understand how much change is introduced to the new product compared to the old one. Therefore the following questions should be answered:

- Will this product contain any new technology with unproven reliability record?
- Is this a revolutionary design as opposed to evolutionary?
- Will this design be significantly different from the old one (e.g. more that 30 % of the content is new)?
- Will this product be used at a different geographic region or be exposed to more extreme environments?
- Does this product have any new requirements (e.g. 15 years life instead of 10 years)?
- Will this product have new applications (e.g. consumer electronics product will be installed on a passenger car)?
- Any new materials used in the design?
- Any changes in the supply chain?
- Will the product be made at a different manufacturing location?
- Will this product be supplied to a new customer?
- Are there any other changes, which can affect reliability?

The more 'yes' answers the higher is the reliability risk for this program. The higher is the risk the more attention should be paid to reliability and therefore more DfR activities should be included in the programme.

7.4 Design

This is the stage where specific design activities begin, such as circuit layout, mechanical drawing, component/supplier selection, and so on. Therefore, a more detailed design picture begins to emerge.

At the DESIGN stage the suppliers are identified and preliminary bill of materials becomes available, therefore initial reliability prediction activities can begin (Chapter 6). Other activities at this stage may include Design Failure Modes Effects and Criticality Analysis (FMECA), analysis of the existing product issues (lessons learned), fault tree analysis (Chapter 6), critical items list, human factors, hazard and operability study (HAZOPS), and design reviews. Also load protection and non-material failure modes should be considered as part of FMECA process or as separate activities.

7.4.1 Computer-Aided Engineering

Computer-aided engineering (CAE) methods are used to perform a wide variety of design tasks. CAE also makes possible the creation of designs which would otherwise be very difficult or uneconomic, for example complex electronic circuits. (CAE for electronic design is often referred to as electronic design automation (EDA)). CAE provides enormous improvements in engineering productivity. Properly used, it can lead to the creation of more reliable designs.

The designer can, in principle, design the system, then ‘build’ it and ‘test’ it, all on the computer screen. The effects of parameter changes or failure modes can be quickly evaluated, and dynamic as well as static operating conditions can be tested.

Specialist CAE software is also available for design and analysis of systems and products incorporating other technologies, such as hydraulics, magnetics and microwave electronics. Multi-technology capability is now also available, so that mixed technology designs can be modelled and analysed.

CAE provides the capability for rapid assessment of different design options, and for analysing the effects of tolerances, variation and failure modes. Therefore, if used in a systematic, disciplined way, with adequate documentation of the options studied and assessments performed, designs can be optimized for costs, producibility and reliability.

However, there are important limitations inherent in most CAE tools. The software models can never be totally accurate representations of all aspects of the design and of its operating environment. For example, electronic circuit simulation programs generally ignore the effects of electromagnetic interference between components, and drafting systems will ignore distortion due to stress or temperature. Therefore it is essential that engineers using CAE are aware of the limitations, and how these could affect their designs.

7.4.2 Failure Modes, Effects and Criticality Analysis (FMECA)

Failure modes effects and criticality analysis (FMECA) (or failure modes and effects analysis (FMEA)), is probably the most widely used and most effective design reliability analysis method. The principle of FMECA is to consider each mode of failure of every component of a system and to ascertain the effects on system operation of each failure mode in turn. Failure effects may be considered at more than one level, for example, at subsystem and at overall system level. Failure modes are classified in relation to the severity of their effects.

An FMECA may be based on a hardware or a functional approach. In the hardware approach actual hardware failure modes are considered (e.g. resistor open circuit, bearing seizure). The functional approach is used when hardware items cannot be uniquely identified or in early design stages when hardware is not fully defined. In this approach function failures are considered (e.g. no feedback, memory lost). Note that a functional failure mode can become a hardware failure effect in a hardware-approach FMECA. An FMECA can also be performed using a combination of hardware and functional approaches.

Figure 7.5 shows a typical worksheet in AIAG-3 format (AIAG, 2003), which serves to highlight failure modes whose effects would be considered important in relation to severity, detectability, probability of occurrence, maintainability or safety. Figure 7.5 includes the Risk Priority Number (RPN) to assess risk.

The RPN procedure includes the following steps: rating of the severity of each effect of failure, rating the likelihood of occurrence for each cause of failure and rating the likelihood of prior detection for each cause of failure (i.e. the likelihood of detecting the problem before it reaches the end user or customer). Rating scales usually range from 1 to 10, with the higher number representing the higher severity or risk. For example, 10 points for severity indicates the worst possible consequence of the failure.

RPN is calculated as the product of the three ratings:

$$\text{RPN} = \text{Severity} \times \text{Occurrence} \times \text{Detection}$$

The RPN can then be used to compare issues within the analysis and to prioritize problems for corrective action.

Figure 7.6 show typical worksheets taken from US MIL-STD-1629 Method 102 (criticality analysis). This method includes consideration of failure rate or probability, failure mode ratio and a quantitative assessment of criticality, in order to provide a quantitative criticality rating for the component or function. The *failure mode criticality number* is

$$C_m = \beta\alpha\lambda_p t \tag{7.1}$$

where: $\beta =$ conditional probability of loss of function or mission.

$\alpha =$ failure mode ratio (for an item, $\Sigma_\alpha = 1$).

$\lambda_\text{p} =$ item failure or hazard rate (can be obtained from the standards-based or other reliability prediction method, Chapter 6).

$t =$ operating or at-risk time of item.

$\lambda_\text{p}\, t$ can be replaced by failure probability, $1 - \exp(-\alpha\lambda_\text{p}t)$.

The *item criticality number* is the sum of the failure mode criticality numbers for the item.

FMECA is widely used in many industries, particularly in those for which failures can have serious consequences, such as military, aerospace, automotive, medical equipment, and so on. There are other published standards besides AIAG-3 and MIL-STD-1629 containing FMECA forms and guidelines. Those standards include ISO/TS16949 (automotive quality management system), SAE J1739 for FMEA (published by Society of Automotive Engineers for the automotive industry), IEC 60812 (Procedure for Failure Mode and Effects Analysis), ARP5580 (Recommended Failure Modes and Effects Analysis Practices for Non-Automobile Applications), P-302-720 (NASA Flight Assurance Procedure) and others.

7.4.3 Steps in Performing an FMECA

An effective FMECA should be performed by a team of engineers having a thorough knowledge of the system's design and application. This team can be augmented by specialists from other functional areas, such as purchasing, tech support, testing, facilities, marketing, and so on. The first step therefore is to obtain all the information available on the design. This includes specifications, drawings, computer-aided engineering (CAE) data, stress analysis, test results, and so on, to the extent they are available at the time. For a criticality analysis, the reliability prediction information must also be available or it might be generated simultaneously.

A system functional block diagram and reliability block diagram (Chapter 6) should be prepared, if not already available, as these form the basis for preparing the FMECA and for understanding the completed analysis.

Process Step/ Function Requirements	Potential Failure Mode	Potential Effect(s) of Failure	Severity	Classification	Potential Cause(s) of Failure	Occurrence	Current Process Controls Prevention	Current Process Controls Detection	Detection	RPN	Recommended Action	Responsibility & Target Completion Date	Action Results				
													Actions Taken & Effective Date	Severity	Occurrence	Detection	RPN

Figure 7.5 FMEA worksheet for AIAG-3 method.

						Failure probability							
Identification number	Item/functional identification (nomenclature)	Function	Failure modes and causes	Mission phase/ operational mode	Severity class	Failure rate data source	Failure effect probability	Failure mode ratio (α)	Failure ratio (λ_p)	Operating time (t)	Failure mode crit $C_m = \beta\alpha\lambda_p$	Item crit $C_r = \Sigma(C_m)$	Remarks

Figure 7.6 MIL-STD-1629 Method 102 worksheet for criticality analysis.

If the system operates in more than one phase in which different functional relationships or item operating modes exist, these must be considered in the analysis. The effects of redundancy must also be considered by evaluating the effects of failure modes assuming that the redundant subsystem is or is not available.

An FMECA can be performed from different viewpoints, such as safety, mission success, availability, repair cost, failure mode or effect detectability, and so on. It is necessary to decide, and to state, the viewpoint or viewpoints being considered in the analysis. For example, a safety-related FMECA might give a low criticality number to an item whose reliability seriously affects availability, but which is not safety critical.

The FMECA is then prepared, using the appropriate worksheet, and working to the item or subassembly level considered appropriate, bearing in mind the design data available and the objectives of the analysis. For a new design, particularly when the effects of failures are serious (high warranty costs, reliability reputation, safety, etc.) the analysis should take account of all failure modes of all components. However, it might be appropriate to consider functional failure modes of subassemblies when these are based upon existing designs, for example, modular power supplies in electronic systems, particularly if the design details are not known.

The FMECA should be started as soon as initial design information is available. It should be performed iteratively as the design evolves, so that the analysis can be used to influence the design and to provide documentation of the eventually completed design. Design options should be analysed separately, so that reliability implications can be considered in deciding on which option to choose. Test results should be used to update the analysis.

FMECA is not a trivial task, and can involve many hours or weeks of work. It can also be difficult to trace the effects of low-level failures correctly through complex systems. The CAE/ EDA software can be used to assist in the analysis, thus aiding the task of working out the effects of component-level failures on the operation of complex systems. Even with aids such as these, FMECA can be an inappropriate method for some designs, such as digital electronic systems in which low-level failures (e.g. of transistors within integrated circuits) are very unlikely, and the effects are dynamic in the sense that they could differ widely depending upon the state of the system.

7.4.4 Uses for FMECA

FMECAs can be used very effectively for several purposes, in addition to the prime one of identifying safety or reliability critical failure modes and effects. These include:

1. Identifying features to be included in the test programme (Chapters 12 and 13).
2. Preparation of diagnostic routines such as flowcharts or fault-finding tables. The FMECA provides a convenient listing of the failure modes which produce particular failure effects or symptoms, and their relative likelihoods of occurrence.
3. Preparation of preventive maintenance requirements. The effects and likelihood of failures can be considered in relation to the need for scheduled inspection, servicing or replacement. For example, if a failure mode has an insignificant effect on safety or operating success, the item could be replaced only on failure rather than at scheduled intervals, to reduce the probability of failure. See Chapter 16.
4. Design of built-in test (BIT), failure indications and redundancy. The failure detectability viewpoint is an important one in FMECA of systems which include these features.
5. For analysis of testability, particularly for electronic subassemblies and systems, to ensure that hardware can be economically tested and failures diagnosed, using automatic or manual test equipment.
6. For development of software for automatic test and BIT.

7 For retention as formal records of the safety and reliability analysis, to be used as evidence if required in reports to customers or in product safety litigation.

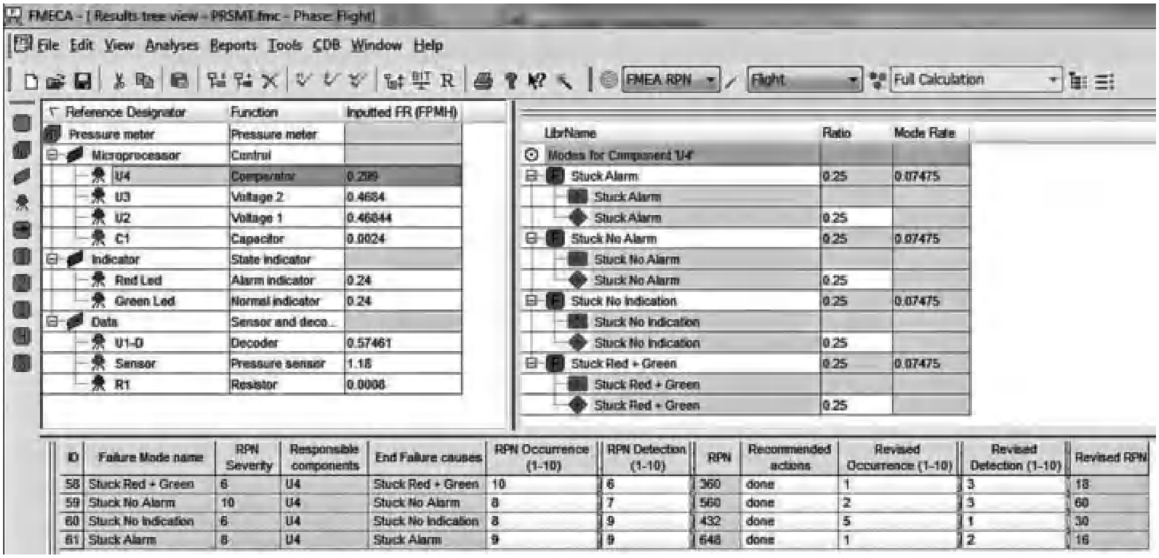
8 To consider the possibility of production-induced failures, for example, wrong diode orientation. Such a *process* FMECA can be very useful in test planning and in design for ease of production.

It is important to coordinate these activities, so that the most effective use can be made of the FMECAs in all of them, and to ensure that FMECAs are available at the right time and to the right people.

7.4.5 FMECA Software Tools

Software has been developed for the performance of FMECA. Using software instead of FMECA worksheets allows FMECAs to be produced more quickly and accurately, and greatly increases the ease of editing and updating to take account of design changes, design options, different viewpoints, and different input assumptions. Like any other computer-aided design technique, computerized FMECA frees engineers to concentrate on engineering, rather than on tedious compilation, so that for the same total effort designs can be more thoroughly investigated, or less effort can be expended for the same depth of analysis.

The FMECA software enables more perceptive analysis to be performed. Failure effects can be ranked in criticality order, at different system levels, in different phases of system operation and from different viewpoints. Report preparation can be partly automated and sensitivity analyses quickly performed. Figure 7.7 shows part of a computerized FMECA, performed using the CARE® software package by BQR Reliability Engineering. It shows three levels of FMECA tree for the comparator U4. Figure 7.7 shows the potential failure mode for this component (on the right) and the RPN analysis (at the bottom). It is also possible, and effective, to use a spreadsheet to create an FMECA. This method has the advantage that the format and type of analysis can be designed to suit the particular design and methods of analysis.


FMECA - [Results tree view - PRSMT.fmc - Phase: Flight]
File Edit View Analyses Reports Tools CDB Window Help
FMEA RPN
Flight
Full Calculation

Reference Designator	Function	Inputted FR (FPMH)
Pressure meter	Pressure meter	
Microprocessor	Control	
U4	Comparator	0.299
U3	Voltage 2	0.4684
U2	Voltage 1	0.46844
C1	Capacitor	0.0024
Indicator	State indicator	
Red Led	Alarm indicator	0.24
Green Led	Normal indicator	0.24
Data	Sensor and deco...	
U1-D	Decoder	0.57461
Sensor	Pressure sensor	1.18
R1	Resistor	0.0008

LibrName	Ratio	Mode Rate
Modes for Component 'U4'		
Stuck Alarm	0.25	0.07475
Stuck Alarm		
Stuck Alarm	0.25	
Stuck No Alarm	0.25	0.07475
Stuck No Alarm		
Stuck No Alarm	0.25	
Stuck No Indication	0.25	0.07475
Stuck No Indication		
Stuck No Indication	0.25	
Stuck Red + Green	0.25	0.07475
Stuck Red + Green		
Stuck Red + Green	0.25	

ID	Failure Mode name	RPN Severity	Responsible components	End Failure causes	RPN Occurrence (1-10)	RPN Detection (1-10)	RPN	Recommended actions	Revised Occurrence (1-10)	Revised Detection (1-10)	Revised RPN
58	Stuck Red + Green	6	U4	Stuck Red + Green	10	6	360	done	1	3	18
59	Stuck No Alarm	10	U4	Stuck No Alarm	8	7	560	done	2	3	60
60	Stuck No Indication	6	U4	Stuck No Indication	8	9	432	done	5	1	30
61	Stuck Alarm	8	U4	Stuck Alarm	9	9	648	done	1	2	16

Figure 7.7 *Part of output listing from CARE®, the FMECA software (Reproduced by permission of BQR Reliability Engineering Ltd.).*

7.4.6 Reliability Predictions for FMECA

Since FMECAs are performed primarily to identify critical failure modes and to evaluate design options, failure rate or reliability values which could be considered as realistic worst cases should be used as it was originally implemented in MIL-STD-1629. Standard methods sometimes stipulate the reliability prediction methods to be used with FMECA, for example, MIL-HDBK-217 for electronics or NSWC-06/LE10 for mechanical systems. However, it is very important to appreciate the large amount of uncertainty inherent in reliability prediction, particularly at the level of individual failure events (see Chapter 6). Therefore, worst-case or pessimistic reliability values should always be used as input assumptions for failure modes which are identified as critical, or which might be critical if the pessimistic assumption proved to be realistic. Alternatively, and preferably unless credible quantitative data are available, a value scale such as 0–1 should be used, with prearranged assignment (e.g. 1 = will definitely occur, 0.5 = will occur occasionally, 0.1 = will rarely occur, 0 = will never occur). Generally, the more critical the failure mode the more pessimistic should be the worst-case reliability assumptions.

7.4.7 Load-Strength Analysis

Load–Strength analysis (LSA) covered in Chapter 5 is a procedure to ensure that all load and strength aspects have been considered in deriving the design, and if necessary in planning of tests. Load-Strength analysis may begin at the early stages of the DESIGN phase and continue through most of the DfR process as more data about system characteristics become available. The LSA should include the following:

- Determine the most likely worst case values and patterns of variation of load and strength.
- Evaluate the safety margin for intrinsic reliability.
- Determine protection methods (load limit, derating, screening, other quality control methods).
- Identify and analyse strength degradation modes.
- Test to failure to corroborate, analyse results.
- Correct or control (redesign, safe life, quality control, maintenance, etc.).

Table 7.1 is an example of a hypothetical load–strength analysis for a mechanical and electrical assembly. The example shows approaches that can be used for different aspects of the analysis. Event probabilities can be expressed as full distributions, or as the likelihood of a particular limiting case being exceeded. The former is more appropriate when the load(s) can cause degradation, or if a more detailed reliability assessment is required. Both examples show typical, though rather simple, cases where the effects of combined loads might have been overlooked but for the analysis. For example, the solenoid might be supplied with a manufacturer's rating of 28 V operating, ±2 V, and a maximum ambient temperature of 45 °C. A room temperature test of the solenoid might have confirmed its ability to function with a 32 V supply without overheating. However, the combined environment of +45 °C and 32 V supply, albeit an infrequent occurrence, could lead to failure.

The mechanical example is less easy to analyse and testing is likely to be the best way of providing assurance, if the assembly is critical enough to warrant it. Where the load–strength analysis indicates possible problems, further analysis should be undertaken, for example, use of probabilistic methods as described in Chapters 4 and 5, and CAE methods. Tests should be planned to confirm all design decisions.

7.4.8 Hazard and Operability Study (HAZOPS)

Hazard and operability study (HAZOPS) is a technique for the systematic determination of the potential hazards that could be generated by a system, and of the methods that should be applied to remove or minimize

Table 7.1 *Load–strength analysis example.*

Item (Matl, function)	Worst case load/ combined load	Frequency/ probability of occurrence	Data source	Combined effect	Strength	Remarks
Rivet (×4) (aluminium, fixing bracket to plastic frame)	1. 50 N total, axial	Continuous	—	—	Plastic frame is weak link	*Life test to confirm Thickness of plastic frame at bracket attachment may be critical feature*
	2. 40 N, lateral impact	See load distribution annex 1	Operating data	Combine with 1 *Degradation*		
	3. Temperature 0–35 °C		NIL	NIL effect		
Solenoid coil	1. 32 V (at 27 °C)	$1/10^4$ h	Data on power supply variation	72 °C	Insulation limited to 70 °C	Overvoltage protection or improved cooling needed
	2.45 °C ambient	$1/10^2$ h				

them. It is used in the development of systems such as petrochemical plants, railway systems, and so on and usually is part of the mandatory safety approval process. Table 7.2 shows an example of the format used.

For the failure/deviation column, a set of 'guidewords' is sometimes applied to help in the identification of things that could possibly go wrong. The usual guidewords are:

—no/not	—part of
—more	—reverse
—less	—other than
—as well as	

HAZOPS should cover the whole range of potential failure causes, including natural hazards, human failures, and so on. HAZOP is also commonly used in risk assessments for industrial and environmental health and safety applications. Additional details on the HAZOP methodology can be found in the standard IEC 61 882, Hazard and Operability Studies (HAZOP) Application Guide.

7.4.9 Parts, Materials and Processes (PMP) Review

All new parts, materials and processes called up in the design should be identified. 'New' in this context means new to the particular design and production organization. The designer is likely to assume that a part or material will perform as specified in the brochure and that processes can be controlled to comply with the design. The reliability and quality assurance (QA) staff must ensure that this faith is well-founded. New parts, materials and processes must therefore be assessed or tested before being applied, so that adequate training for production people can be planned, quality control safeguards set up and alternative sources located. New parts, materials and processes must be formally approved for production and added to the approved lists.

Materials and processes must be assessed in relation to reliability. The main reliability considerations include:

1 *Cyclical loading*. Whenever loading is cyclical, including frequent impact loads, fatigue must be considered.
2 *External environment*. The environmental conditions of storage and operation must be considered in relation to factors such as corrosion and extreme temperature effects.
3 *Wear*. The wear properties of materials must be considered for all moving parts in contact.

There is such a wide variation of material properties, even amongst categories such as steels, aluminium alloys, plastics and rubbers, that it is not practicable to generalize about how these should be considered in relation to reliability. Material selection will be based upon several factors; the design review procedure should ensure that the reliability implications receive the attention appropriate to the application. Chapter 8 covers mechanical design for reliability in more detail.

7.4.10 Non-Material Failure Modes

Most reliability engineering is concerned with material failure, such as caused by load–strength interference and strength degradation. However, there is a large class of failure modes which are not related to this type of material failure, but which can have consequences which are just as serious. Examples of these are:

1 Fasteners which secure essential panels and which can be insecurely fastened due to wear or left unfastened without being detected.

Table 7.2 *HAZOPS on motion system (partial).*

Component/Function	Failure/Deviation	Possible Cause/s	Consequences/ Event No	Safeguards	Action
Electrical power	No power	1. Main power fail 2. Connector	System failure (1)	Provide standby power?	System design
Hydraulic supply	Main AND standby fail	1. Main AND pressure sensor fail.	System failure (2)	Checks on maintenance	Maintenance schedule
		2. Main AND changeover valve fail	System failure (2)		
PWM circuit	Permanent 'on'	See FMEA	System failure (3)	*To be determined*	Analysis, test
Solenoid valve	Stuck open	Corrosion	System failure (3)	*To be determined*	Test

2 Wear in seals, causing leaks in hydraulic or pneumatic systems.
3 Resistance increase of electrical contacts due to arcing and accretion of oxides.
4 Failure of protective surfaces, such as paints, metal plating or anodized surfaces.
5 Distortion of pins, or intermittent contact, on multipin electrical connectors.
6 Drift in electronic component parameter values.
7 Electromagnetic interference (EMI) and timing problems in electronic systems.
8 Other personnel-induced failures such as faulty maintenance, handling or storage, for example, omitting to charge electrolytic capacitors kept in long-term storage, which can result in reduced charge capacity in use.
9 Interface problems between sub-systems, due to tolerance mismatch.

All of these modes can lead to perceived failures. Failure reporting systems always include a proportion of such failures. However, there is usually more scope for subjective interpretation and for variability due to factors such as skill levels, personal attitudes and maintenance procedures, especially for complex equipment.

Non-material failures can be harder to assess at the design stage, and often do not show up during a test programme. Design reliability assessments should address these types of failure, even though it may be impracticable to attempt to predict the frequency of occurrence in some cases, particularly for personnel-induced failures.

7.4.11 Critical Items List

The critical items list is a summary of the items shown by the other analyses to be likely either to have an appreciable effect on the product's reliability or to involve uncertainty. Its purpose is to highlight these items and summarize the action being taken to reduce the risks. The initial list will be based upon the design analyses, but updates will take account of test results, design changes and service data as the project develops. The critical items list is a top document for management reporting and action as it is based upon the 'management by exception' principle and summarizes the important reliability problems. Therefore, it should not usually include more than ten items and these should be ranked in order of criticality, so that management attention can be focused upon the few most important problems. It could be supported by a Pareto chart (Chapter 13) to show the relative importance, when there are sufficient data. The critical items list should provide only identification of the problem and a very brief description and status report, with references to other relevant reports.

7.4.12 Load Protection

Protection against extreme loads is not always possible, but should be considered whenever practicable. In many cases the maximum load can be pre-determined, and no special protection is necessary. However, in many other loading situations extreme external loads can occur and can be protected against. Standard products are available to provide protection against, for example, overpressure in hydraulic or pneumatic systems, impact loads or electrical overload. When overload protection is provided, the reliability analysis is performed on the basis of the maximum load which can be anticipated, bearing in mind the tolerances of the protection system. In appropriate cases, loads which can occur when the protection system fails must also be considered.

However, in most practical cases it will be sufficient to design to withstand a predetermined load and to accept the fact that loads above this will cause failure. The probability of such loads occurring must be determined for a full reliability analysis to be performed. It may not always be practicable to determine the distribution of such extreme events, but data may be available either from failure records of similar items, or from test or other records.

Where credible data are not available, the worst design load case must be estimated. The important point is that the worst design case is estimated and specified. A common cause of failure is the use of safety factors related to average load conditions, without adequate consideration having been given to the extreme conditions which can occur during use of the product.

7.4.13 Protection against Strength Degradation

Strength degradation, in its many forms, can be one of the most difficult aspects to take into account in design reliability analysis. Strength degradation due to fatigue in metals is fairly well understood and documented, and therefore reliability analysis involving metal fatigue, including the effects of stress raisers such as notches, corners, holes and surface finish, can be performed satisfactorily, and parts can be designed to operate below the fatigue limit, or for a defined safe life.

However, other weakening mechanisms are often more complex. Combined stresses may accelerate damage or reduce the fatigue limit. Corrosion and wear are dependent upon environments and lubrication, the effects of which are therefore often difficult to forecast. If complete protection is not possible, the designer must specify maintenance procedures for inspection, lubrication or scheduled replacement.

Reliability analysis of designs with complex weakening processes is often impracticable. Tests should then be designed to provide the required data by generating failures under known loading conditions. Chapter 8 covers these aspects in more detail.

7.4.14 Design Reviews

The review techniques described must be made part of a disciplined design sequence, or they will merely generate work and not advance the objective of more reliable design. To be effective, they must be performed by the people who understand the design. This does not necessarily mean the designers, for two reasons. First, the analyses are an audit of their work and therefore an independent assessment is generally more likely to highlight aspects requiring further work than would be the case if the designers were reviewing their own work. Second, the analyses are not original work in the same sense as is the design. The designers are paid to be creative and time spent on reassessing this effort is non-productive in this sense. The designers may, however, be the best qualified to perform much of the analysis, since they know the problems, assessed the options, carried out all the design calculations and created the solutions. On the other hand, the creative talent may not be the best at patiently performing the rather tedious review methods.

The best solution to this situation is the peer review format, where the engineers performing the reviews work closely with the designers and act as their 'Devil's advocate' during the creative process. In this way, designers and the reviewers work as a team, and problem areas are highlighted as early as possible. The organization of reliability engineering staff to provide this service is covered in Chapter 17. The reviewer should ideally be a reliability engineer who can be respected by the designer as a competent member of a team whose joint objective is the excellence of the design. Since the reliability engineer is unlikely to spend as much time on one design as the designer, one reliability engineer can usually cover the work of several designers. The ratio obviously depends upon the reliability effort considered necessary on the project and on the design disciplines involved.

By working as a team, the design and reliability staff can resolve many problems before the formal analysis reports are produced, and agreement can be reached on recommendations, such as the tests to be performed. Since the reliability engineer should plan and supervise the tests, the link is maintained. Also, the team approach makes it possible for designs to be adequately reviewed and analysed before drawings are signed off, beyond which stage it is always more difficult and more expensive to incorporate changes.

Unfortunately, this team approach is frequently not applied, and design and reliability staff work separately in preparing analyses and criticizing one another's work at a distance, either by email, teleconferencing, or over the conference table. Design review techniques then lose credibility, as do reliability staff. The main victim is the design itself, since the protagonists usually prosper within their separate organizations.

To be of continuing value, the design analyses must be updated continually as design and development proceed. Each formal review must be based upon analyses of the design as it stands and supported by test data, parts assessments, and so on. The analyses should be scheduled as part of the design programme, with design reviews scheduled at suitable intervals. The reviews should be planned well in advance, and the designers must be fully aware of the procedure. All people attending must also be briefed in advance, so they do not waste review time by trying to understand basic features. To this end, all attendees must be provided with a copy of all formal analysis reports (reliability prediction, load–strength analysis, PMP review, maintainability analysis, critical items list, FMECA, FTA) and a description of the item, with appropriate design data such as drawings. The designer should give a short presentation of the design and clear up any general queries. Each analysis report should then form a separate agenda item, with the queries and recommendations as the subjects for discussion and decision. If experience has generated a checklist appropriate to the design, this could also be run through, but see the comments that follow.

With this procedure, nearly all aspects requiring further study or decision will have been discussed before, during the continuous, informal process of the team approach to preparing the analyses. The formal review then becomes a decision-making forum, and it is not bogged down with discussion of trivial points. This contrasts markedly with the type of design review meeting which is based largely upon the use of checklists, with little preparatory work. Such reviews become a stolid march through the checklist, many of whose questions might be irrelevant to the design. They can become a substitute for thinking.

Three golden rules for the use of checklists should be:

1 Use them in the design office, not during the formal design review meetings.
2 Ensure that they are relevant and up to date.
3 Avoid vague questions such as 'Has maintenance been considered?', or even 'Are the grease points accessible?' 'What access is provided for lubrication?' would be a better question, since it calls for detailed response, not a simple affirmative.

The design review team should consist of staff from sales, production, QA and specialists in key design areas. The people on the spot are the designers and the reliability engineering team member (who may belong to the QA department). The chairman should be the project manager or another person who can make decisions affecting the design, for example, the chief designer. Sometimes design reviews are chaired by the procuring agency, or it may require the option of attending. A design review which is advisory and has no authority is unlikely to be effective, and therefore all those attending must be concerned with the project (apart from specialists called in as advisers).

Formal design review meetings should be scheduled to take place when sufficient information is available to make the meeting worthwhile and in time to influence future work with the minimum of interference with project schedules and budgets. Formal and informal design reviews should begin at the IDENTIFY stage of the DfR process and continue virtually through all of its phases, although with different intensity. Three formal reviews are typical, based upon initial designs, completion of development testing and production standard drawings. Each review authorizes transition to the next phase, with such provisos deemed to be necessary, for example, design changes, additional tests. The design reviews should be major milestones in a project's evolution. They are not concerned solely with reliability, of course, but reliability engineers have considerably influenced the ways that modern design reviews are conducted, and design reviews are key events in reliability programmes.

7.4.15 Design Review Based on Failure Modes (DRBFM)

When the design is evolutionary and does not involve many changes, a technique called Design Review Based on Failure Modes (DRBFM) can be applied. This tool was originally developed by Toyota engineers on the premise that reliability problems occur when changes are made to existing designs that have already been proven successful. DRBFM can be considered as a narrowed down FMECA with the attention focusing on the new points and changed points from the existing design. DRBFM encourages design teams to discuss the potential design problems or weaknesses from a cross functional multi-perspective approach, and to develop corrective actions.

DRBFM is performed based on FMECA while focusing specifically on product changes, both intentional and incidental. Therefore DRBFM activities are similar to FMECA and use similar format worksheets. The DRBFM worksheet can vary but usually requires the information about the following: component, its function, change point(s), reasons for change, potential failure modes, condition of their precipitation, effect on the customer, design steps to prevent that failure, recommended actions (result of DRBFM) and action results (conclusion of DRBFM). The Society of Automotive Engineers (SAE) has published standard J2886 containing an explanation of what the DRBFM process is, the recommended steps and examples of how to conduct this process.

7.4.16 Human Reliability

The term ‘human reliability’ is used to cover the situations in which people, as operators or maintainers, can affect the correct or safe operation of systems. In these circumstances people are fallible, and can cause component or system failure in many ways.

Human reliability must be considered in any design in which human fallibility might affect reliability or safety. Design analyses such as FMECA and FTA should include specific consideration of human factors, such as the possibility of incorrect operation or maintenance, ability to detect and respond to failure conditions, and ergonomic or other factors that might influence them. Also, where human operation is involved, product design should be made in full consideration of physiological and psychological factors in order to minimize the probability of human error in system operation.

Attempts have been made to quantify various human error probabilities, but such data should be treated with caution, as human performance is too variable to be credibly forecast from past records. Human error probability can be minimized by training, supervision and motivation, so these must be considered in the analysis. Of course in many cases the design organization has little or no control over these factors, but the analyses can be used to highlight the need for specific training, independent checks, or operator and maintainer instructions and warnings. More on human factors in engineering can be found in Wickens *et al*. (2003).

7.5 Analyse

It is important at this phase to further address all the potential sources of product failure. Various types of analysis can be done after the first draft of design is created, including physics of failure, finite element analysis (FEA), warranty data analysis, continued DRBFM, reliability prediction (Chapter 6), study of the lessons learned on previous programmes, design of experiments (DOE) (Chapter 11), derating analysis (Chapter 9). When a mock up, proof of concept, or engineering development unit is built it will make it easier to verify the results of the analysis and improve on the design.

Finite element analysis (FEA) is one of the important tools for physics of failure (PoF) analysis discussed in Chapter 6. FEA can be utilized to calculate the stresses caused by thermal expansion, vibration, accidental drop, and other environments. It can also be used to estimate fatigue life for products subjected to thermal cycling or vibration. More on fatigue is covered in Chapter 8.

As shown in Figure 7.2, DfR flow has an iterative pattern, especially in the DESIGN-ANALYZE-VERIFY sequence, thus the same design tools can be used at the different phases.

7.5.1 Field Return and Warranty Data Analysis

Field return and warranty data analysis can be an invaluable source in identifying and addressing potential reliability problems. In the cases where the new product design is not significantly different from the existing one, failures experienced in the field can be relevant to the new design. Engineering feedback from the field is essential to successful product design and development. The new product development process needs to be attuned to the engineering analysis of returned parts to prevent old problems from recurring in new products.

Depending on the complexity of the returned parts, the engineering analysis tasks can be accomplished by failure analysis or structured problem solving, or by using a combination of existing continuous improvement tools.

Engineering analysis, for example, can determine that a failure occurred due to an assembly problem, end-user abuse, software malfunction, electronic or mechanical component failure, corrosion, overheating or vibration. The analysis should narrow down the design related failure and identify the problems which may repeat themselves in the new design. For example, if the same faulty parts will be used in the new product, or if the parts would go through the same soldering process at the same assembly plant, where the field return parts came from.

Additionally, warranty data are routinely used for reliability and warranty prediction in new product development. Field failure data of existing products can often be more accurate reliability predictors of future products than some of the traditional methods, such as reliability growth models (Chapter 14) or standards-based predictions.

7.6 Verify

At this stage hardware prototype is available and the design verification activities can begin. These activities include *accelerated life testing* (ALT) and *highly accelerated life testing* (HALT) (both Chapter 12); life data analysis (Chapter 3), degradation analysis, configuration control, sub-system level testing, reliability growth modelling (Chapter 14). After a problem is detected, root cause analysis and structured problem solving can be applied using tools like the Ishikawa diagram and other techniques covered in Chapter 15.

7.6.1 Degradation Analysis

Product verification often involves test to failure and life data analysis. However in many cases it may take too long for the product to fail operating under normal or even accelerated stress conditions. Degradation analysis is one way to shorten the test time and assess if the product is meeting reliability specifications. It can also be a way to assess the strength degradation covered in Section 7.4.9. Degradation analysis involves measurement of the degradation of a certain product characteristic over time (current reduction, wear, etc.), and extrapolation of this degradation data to estimate the eventual failure time for the product. More on degradation analysis will be covered in Chapter 14.

7.6.2 Configuration Control

Configuration control is the process whereby the exact design standard of a system is known. Configuration control applies to hardware and to software. Effective configuration control ensures that, for example, the specifications and sources of components, and the issue numbers of drawings, can be readily identified for a particular system. Configuration control is very important in the development and production of systems, and it is mandatory for projects such as in aerospace (civil and military) and defense. Formal control should start after the first design review.

Configuration control is important to reliability, since it allows failures to be traced back to the appropriate design standard. For example, failures might occur in a component machined to a particular tolerance; the configuration control system should enable this cause to be identified.

7.7 Validate

The term ‘*validation*’ might have different meanings depending on the engineering area it is applied in. For example in systems engineering ‘validation’ is defined as the process to ensure that the system meets all the customer requirements and specifications with the emphasis on the system’s functionality. In reliability engineering validation usually deals with both functional and environmental specifications and it is also set up to ensure that all the reliability requirements of the system are met.

At the VALIDATE stage hardware design and verification are complete, software is debugged and the system is fully functional. The goal of validation is to successfully resolve design and manufacturing issues in case they had been overlooked at the previous design phases. Validation usually involves functional and environmental testing at a system level with the purpose of ensuring that the design is production-ready. These activities may include test to failure or test to success (Chapter 14) and are usually conducted at field stress levels or as accelerated life testing (ALT). It is also important at this stage to have the software tested, released, and be production-ready. Reliability requirements also need to be demonstrated at this stage.

Product validation is often done in two phases *design validation* (DV) and *process validation* (PV). DV activities usually include the environmental, durability, capability and functional tests and are executed on a prototype. The PV tasks are similar to DV but are executed on pilot or production parts, preferably manufactured at the intended production facilities. The intent is to validate that the production processes are fully capable of repeatedly producing products that meet specifications. More on reliability testing, reliability data analysis and reliability demonstration are covered in Chapters 12–14.

7.8 Control

The goal at this stage is to keep the manufacturing process under control and maintain low process variability (Chapter 15). The engineering tools applied at this stage include automatic inspections, control charts, audits, human factor, burn-in, analysis of the known production issues, environmental stress screening (ESS) (Chapter 15), highly accelerated stress screening (HASS) (Chapter 12), process FMEA, and other production-focused activities. Many of the activities discussed in this section should be started at the earlier phases of the design process (sometimes as early as DESIGN stage) and run in parallel to the other design activities, however since they directly affect production processes they are listed and explained in this section.

7.8.1 Design Analysis for Processes

The processes that will be used to manufacture and maintain the product must be understood and optimized. It is essential that all of the processes are capable of being performed correctly and efficiently, whether by people or by machines. Therefore the designers must know the methods that will be used and their capabilities and limitations, and must design both the product and the processes accordingly. The test programme must include tests of all of the processes that have been shown by the analyses to be critical or important.

The applicable analysis methods that can be used are described below.

7.8.2 Variation

As discussed in Chapter 2, all manufacturing processes are subject to variation, as are parameter values and dimensions of parts and subassemblies. Production people and processes inevitably vary in their performance in terms of accuracy and correctness. The design must take all of these into account, and must minimize the possibility of failure due to production-related causes.

The use of FMECA to help to identify such causes has been mentioned above. Techniques for evaluating the effects of variation in production processes are described later. These methods are sometimes referred to as *process* design, to distinguish them from those aspects of the design which address the product specification and its environment. Product and process design, using an integrated approach, including the test and analysis techniques described later, are sometimes referred to as *off-line quality control*. Process design leads to the setting of the correct controls on the production processes, for monitoring as part of *on-line quality control*, described in Chapter 15.

There are two possible approaches to designing for parameter variation and tolerances, the 'worst case' approach, and statistical methods. The traditional approach is to consider the worst case. For example, if a shaft must fit into a bore, the shaft and bore diameters and tolerances might be specified as: shaft, 20 ± 0.1 mm; bore, 20.2 ± 0.1 mm, in order to ensure that all shafts fit all bores. If the tolerance limits are based upon machining processes which produce parts with normally distributed diameters, of which 2.5 % are oversize and 2.5 % are undersize (2σ limits), the probability of a shaft and bore having an interference would be

$$0.025 \times 0.025 = 0.000625$$

On the other hand, most combinations will result in a fairly loose fit. Figure 7.8 shows the situation graphically. 25 % of combinations will have fits greater than 0.2 mm.

If, however, statistical tolerancing were used, we could design for much closer nominal diameters and still have an acceptably low probability of interference. If the shaft nominal diameter was set at 20.1 mm (dotted line on Figure 7.8), the interference probability P_{I} can be calculated as:

$$P_{\mathrm{I}} = 1 - \Phi\left[\frac{D_1 - D_2}{\left(\sigma_1^2 + \sigma_1^2\right)^{1/2}}\right]$$

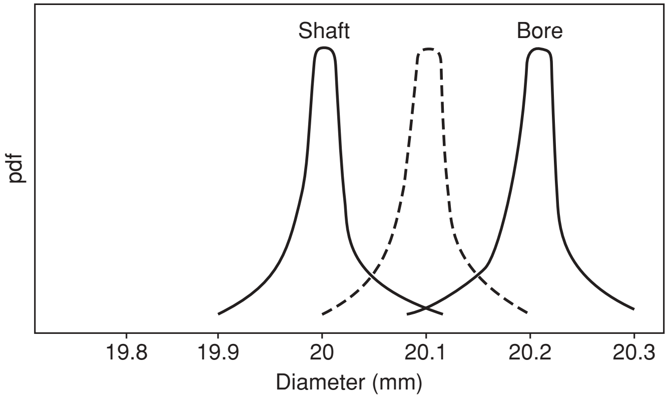


Figure 7.8 *Shaft–bore interference.*

In this case, if 2σ is 0.1 mm then σ is 0.05 mm. Therefore,

$$\begin{aligned} P_{\mathrm{I}} &= 1 - \Phi\left[\frac{20.2 - 20.1}{(0.05^2 + 0.05^2)^{1/2}}\right] \\ &= 0.08 \end{aligned}$$

This is a very simple example to illustrate the principle. Ryan (2000) covers statistical tolerancing in more detail. For systems such as electronic circuits, where tolerances of many components must be considered, statistical analysis of parameter tolerances or drift can provide more economic designs, since the probability of several parts being near their specification limits is much lower than for one or two parts. Statistical tolerancing can also result in lower production costs, since part and subassembly test specifications do not need to be as tight and thus there will be fewer test rejects. However, statistical tolerancing must be based on the correct models, and it is not always safe to assume that variables are normally distributed. For example, many electronic parts are sorted, with parts whose values lie close to the nominal being sold as ‘precision’ parts. Thus the distribution of values of a lot from which such parts have been removed would be bimodal, with no parts being within the sorted range of the nominal value (see Chapters 2 and 9, Figure 9.11).

Methods for analysing multiple simultaneous variations are described in Chapter 11.

7.8.3 Process FMECA

A *process FMECA* (PFMECA) is performed as described above for FMECA, but instead of asking ‘how can the component or function fail?’ we ask ‘how can the process fail?’ Each process task is considered in turn, with the objective of identifying potential problems so that improved methods or controls can be set up. For an example of process FMECA see ReliaSoft (2003).

7.8.4 ‘Poka Yoke’

Poka Yoke is the Japanese expression for ‘mistake proofing’. It is a design approach that considers the ways in which processes might be incorrectly performed, and then making it difficult or impossible to do so. Examples are templates to ensure that directional components cannot be connected the wrong way round.

7.8.5 Testability Analysis

Electronic circuits and systems must be tested after assembly to ensure that they are correct, and to indicate the sources of failure. Testability and testability analysis are described in Chapter 9.

7.8.6 Test Yield Analysis

By using the Monte Carlo capabilities of design software (similar to the methods covered in Chapter 4) and knowledge of test or measurement criteria, yield prediction analyses can be performed on designs that are described in appropriate models.

7.8.7 Maintainability Analysis

Maintenance tasks that might be necessary, such as lubrication, cleaning, replenishment, calibration, failure diagnosis and repair, must all be analysed to ensure that they can be performed correctly by the people likely to be involved. Aspects that should be covered include physical accessibility, time to perform the tasks, skill levels, training requirements, special tools and equipment, and the need for special facilities. Maintenance and maintainability are covered in Chapter 16.

7.9 Assessing the DfR Capability of an Organization

In order to successfully implement the DfR process and deliver reliable products an organization should possess the required tool set, needed expertise, the resources, and the reliability-focused priorities. Therefore an organization needs to be able to evaluate itself and its suppliers on the basis of how capable it is to conduct DfR activities. The evaluation methods for organizational reliability processes are *reliability capability* and *reliability maturity* assessments. They will be covered in Chapter 17.

7.10 Summary

The methods comprising the DfR process can be expensive in engineering time, and this can be a constraint on their effective application. They all involve the detailed consideration of many aspects, such as product characteristics in QFD, system simulation, LSA, failure modes in FMECA, sneak analysis (Chapter 9) and FTA (Chapter 6), and process analysis. However, in addition to improving reliability, DfR can help to reduce overall project time and cost if it is applied effectively, particularly since the prime objective is always to identify and prevent or correct problems early. If product timescales are very tight, which is often the case, it is important that early decisions are made on which methods will be applied and how the results will be used. It might be appropriate to limit the scope of the analyses. For example, the QFD might be limited to a small number of critical design requirements, including variations if appropriate, and the LSA and FMECA to a few identified critical components rather than applied to all. Risk factors described in Section 7.3.5 should be used as guiding principles on how many DfR tools to use and to what extent. The higher the risk of a programme the closer the DfR process should be followed.

An important result of the development of modern engineering design and analysis software is the possibility of reducing the need for development testing. This can be a very worthwhile objective, since projects are nearly always under pressure to reduce development costs and time. However, no design analysis software can deal with the whole range of possible operating stresses, environments, variations and degradation mechanisms that can cause failures. All design analysis and simulation methods involve assumptions and simplifications that can, to varying degrees, generate erroneous or misleading results.

Whilst it is always possible in theory to analyse the effects of variation by performing analyses with parameter values set at, say, tolerance limits or over tolerance ranges, and most CAE software includes facilities for tolerance analysis, it is often difficult and time-consuming to perform such analyses effectively. In particular, analysis implies that distribution parameters and interaction effects are known. As explained in Chapter 2, these aspects are often very uncertain.

Ultimately, only the actual hardware embodies the whole truth of the design, particularly aspects which might have been neglected or misrepresented in the analysis. *The need for testing is a direct consequence of the uncertainty gaps that arise as a result of the limitations inherent in all design analysis.* Therefore the results of the analyses should be used to help to plan and prioritize the tests, and the engineers involved should be part of the test team. Test methods are described in Chapter 12, and we will discuss the management aspects of integrating the design and test activities in Chapter 17.

Questions

1. Produce a failure mode and effect analysis (FMEA) for five components in *one* of the following systems: (i) a domestic washing machine; (ii) the braking system of an automobile; (iii) a simple camera; (iv) a portable transistor radio; or (v) any other system with which you are familiar (giving a brief explanation of its function). Your answer should be properly laid out as if it formed part of a complete FMEA on the system. Explain the additional considerations that would be included to convert your FMEA into a FMECA.
2. Describe the main uses to which a completed FMECA can be applied.
3. Describe three methods that can be used to analyse and improve the processes that will be used for the manufacture of a new product design.
4. Explain briefly, using diagrams if appropriate, the following methods:
   a Quality function deployment.
   b Process FMEA.
   c Hazard and operability study.
   d Poka yoke.
   e Critical items list.
5. Comment on the values you would apply to the likelihood of failures if you are performing a FMECA for each of the following (consider the availability of data, its credibility and the purpose of the analysis):
   a Operation of a car seat positioning system.
   b Operation of a train braking system.
   c Test coverage of an electronic circuit being used in a mobile telephone.
   d Operation of the fuse of a hydrogen bomb.
6. Give four questions that would be appropriate for the reliability aspects of a design review of either a high speed mechanism involving bearings, shafts and gears and high mechanical stress, or for design of a DC electrical power supply unit which uses a standard AC power input.

7. Discuss the ways by which the design review process should be managed in order to provide the most effective assurance that new product designs are reliable, produceable and maintainable. Comment on the organizational and procedural aspects, as well as the actual conduct of the review.
8. Discuss the factors you would consider in producing a Reliability Critical Items List for
   a a modern electronically controlled washing machine.
   b a fighter aircraft electronic box.
9. Develop a QFD House of Quality for a common system such as a coffee maker, a hand calculator, a digital camera, a water heater, a lawnmower, a microwave oven, an electronic thermostat, a TV remote control, a vacuum cleaner or other everyday product. Use your own or commonly available technical knowledge about the device.
10. Research the Internet on the subject of DFSS (Design for Six Sigma). Make a list of common features and the list of major differences between DFSS and DfR.
11. What would you consider as the very first step in the FMECA process?
12. Since the RPN index in FMECA is somewhat subjective, how would you determine the minimum RPN value for the critical items needing attention and corrective actions?
13. Consider design situations where DRBFM would be beneficial and where it would not be beneficial. Give one example of each.
14. Consider the DfR process flow Figure 7.2. Describe the cost factors for each of the DfR stages. What would be the contributing cost factors to correcting a design error at each of the six DfR stages?
15. How would you consider the role of industry standards in successful implementation of the Design for Reliability process?

Bibliography

AIAG (2003) *Potential Failure Mode & Effects Analysis: FMEA-3*. Available at www.aiag.org.

Akao, Y. (1990) *QFD: Quality Function Deployment - Integrating Customer Requirements into Product Design*, Productivity Press.

Allan, L. (2008) *Change Point Analysis and DRBFM: A Winning Combination. Reliability Edge* (published by ReliaSoft), volume 9, issue 2. Available at http://www.reliasoft.com/newsletter/v9i2/drbfm.htm.

British Standard, BS 5760. *Reliability of Systems, Equipment and Components*. British Standards Institution, London.

Brombacher, A. (1999) *Maturity Index on Reliability: Covering Non-technical Aspects of IEC61508*. Reliability Certification Reliability Engineering & System Safety, **66**(2), 109–120.

Clausing, D. (1994) *Total Quality Development: a Step by Step Guide to World Class Concurrent Engineering*, ASME Press.

Creveling C., Slutsky, J. and Antis, D. (2003) *Design for Six Sigma in Technology and Product Development*, Prentice Hall.

Crowe, D. and Feinberg, A. (2001) *Design for Reliability (Electronics Handbook Series)*, CRC Press, Boca Raton.

Mettas, A. (2010) *Design for Reliability: Overview of the Process and Applicable Techniques*. International Journal of Performability Engineering (IJPE), **6**(6), November 2010 - Paper 4 - 577–586.

ReliaSoft (2001) *Using Degradation Data for Life Data Analyses*, Reliability Edge, Volume 2, Issue 2. Available at: http://www.reliasoft.com/newsletter/2q2001/degradation.htm.

ReliaSoft (2003) *Process FMEA sample (Front Door L.H)*, Available at http://www.weibull.com/basics/fmea_fig1.htm.

ReliaSoft (2007) *Fundamentals of Design for Reliability: RS 560 Course Notes*. Available at http://www.reliasoft.com/seminars/gencourses/rs560.htm.

ReliaSoft (2008) *Design for Reliability: Overview of the Process and Applicable Techniques*. Reliability Edge, Volume 8, Issue 2. http://www.reliasoft.com/newsletter/v8i2/reliability.htm.

Ryan, T. (2000) *Statistical Methods for Quality Improvement, Wiley*.

SAE J-1739, *Potential Failure Mode and Effects Analysis*. Society of Automotive Engineers, USA.

Silverman, M. (2010) *How Reliable is Your Product? 50 Ways to Improve Product Reliability*, Super Star Press, Silicon Valley, California.

Stamatis, D. (2003) *Failure Mode and Effect Analysis: FMEA from Theory to Execution*, 2nd edn, American Society for Quality (ASQ) Press.

US MIL STD1629A: *Failure Mode and Effects Analysis*. National Technical Information Service, Springfield, Virginia.

Wickens, C., Lee, J., Liu Y. and Gordon-Becker, S. (2003) *Introduction to Human Factors Engineering*, 2nd edn, Prentice Hall.

8

Reliability of Mechanical Components and Systems

8.1 Introduction

Mechanical components can fail if they break as a result of applied mechanical stresses. Such failures occur primarily due to two causes:

1 Overstress leading to fracture. Stresses may be tension, compression or shear. Bending stresses cause tensile and compressive forces, but fracture usually occurs in tension.
2 Degradation of strength, so that working stresses cause fracture after a period of time.

For example, a pressure vessel will burst if the pressure exceeds its design burst strength, or if a crack or other defect has developed to weaken it sufficiently.

Mechanical components and systems can also fail for many other reasons, such as (though this list is by no means exhaustive):

– Backlash in controls, linkages and gears, due to wear, excessive tolerances, or incorrect assembly or maintenance.
– Incorrect adjustments on valves, metering devices, and so on.
– Seizing of moving parts in contact, such as bearings or slides, due to contamination, corrosion, or surface damage.
– Leaking of seals, due to wear or damage.
– Loose fasteners, due to incorrect tightening, wear, or incorrect locking.
– Excessive vibration or noise, due to wear, out-of-balance rotating components, or resonance.

Designers must be aware of these and other potential causes of failure, and must design to prevent or minimize their occurrence. Appreciation of 'Murphy's Law' ('if a thing can go wrong, it will') is essential, particularly in relation to systems which are maintained and which include other than simple operator involvement. This chapter will describe overload and strength deterioration, and relevant aspects of component and material selection and manufacturing processes.

Practical Reliability Engineering, Fifth Edition. Patrick D. T. O'Connor and Andre Kleyner.

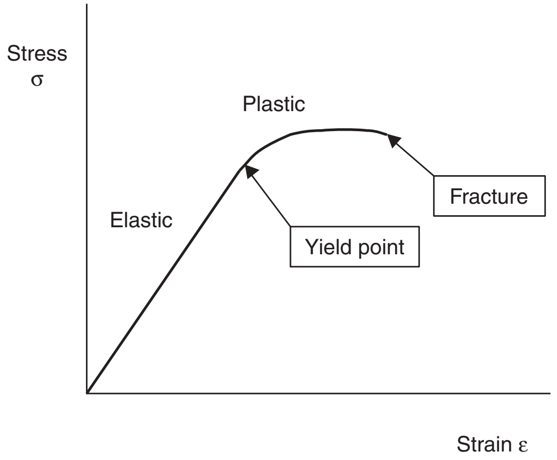


Figure 8.1 *Material behaviour in tensile stress.*

8.2 Mechanical Stress, Strength and Fracture

Mechanical stress can be either *tensile*, *compressive* or *shear*. Tensile stress is caused when the material is pulled, so that the stress attempts to overcome the internal forces holding the material together. Typical material behaviour in tension is shown in Figure 8.1. This shows that, as stress increases, the material stretches proportionally to the stress (the *elastic region*), then begins to stretch more rapidly (the *plastic region*), and finally fractures. In the elastic region the material will return to the original unstressed length if the stress is removed. The amount of deformation is called the *strain*. In the plastic region the material will retain some or all of the deformation if the stress is removed. Fracture occurs when sufficient energy has been applied to overcome the internal forces.

Stress is the load per unit cross-sectional area, conventionally expressed as σ, and is measured in kg/m^2, lbs/in^2 (psi), or pascals (Pa) (N/m^2). The strain (ε) is the ratio of the change in length to the original length. The relationship between stress and strain is described by *Hooke's Law:*

$$\sigma = E\varepsilon \tag{8.1}$$

where E is *Young's Modulus*, or the *modulus of elasticity* for the material. A high value of E indicates that the material is *stiff*. A low value means that the material is soft or *ductile*.

The strength of a material in tension is measured by its *Yield Strength* (the stress at which irreversible plastic deformation begins) or *Ultimate Tensile Strength* (UTS), the stress at which fracture occurs. Note that the UTS might be lower than the yield strength.

When a specimen is subjected to tensile stress narrowing or 'necking' occurs, so that the cross-sectional area is reduced. This causes the 'true' stress level to increase compared with the 'engineering' stress calculated on the basis of the original unstressed cross-section. However, engineering design practice generally limits stress to not more than about 0.2 % strain, so the engineering stress–strain relationships are mostly used.

The elastic/plastic/fracture behaviour of a material is determined by its atomic or molecular structure. Atoms in solids are bound together by the interatomic or intermolecular attractive forces. E is proportional to the interatomic spacing, and it is reduced if temperature is increased. Elastic deformation extends the interatomic distances. Plastic deformation occurs when the energy applied is sufficient to cause the atomic planes, for example in a crystal, to slip along the lattice structure and take up new stable conditions. Material surfaces contain energy, in the same way as the surface of a liquid possesses surface tension, due to the fact that the interatomic attractive forces between atoms at the surface can act in only two dimensions and so do not cancel as they do within the bulk material. In solids this energy is much higher than in liquids. When fracture occurs, two new surfaces are created. This extra energy is imparted by the applied stress which causes the fracture. Knowing the surface energy of a material enables us to determine the theoretical strength. This far exceeds what we actually measure, by factors of 1000 to over 10 000. The reason for the difference is that some plastic deformation occurs at stresses much lower than the theoretical elastic limit, as actual materials contain defects that create stress concentrations, for example dislocations within crystal planes of crystalline materials (metals, metal alloys, silicon, carbon, etc.), and between molecular boundaries in amorphous materials like plastics. Very pure single crystals, such as carbon fibres, can be produced with strengths that approach the theoretical values. The practical strength of a material can be determined only by tests to failure, though theoretical knowledge of aspects such as crystal structure, uniformity, and so on, enable materials scientists to make approximate forecasts of strength.

Another important material property is *toughness*. Toughness is the opposite of *brittleness*. It is the resistance to fracture, measured as the energy input per unit volume required to cause fracture. This is a combination of strength and ductility, which is represented by the area under the stress–strain curve. Figure 8.2 shows this schematically (and very generally) for different material types.

The different patterns of behaviour represent the properties of ductility, brittleness and toughness. A ductile, weak material like pure copper will exhibit considerable strain for a given stress, and will fracture at low stress. A tough material like kevlar or titanium will have little strain and a high UTS. A brittle material like cast iron, glass or ceramic will show very little strain, but lower resistance to rapid stress application such as impact loads. Material properties, especially of metals, vary widely as a result of processes such as heat treatment and machining. In practice materials are applied so that the maximum stress is always well below the yield strength, by a factor of at least 2.

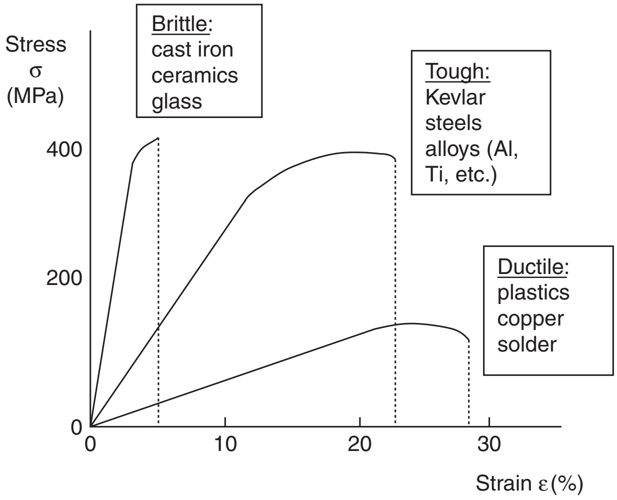


Figure 8.2 *Stress–strain for different materials (generalized).*

A crack will grow if the energy at the crack tip is sufficient to overcome the interatomic forces, and thus open it further. *Griffith's Law* expresses this:

$$\sigma = \sqrt{\frac{2E\gamma}{\pi a}} \tag{8.2}$$

where: $\sigma =$ maximum stress at crack tip (note: not the average stress in the material).
$E =$ modulus of elasticity.
$\gamma =$ surface energy.
$a =$ half crack length.

The maximum stress at the tip of a crack (or at any other defect or *stress raiser*) is proportional to the applied total stress, the size of the crack or defect, and the sharpness of the tip around which the stress is applied. The ratio of maximum stress to applied stress is the *stress concentration factor*. Whilst the total or average stress can usually be determined quite accurately, using methods such as finite element analysis, the maximum stress around a stress raiser, and thus the strength, is often much less certain. Stress concentrations can be reduced by designing to provide adequate radii of curvature on corners of stressed components, ensuring that material surfaces are smooth, and, in the case of cracks in sheet material, by drilling a hole at the tip of the crack to increase the radius.

Compressive strength is much more difficult to analyse and predict. It depends upon the mode of failure (usually buckling for most engineering materials and components such as steel or aluminium alloy vehicle panels, struts, and electrical connector pins) and the shape of the component. Compressive fracture can also occur, particularly in brittle materials.

Structures that have bending loads applied are subjected to both tensile and compressive stress. The upper part of a loaded cantilevered beam will be in tension, whilst the lower part will be in compression.

Stress can also be applied in *shear*. A common practical example is solder joints that connect surface-mounted electronic components (integrated circuit packages) to circuit boards: during operation the temperature rise in the component causes thermal expansion relative to the circuit board, thus applying shear stresses to the solder joints.

The discussion above has presented a very brief overview of the topic. In most cases of applied stress the material behaviour is more complex, since combined effects occur. For example, a component in tensile stress will be caused to be compressed in the directions perpendicular to the tensile stress, so there will be a compressive stress also. Bending loads cause varying tensile and compressive stress from top to bottom of the beam, and therefore shear stress within the beam. Fracture in compression might be caused by shear stresses generated in the material. Finite element analysis (FEA), using modern software, can be used to analyse complex loading situations. However, it is nearly always necessary to test structural components to determine their true strength, especially if the designs are not simple.

8.3 Fatigue

Fatigue damage within engineering materials is caused when a repeated mechanical stress is applied, the stress being above a limiting value called the fatigue limit. Fatigue damage is cumulative, so that repeated or fatigue limit above the fatigue limit will eventually result in failure. For example, a spring subjected to cyclic extension beyond the fatigue limit will ultimately fail in tension.

Fatigue is a very important aspect of reliability of structures subject to repetitive stress, for example from repeated load application, aerodynamic loading, and vibration, since the critical stress can be less than a quarter of the static fracture strength, and fracture can occur after 10^7 to 10^8 cycles when the applied stress is less than half the static strength.

The fatigue damage mechanism is the formation of microcracks resulting from the energy imparted to crystal boundaries by the cyclic stresses. The cracks then continue to extend along these lines of weakness, which act as stress concentrators. Like static fracture mechanics, quantification and prediction is largely empirical and based on experiment, but the degree of uncertainty is much higher.

Initiation and growth rate of the cracks varies depending upon the material properties and on surface and internal conditions. The material property that imparts resistance to fatigue damage is the toughness. As described above, the stresses around the tip of a crack or other defect (such as a machining scratch on a component or a void or inclusion in a casting or forging) are much higher than those in the bulk of the material, so concentrating the energy at these locations. We can demonstrate this easily by repeatedly bending a straightened paperclip through 180 °. Being of ductile material, the clip will not fracture on the first bending. However, the alternating tensile and compressive stresses will generate cumulative fatigue damage, leading to fracture after typically about 20 cycles. If we now repeat the experiment, but now test paperclips which have been lightly cut with a sharp modelling knife, they will fracture in typically five cycles or fewer.

Figure 8.3 shows the general, empirical relationship between stress and cycles to fracture. This is a log-log plot of the stress σ and the number of cycles N to failure, which is called the *S–N* curve. Below the fatigue limit the life is indefinite, but higher stress levels induce cumulative damage, leading ultimately to failure. The S–N curve indicates the cycles to failure at any cyclic stress value between the ultimate stress and the fatigue limit σ' and analytically expressed by Eq. (8.3).

$$N = A\sigma^{-b} \tag{8.3}$$

where: N = Number of cycles to failure.
σ = cyclic stress value.
b = fatigue exponent.
A = empirical constant.

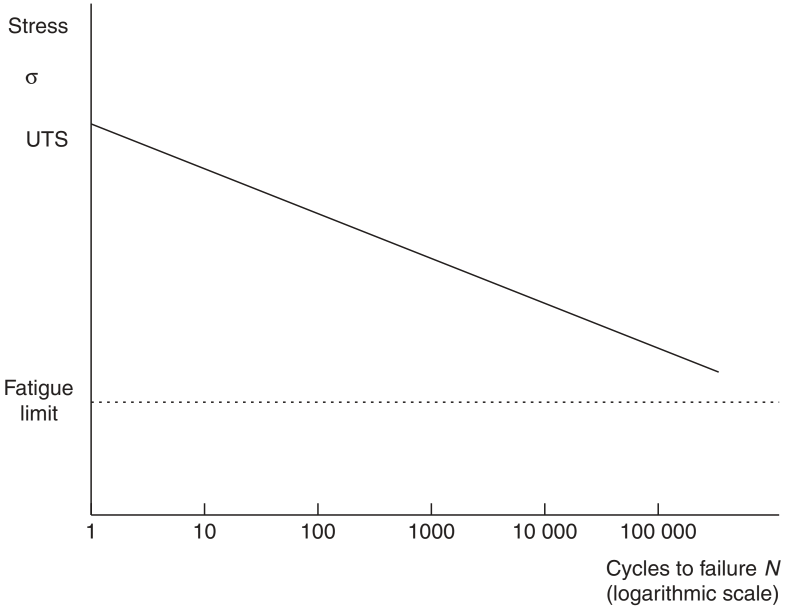


Figure 8.3 *S–N curve.*

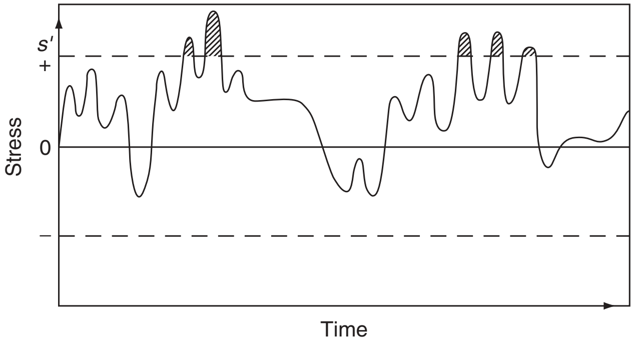


Figure 8.4 *Random overload.*

The curve indicates the mean value of cyclic load for a given number of cycles to failure (or vice versa). The population cycles to failure would in fact be distributed.

The basic *S–N* curve shows the simplest situation, in which a uniform cyclic load is applied. In the more general practical case, with randomly distributed stresses as shown in Figure 8.4, the population distribution of cycles to failure will have an additional variability and we will not know how much damage has been inflicted.

The fatigue life of an item subject to varying stress can be estimated using Palmgren-Miner's Law (more often referred as Miner's Rule). This is expressed as

$$\frac{n_1}{N_1}+\frac{n_2}{N_2}+\frac{n_3}{N_3}+\cdots+\frac{n_k}{N_k}=1$$
$$\sum_{i=1}^{k}\frac{n_i}{N_i}=1 \qquad (8.4)$$

where n_i is the number of cycles at a specific stress level, above the fatigue limit, and $\mathrm{n_i}$ is the median number of cycles to failure at that level, as shown on the *S–N* curve.

The fatigue life of an item subject to an alternating stress with a mean value of zero is

$$N_\mathrm{e}=\sum_{i=1}^{k} n_i \qquad (8.5)$$

N_e is called the *equivalent life*, and when used with the *S–N* diagram gives an equivalent steadily alternating stress, at which damage will occur at the same rate as under the varying stress conditions.

Example 8.1

Load data on a part indicate that there are three values which exceed the fatigue limit stress of $4.5 \times 10^8\ \mathrm{Nm}^{-2}$. These values occur during operation in the following proportions:

$5.5 \times 10^8\ \mathrm{Nm}^{-2}$: 3
$6.5 \times 10^8\ \mathrm{Nm}^{-2}$: 2
$7.0 \times 10^8\ \mathrm{Nm}^{-2}$: 1

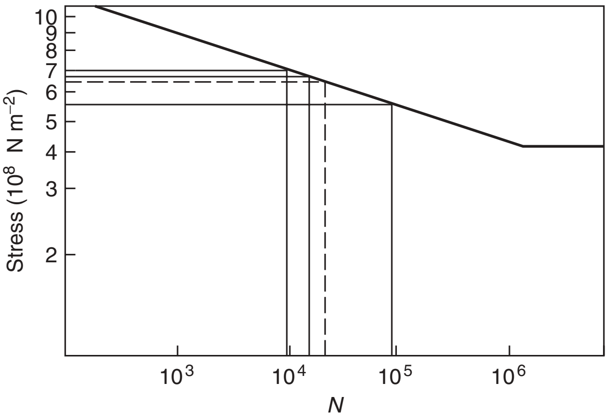


Figure 8.5 *S–N diagram for the part in Example 8.1.*

Evaluate the equivalent constant dynamic stress.

The *S–N* diagram for the material is shown in Figure 8.5. The cycles to failure at each overstress level are:

$$5.5 \times 10^8\,\text{Nm}^{-2} : 9.5 \times 10^4 \text{ cycles}$$
$$6.5 \times 10^8\,\text{Nm}^{-2} : 1.5 \times 10^4 \text{ cycles}$$
$$7.0 \times 10^8\,\text{Nm}^{-2} : 0.98 \times 10^4 \text{ cycles}$$

Therefore, from (8.4) where *C* is an arbitrary constant,

$$\frac{3C}{9.5 \times 10^4} + \frac{2C}{1.5 \times 10^4} + \frac{1C}{0.98 \times 10^4} = 1$$
$$C = 3746$$

From (8.5)

$$\begin{aligned} N_\mathrm{e} &= 3C + 2C + C \\ &= 2.25 \times 10^4 \text{ cycles} \end{aligned}$$

From the *S–N* diagram, the equivalent constant dynamic stress is $6.3 \times 10^8 \,\dfrac{N}{m^2}$.

Figure 8.6 shows an *S–N* diagram for a population of items, with the strength and applied stress distributions also shown. The stress distribution tail extends beyond S′, thus generating fatigue damage, and the mean of the strength distribution is therefore reduced. The strength distribution variance increases as items incur different amounts of fatigue damage. At *N′*, the tails of the load and strength distribution interfere, and we enter the increasing hazard rate period.

Population times to failure in fatigue are typically log normal or Weibull distributed, as shown by the pdf. The variance is large, typically an order of magnitude even under controlled test conditions, and much larger under random service environments, particularly when other factors such as temperature stress, corrosion,

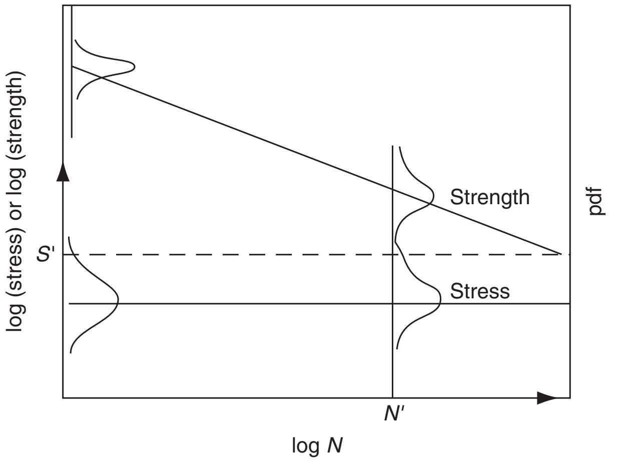


Figure 8.6 *Strength deterioration with cyclic stress.*

damage, or production variation extend the left-hand tail of the life distribution. Therefore fatigue lives are predicted conservatively, particularly for critical components and structures. However, reliability values calculated in this way are subject to considerable uncertainty. The usual practice in designing for a safe life is to estimate the equivalent cycles to failure and to assign a safe life based upon the expected variation in N_e. However, the predicted safe life should always be confirmed by carrying out life tests, using simulated or actual environments, and actual production items.

Analyses of time to failure in test and service situations can be performed using Weibull probability plots. Weibull distributions of times to failure show a positive failure-free life (γ) and slope (β) values greater than 1 (typically 2–3.5), that is, an increasing hazard rate with failures starting after the safe life interval. This life is sometimes referred as the '*B*-life' (see Chapter 3, Section 3.4.5 for more details). *B*-lives are also used to define the lives of components subject to wear, e.g. bearings.

Generally for metals the fatigue life is not affected by the rate at which stress cycling is applied. This is due to the fact that, since they are good thermal conductors, any energy converted to heat is readily conducted away so there is little or no temperature rise. However, plastics generally are more likely to be locally heated by high rates of stress reversals, and this, coupled with their lower melting points and other properties, such as the glass transition temperature, can result in reduced fatigue lives at high cycle rates.

Composite materials, such as fibre-reinforced structural components, can be designed and manufactured to have tailored mechanical properties, since the stresses are transmitted primarily through the fibres rather than through the bulk material. Failure of composite components can be due to delamination or separation of the fibres, or fracture of the whole component.

Fatigue life is affected by other factors, mainly temperature and corrosion. High temperature accelerates crack growth rates, by maintaining the critical energy levels at the crack tips. Corrosion can greatly accelerate crack propagation. Complex loading situations, for example vibration superimposed on a static load, can also reduce fatigue life.

The fracture surfaces of fatigue failures typically show characteristic 'rings' spreading out from the initial fatigue crack as it progressively grows, and a granular area where the final fracture occurs (Figure 8.7).

The fatigue behaviour of designs can be analysed using software that combines FEA and material property data. However, the software assumes (unless otherwise instructed) that material surfaces are smooth and not damaged, and that no other effects such as corrosion are present. Software for fatigue life prediction

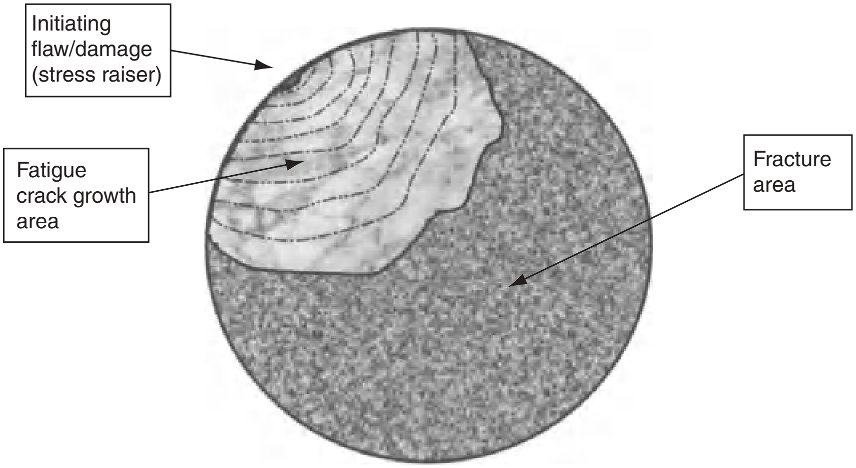

Initiating flaw/damage (stress raiser)
Fatigue crack growth area
Fracture area

Figure 8.7 *Typical fatigue failure (schematic).*

evaluates expected (average) lifetimes and variations around these, not the possible time to the first failure. The correctness of the outputs depends on the correctness of the input descriptions such as surface conditions, the adequacy of the mesh being used, and understanding of the underlying mechanics and physics. Small errors or omissions in the mesh or other inputs can diverge and result in large errors in the predicted behaviour. The references under fracture mechanics in the bibliography provide good introductions to the material properties described above.

8.3.1 Design against Fatigue

Design for reliability under potential fatigue conditions means either ensuring that the distributed load does not exceed the critical load or designing for a limited 'safe life', beyond which the item is not likely to be used or will be replaced in accordance with a maintenance schedule. If we can ensure these conditions, then failure should not occur.

However, in view of the wide variation of fatigue lives and the sensitivity to stress and other environmental and material conditions, this is not easy. The following list gives the most important aspects that must be taken into account:

1 Knowledge must be obtained on the material fatigue properties, from the appropriate data sources, and, where necessary, by test. This knowledge must be related to the final state of the item, after all processes (machining, etc.) which might affect fatigue.
2 Stress distributions must be controlled, by careful attention to design of stress concentration areas such as holes, fixings and corners and fillets. The location of resonant anti-nodes in items subject to vibration must be identified. Finite element and nodal analysis methods are used for this work. (See later section on vibration.)
3 Design for 'fail safe', that is, the load can be taken by other members or the effect of fatigue failure otherwise mitigated, until the failed component can be detected and repaired or replaced. This approach is common in aircraft structural design.
4 Design for ease of inspection to detect fatigue damage (cracks), and for ease of repair.

5 Use of protective techniques, such as surface treatment to relieve surface stresses (shot peening, heat treatment), increasing surface toughness (nitriding of steels, heat treatment), or provision of 'crack stoppers', fillets added to reduce the stress at crack tips.
6 Care in manufacture and maintenance to ensure that surfaces are not damaged by scratches, nicks, or impact.

8.3.2 Maintenance of Fatigue-Prone Components

It is very important that critical components subject to fatigue loading can be inspected to check for crack initiation and growth. Maintenance techniques for such components include:

1 Visual inspection.
2 Non-destructive test (NDT) methods, such as dye penetrants, X-ray and acoustic emission tests.
3 Where appropriate, monitoring of vibration spectra.
4 Scheduled replacement before the end of the fatigue life.

The scheduling and planning of these maintenance techniques must be based upon knowledge of the material properties (fatigue life, crack propagation rates, variability), the load duty cycle, the effect of failure, and test data. See Chapter 16 for a more detailed discussion of maintenance planning principles.

8.4 Creep

Creep is the gradual increase in length of a component that is subjected to combined continuous or cyclic tensile stress and high temperature. Creep is a plastic (i.e. permanent) deformation, which occurs when the material temperature exceeds about 50 % of the melting point, on the absolute temperature scale. The effect is significant with components like turbine discs and blades in gas turbine engines, due to the combined very high temperatures and centrifugal forces. It has recently become a problem in electronics assemblies using surface mount components. Since solder melts at about 183 °C, system operating temperatures are generally within the creep temperature range. Therefore, permanent deformation takes place due to the shear stresses imposed by thermal cycling. The deformation in turn can result in higher shear stresses, thus accelerating the fatigue mechanism.

8.5 Wear

8.5.1 Wear Mechanisms

Wear is the removal of material from the surfaces of components as a result of their movement relative to other components or materials. Wear can occur by a variety of mechanisms, and more than one mechanism may operate in any particular situation. The science and methods related to understanding and controlling wear in engineering comprise the discipline of *tribology*. The main wear mechanisms are described below.

Adhesive wear occurs when smooth surfaces rub against each other. The contact load causes interactions between the high spots on the surfaces and the relative motion creates local heating and dragging between the surfaces. This results in particles being broken or scraped off the surfaces, and loose particles of wear debris are generated.

Fretting is similar to adhesive wear, but it occurs between surfaces subject to small oscillatory movements. The small movements prevent the wear debris from escaping from the wear region, so the particles are broken up to smaller sizes and might become oxidized. The repeated movements over the same parts of the surface also result in some surface fatigue, and corrosion also contributes to the mechanism.

Abrasive wear occurs when a relatively soft surface is scored by a relatively hard surface. The wear mechanism is basically a cutting action often with displacement of the soft material at the sides of grooves scored in the soft material.

Fluid erosion is caused to surfaces in contact with fluids, if the fluid impacts against the surfaces with sufficient energy. For example, high velocity fluid jets can cause this type of damage. If the fluid contains solid particles the wear is accelerated. Cavitation is the formation and violent collapse of vacuum bubbles in flowing liquids subject to rapid pressure changes. The violent collapse of the vacuum bubbles on to the material surfaces causes fluid erosion. Pumps, propellers and hydraulic components can suffer this type of damage.

Corrosive wear involves the removal of material from a surface by electrolytic action. It is important as a wear mechanism because other wear processes might remove protective films from surfaces and leave them in a chemically active condition. Corrosion can therefore be a powerful additive mechanism to other wear mechanisms.

8.5.2 Methods of Wear Reduction

The main methods of wear reduction are:

1 Minimize the potential for wear in a design by avoiding as far as practicable conditions leading to wear, such as contact of vibrating surfaces.
2 Selection of materials and surface treatments that are wear-resistant or self-lubricating.
3 Lubrication, and design of efficient lubricating systems and ease of access for lubrication when necessary.

When wear problems arise in use, an essential starting point for investigation is examination of the worn surfaces to determine which of the various wear mechanisms, or combinations of mechanisms, is involved. For example, if a plain bearing shows signs of adhesive wear at one end, the oil film thickness and likely shaft deflection or misalignment should be checked. If the problem is abrasive wear the lubricant and surfaces should be checked for contamination or wear debris.

In serious cases design changes or operational limitations might be needed to overcome wear problems. In others a change of material, surface treatment or change of lubricant might be sufficient. It is also important to ensure that lubricant filtration, when appropriate, is effective.

8.5.3 Maintenance of Systems Subject to Wear

The life and reliability of components and systems subject to wear are very dependent upon good maintenance. Maintenance plans should be prepared, taking into account cleaning and lubrication requirements, atmospheric and contamination conditions, lubricant life and filtration, material properties and wear rates, and the effects of failure. In appropriate cases maintenance also involves scheduled monitoring of lubricant samples, using magnetic plugs to collect ferrous particles and spectroscopic oil analysis programmes (SOAP) to identify changes in levels of wear materials. Vibration or acoustic monitoring is also applied. These techniques are used in systems such as industrial and aero engines, gearboxes, and so on.

Neale (1995) and Summers-Smith (1994) are excellent introductions to tribology and wear.

8.6 Corrosion

Corrosion affects ferrous and some other non-ferrous engineering metals, such as aluminium and magnesium. It is a particularly severe reliability problem with ferrous products, especially in damp environments. Corrosion can be accelerated by chemical contamination, for example by salt in coastal or marine environments.

The primary corrosion mechanism is oxidation. Some metals, particularly aluminium, have oxides which form as very hard surface layers, thus providing protection for the underlying material. However, ferrous alloys do not have this property, so oxidation damage (rust) is cumulative.

Galvanic corrosion can also be a problem in some applications. This occurs when electromotive potentials are built up as a result of dissimilar metals being in contact and conditions exist for an electric current to flow. This can lead to the formation of intermetallic compounds and the acceleration of other chemical action. Also, electrolytic corrosion can occur, with similar results, in electrical and electronic systems when induced currents flow across dissimilar metal boundaries. This can occur, for example, when earthing or electrical bonding is inadequate. Electrolytic corrosion affects the most electrically active element in the circuit.

Stress corrosion is caused by a combination of tensile stress and corrosion damage. Corrosion initiates surface weaknesses, leading to crack formation. Further corrosion and weakening occurs at crack tips, where the metal is in a chemically active state and where the high temperatures generated accelerate further chemical action. Thus the combined effect can be much faster than either occurring alone.

Design methods to prevent or reduce corrosion include:

1 Selection of materials appropriate to the application and the expected environments.
2 Surface protection, such as anodizing for non-ferrous metals, plasma spraying, painting, metal plating (galvanizing, chrome plating), and lubrication.
3 Other environmental protection, such as the use of dryers or desiccators.
4 Avoidance of situations in which galvanic or electrolytic corrosion can occur.
5 Awareness and avoidance of conditions likely to generate stress corrosion.

Correct maintenance is essential to ensure the reliability of corrosion-prone components. Maintenance in these situations involves ensuring the integrity of the protective measures described above. Since corrosion damage is usually extremely variable, scheduled maintenance should be based upon experience and criticality.

Revie (2011) describes corrosion in detail.

8.7 Vibration and Shock

Components and assemblies can be subjected to vibration and shock inputs, during use, transport or maintenance. Vibration and shock can cause:

- Fracture due to fatigue, or due to mechanical overstress.
- Wear of components such as bearings, connectors, and so on.
- Loosening of fasteners, such as screws, bolts, and so on.
- Leaks in hydraulic and pneumatic systems, due to wear of seals or loosening of connectors.
- Acoustic noise (10–10 000 Hz).

Common vibration inputs are:

- Reciprocating or rotating machinery. The dominant vibration frequency (Hz) generated by rotating masses will be rm./60.
- Wheel vibration, on road and rail vehicles.
- Aerodynamic effects on aircraft and missile structures.

- Pressure fluctuations in hydraulic and pneumatic systems.
- Acoustic noise.

Vibration of a structure may occur at a fixed frequency, at different frequencies over time, or simultaneously over a range of frequencies. Vibration over a wide range of simultaneous frequencies is called *broad band* vibration. Vibration can occur in or about different linear and rotating axes.

The important units in relation to sinusoidal vibration are:

- Frequency (Hz).
- Displacement (mm), defined as peak or peak-to-peak values.
- Velocity (m/s), defined as peak values.
- Acceleration (m/s^2 or g_n), defined as peak values.

Every structure has one or more resonant frequencies, and if the vibration input occurs at these, or at harmonics, vibration displacements will be maximized. The locations at which zero vibration displacements occur are called *nodes*, and maximum displacement amplitudes occur at the *anti-nodes*. There may be more than one resonant frequency within the expected environmental range, and these may exist along different axes. There may also be more complex resonance modes, such as torsional or combinations of mechanical, acoustic, rotational or electromechanical modes. Sometimes simultaneously occurring resonances might be important, for example two components that vibrate in different modes or at different frequencies and in so doing impact one another. Examples are electronic circuit boards, hydraulic pipes and vehicle panels.

The resonant frequency is proportional to the stiffness of the structure, and inversely proportional to the inertia. Therefore, to ensure that resonant frequencies are well above any input vibrations that might be applied, structures must be sufficiently stiff, especially where there are relatively heavy parts, such as large components on circuit boards.

The vibration amplitude at any frequency is reduced by damping. Damping can also change the resonant frequency. Damping is provided in hydraulic and pneumatic systems by accumulators, in suspension and steering systems by mechanical dampers, and by using anti-vibration mountings for motors, electronic boxes, and so on.

The pattern of vibration as a function of other parameters, such as engine speed, can be shown on a *waterfall plot*. Figure 8.8 is an example. Waterfall plots help to indicate the sources of vibration and noise. For example, a resonance, that is independent of speed shows as a vertical line, and one that is generated at a particular speed shows peaks running horizontally. The peak heights (or colours on colour map displays) indicate the amplitudes.

Shock loads can cause vibration, though the amplitude is usually attenuated due to inherent or applied damping. Shock loads are only a particular type of vibration input: relatively high intensity and frequency, for short intervals.

Piersol and Paez (2009) provides a comprehensive treatment of the subject and Steinberg (2000) describes applications to electronics. Testing methods for vibration and shock will be covered in Chapter 12.

8.8 Temperature Effects

Failures can be caused by materials being subjected to high or low temperatures. The main high temperature failure modes are:

- Softening and weakening (metals, some plastics).
- Melting (metals, some plastics).

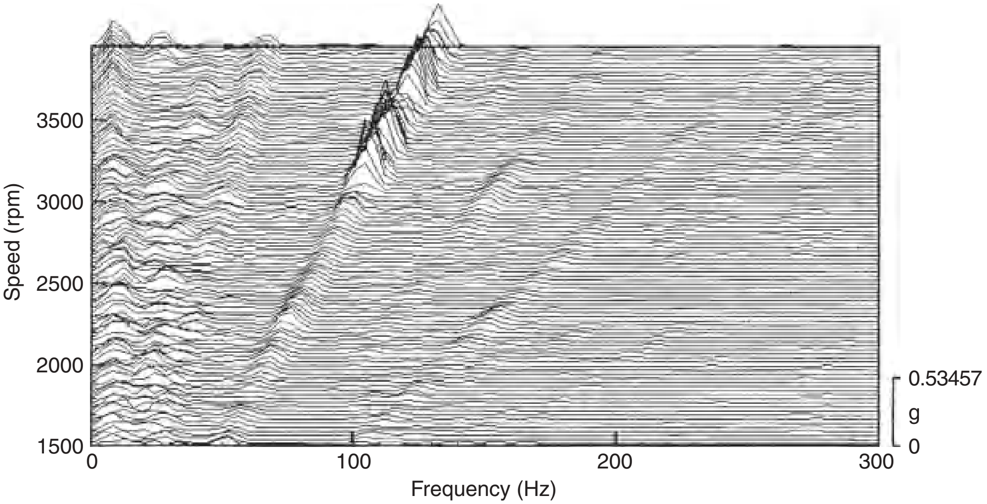


Figure 8.8 *Waterfall plot.*

- Charring (plastics, organic materials).
- Other chemical changes.
- Reduced viscosity or loss of lubricants.
- Interaction effects, such as temperature-accelerated corrosion.

Low-temperature effects can include embrittlement of plastics, increasing viscosity of lubricants, condensation and freezing of condensation or coolants.

Most temperature effects are deterministic (melting points, condensation temperatures, freezing points, viscosities). Effects such as these are not cumulative, so time and numbers of temperature cycles do not directly affect reliability. However, secondary effects might be cumulative, for example the effects of lubricant viscosity on rate of wear.

All materials have a *thermal coefficient of expansion* (TCE). If two components with different TCEs are attached to one another, or two attached components can experience different temperatures, then mechanical stresses will be set up. An important example of this situation is the attachment of electronic components to circuit boards or other substrates, particularly surface mount integrated circuit packages as described in Chapter 9. When the IC is powered and operated heat is generated, so the package temperature rises. The heat is transferred through the package and the solder joints to the circuit board (which might include a ‘heat plane’, to improve heat dissipation). The thermal resistance of the heat flow path from the package to the eventual heat sink will result in the package being hotter than the board, and in the board temperature rise lagging that of the package. If power is cycled, the temperature differences will be also. This will result in cyclic shear stresses being imparted to the solder joints. The magnitude of these stresses can lead to fatigue failures, in the form of cracks running through the joint. These in turn can cause electrical failure, often of an intermittent nature, after a sufficient number of cycles. This type of failure is particularly important in electronic systems that must withstand many on–off cycles, such as engine control systems. If the systems are also subjected to vibration, the combined effects of thermal and vibration cycling can be highly interactive.

Chemical reactions, gaseous and liquid diffusion and some other physical processes are accelerated by increasing temperature. Arrhenius' Law expresses this phenomenon:

$$R = K \exp[-E_A/kT] \tag{8.6}$$

where: R = process rate.
K = constant.
E_A = activation energy for the process (varies depending on the material and/or failure mechanism).
K = Boltzmann's constant.
T = Absolute temperature, K.

Typically, chemical process rates increase by a factor of 2 for every 10–20°C rise in temperature. An important group of processes that can be thermally accelerated is corrosion, particularly rusting of iron and steel.

8.8.1 Humidity and Condensation

Damp environments can cause or accelerate failure processes such as corrosion and mould growth. Temperature and humidity are closely related, humidity being inversely proportional to temperature until the dew point is reached, below which moisture condenses on to surfaces. Liquid water can cause further failures, including:

- Chemical corrosion, if contamination is also present.
- Electrolytic corrosion, by providing an electrolyte.
- Short circuiting of electrical systems, particularly within connectors.
- Mould growth.

Plastics are generally hygroscopic, that is, they absorb moisture, whether above or below the dew point. Therefore any components that are encapsulated in plastics, particularly electronic components and assemblies, are in principle prone to moisture ingress. This presented until fairly recently a major limitation on the application of plastic encapsulated components, since they suffered corrosion of the aluminium conductor metalization when used in high humidity environments. Their use in military and aerospace systems was banned. However, modern components such as integrated circuits have much improved protection against moisture, due to better control of the chip's surface protective layer and control of the plastic material purity and the encapsulating process, so that today there are few limitations on their application, and moisture-related failures are very rare.

8.9 Materials

Selection of appropriate materials is an important aspect of design for reliability, and it is essential that designers are aware of the relevant properties in the application environments. With the very large and increasing range of materials available this knowledge is not easy to retain, and designers should obtain data and application advice from suppliers as well as from handbooks and other databases. A few examples of points to consider in selecting engineering materials for reliability are given below. The list is by no means exhaustive, and it excludes obvious considerations such as strength, hardness, flexibility, and so on, as appropriate to the application.

8.9.1 Metal Alloys

1 Fatigue resistance.
2 Corrosion environment, compatibility.
3 Surface protection methods.
4 Electrochemical (electrolytic, galvanic) corrosion if dissimilar metals in contact.

8.9.2 Plastics, Rubbers

1 Resistance to chemical attack from materials in contact or in the local atmosphere (lubricants, pollutants, etc.).
2 Temperature stability (dimensional, physical), and strength variation at high and low temperature.
3 Sensitivity to ultraviolet radiation (sunlight).
4 Moisture absorption (all plastics are hygroscopic).

8.9.3 Ceramics

Brittleness, fracture toughness.

8.9.4 Composites, Adhesives

1 Impact strength.
2 Erosion.
3 Directional strength.

Crane *et al*. (1997) is an excellent source of information on selection of engineering materials.

8.10 Components

The range of mechanical components is vast, ranging from springs, seals and bearings to engines, pumps and power transmission units. Even amongst the most basic components there is little standardization, and new products and concepts are constantly being developed. It would not be feasible to attempt to provide guidance on the detailed reliability aspects of such a range in this chapter, but some general principles should be applied:

1 All relevant aspects of the component's application must be carefully evaluated, using the techniques described in the previous chapters. Where experience exists of application in another system, all data on past performance should be used, such as modes and causes of failure, application conditions, durability, and so on. It is essential to discuss the application fully with the supplier's applications engineers, to the extent of making them effectively part of the design team, with commitment to the success of the product.
2 Use mature components in preference to new ones unless there are clear overriding reasons of cost, performance, and so on. Novelty, even when the risks seem insignificant, often introduces unpleasant surprises. All new components should be placed on the critical items list (see Chapter 6).

3 Minimize the number of components and of component types. Whether a spring or a hydraulic pump, this approach not only reduces costs of the product and of assembly, but also can improve reliability. For example, where a mechanism such as a paper feed requires springs, cams and levers, careful study of the problem can often reveal ways in which one component can perform more than one function.

4 Pay attention to detail. It is very often the simple design problems which lead to unreliability, because insufficient attention was paid to them. For example, spring attachment lugs on plastic components that break as a result of the hard spring material cutting through (a metal bush could be a solution), and the location of components so that they suffer contamination from water or oil, or are difficult to fit and adjust, are common examples of failure to apply design skills and experience to the ‘simple’ jobs.

8.11 Processes

Designers must be aware of the reliability aspects of the manufacturing processes. Machining processes create variations in dimensions, which can affect wear and fatigue properties. Processes designed to improve material properties must be considered and designed for. For example, heat treatment, metal plating, anodizing, chemical treatment, and painting require careful control if they are to be effective, and the design of the product and of its methods of assembly must ensure that these processes can be applied correctly and efficiently.

Other processes that can affect reliability include:

8.11.1 Fasteners

A huge range of different fastening methods and systems is available, including rivets, bolts and nuts, clamps, adhesives, and so on. Fasteners can loosen under vibration or as a result of temperature cycling. Fasteners can fail due to fatigue, and fatigue cracks can start at holes for rivets and bolts.

Bolts and nuts can be combined with locking devices to prevent loosening. These include deformable plastic inserts, spring washers, crush washers, split pin retainers, adhesives, locking wires, and so on. The integrity of many locking devices can be degraded if they are used more than once. Bolts and nuts used in some applications must be accurately torque loaded to ensure that the correct holding force is applied, and that the fasteners are not over-stressed on assembly.

8.11.2 Adhesives

Adhesives are used for many assembly operations, including aircraft and other vehicle structures, electronic component mountings on to heat sinks, locking of bolts and nuts, and so on. The most commonly used industrial adhesives are epoxy plastics and cyanoacrylics. Epoxies are two-component adhesives which must be mixed shortly before use. Cyanoacrylics are contact adhesives that form an instant bond. Other adhesive compounds and systems include elastomerics (used in applications such as vibration isolation) and adhesive tapes.

All adhesives require careful preparation and cleaning of the surfaces to be bonded, and they all have limitations in relation to the kinds of materials they can bond. Adhesives also have temperature limits, and generally cannot withstand temperatures above 200 °C.

8.11.3 Welding and Soldering

Metals can be joined by welding, and several different welding methods are used, depending on the materials and the application. Steel structures are welded with electric arcs or oxy-acetylene gas torches. Alloys such as of aluminium and magnesium, which burn in oxygen, are arc welded in inert gas (argon). Car assemblies are spot welded by robots applying pressure and high electric current to form resistance welds. Surfaces can also be welded by friction (high pressure and vibration, including ultrasonic welding of gold wire bonds on microelectronic assemblies).

Tin–lead solder has been by far the most common method for connecting electronic and electrical components within systems, though lead-free solders are becoming more widely used. It also serves as a structural connection. Soldering for electronics assembly is described in Chapter 9.

8.11.4 Seals

Seals are used to prevent leaks in systems such as water, oil hydraulic and pneumatic components and pipe connections, around rotating shafts and reciprocating actuator rams, and to protect items in sealed containers. Special seals include those to block electromagnetic radiation from or into electronic equipment enclosures.

The effectiveness of seals is always influenced by control of assembly operations, and often also by maintenance. They are always affected by usage (wear, erosion, etc.), so they tend to degrade over time and use.

Summers-Smith (1994) is a good introduction to engineering seals.

Chapter 15 covers the control of manufacturing processes. However, it is essential that the capabilities and problems of these are given as much consideration in design as aspects such as performance and cost. The manufacturing operations affect these aspects also, so a fully integrated approach, as described in Chapters 7 and 15, must be followed. Production and quality engineers must be included in the design team, and not left to devise production methods and quality standards after the design has been finalized.

Questions

1. Sketch and annotate the general strain behaviour of materials subjected to tensile stress. Show how this differs for brittle, tough and ductile engineering materials, and give examples of each.
2. Explain why the actual mechanical strength of engineering components is very much less than the theoretical strength. How does this difference affect the predictability of strength?
3. Briefly describe the three most common causes of strength degradation of mechanical components. Give examples of each, with descriptions of methods used to prevent or reduce the chances of failure.
4. Miner's rule is used to predict the expected time to failure in fatigue:

   a Write down the mathematical expression for Miner's rule.

   b A component was tested in the laboratory to determine its fatigue life. The test results were as follows:

Stress level ($\times 10^8$ N/m^2)	6.8	8.0	10.0
Mean cycles to failure ($\times 10^5$)	12.7	4.2	0.6

The component will be used in service with these stress levels occurring in the following proportions, respectively:

Proportion of cycles	0.5	0.3	0.2

What will be the expected time to failure in service, if the stress cycle rate is 1000 per hour?

c Comment on the factors that would influence the accuracy of this prediction.

5. A component designed for a cyclic mechanical stress application has been analysed to determine its likely fatigue life. Comment on the approach that you would apply for ensuring that failures do not occur if the component is:
   a A steel mounting bracket for an actuator on an earthmoving machine.
   b An aluminium alloy mounting bracket for a flight control actuator on an aircraft.
   c A plastic part in a copying machine.
6. Two basic approaches can be applied in the design of components and structures that can fail as a result of fatigue damage. These are the fail-safe and safe-life approaches. Describe these, discuss the factors that would determine which approach is appropriate, and give examples of their application.
7. Describe briefly three methods that can be applied to reduce the likelihood of failure of components and structures owing to fatigue.
8. Describe three types of wear processes that can lead to failure of surfaces in moving contact. Describe how one of these can be minimized by designers.
9. Corrosion can cause failure of metallic parts. Describe three corrosion processes. How can each be minimized by designers?
10. You are designing an electronic unit that will be used on an agricultural machine. What failures might be caused by the vibration environment? What steps would you take to minimize these?
11. Describe briefly the effect of temperature on (give temperature values where appropriate, and consider also the effects of temperature cycles):
    a The strength of a solder joint used to retain a heavy electronic component.
    b The properties of a lubricating oil.
    c An electronic unit located outdoors.
    d Corrosion.
12. Fatigue testing of a metal alloy resulted in the S-N curve:

$$N = 1.2 \times 10^{26}\sigma^{-7.2}$$

    Where N is number of cycles to failure and σ is the stress amplitude in MPa. The alloy is used in the design of an aircraft engine that under normal use will experience 120 cycles per second under a stress amplitude of 72.2 MPa. The typical engine will operate 450 hours per year. If the engine is being designed for a 12 yr life, is this the right choice of material?

Bibliography

Fracture mechanics

Anderson, T.L. (2005) *Fracture Mechanics*, CRC Press.
Collins, J.A. (1981) *Failure of Materials in Mechanical Design*, J. Wiley.
Dowling, N.E. (2006) *Mechanical Behaviour of Materials*, 2nd edn, Pearson Education.
Gordon, J.E. (1991) *The New Science of Strong Materials*, Penguin Books.

Wear

Neale, M.J. (1995) *The Tribology Handbook*, 2nd edn, Butterworth-Heinemann.
Summers-Smith, J.D. (1994) *An Introductory Guide to Industrial Tribology*, Mechanical Engineering Publications.

Corrosion

Revie, R.W. (2011) *Uhlig's Corrosion Handbook*, 3rd edn, J. Wiley.

Vibration and shock

Piersol, A.G. and Paez, T.L. (2009) *Harris' Shock and Vibration Handbook*, 6th edn, McGraw-Hill.

Steinberg, D. (2000) *Vibration Analysis for Electronic Equipment*, 3rd edn, Wiley.

Materials and components

Brostow, W. and Corneliussen, R. (1986) *Failure of Plastics*, Hanser.

Crane, F.A., Charles, J.A. and Furness, J.A. (1997) *Selection and Use of Engineering Materials*, 3rd edn, Butterworth-Heinemann.

Summers-Smith, J. (1992) *Mechanical Seal Practice for Improved Performance*, 2nd edn, Mechanical Engineering Publications.

9

Electronic Systems Reliability

9.1 Introduction

Reliability engineering and management grew up largely in response to the problems of the low reliability of early electronic equipment, and many of the techniques have been developed from electronics applications. The design and construction of an electronic system, more than any other branch of engineering, involves the utilization of very large numbers of components which are similar, but over which the designer and production engineer have relatively little control. For example, for a given logic function a particular integrated circuit device might be selected. Apart from choosing a functionally identical device from a second source, the designer usually has no option but to use the catalogued item. The reliability of the device used can be controlled to a large extent in the procurement and manufacturing phases but, as will be explained, mainly by quality control methods. The circuit designer generally has little control over the design reliability of the device. This trend has become steadily more pronounced from the time that complex electronic systems started to be produced. As the transistor gave way to the integrated circuit (IC) and progressively with the advent of large scale integration (LSI) and very large scale integration (VLSI), the electronic system designer's control over some of the major factors influencing reliability has decreased. However, this is changing in some respects as system designs are increasingly implemented on custom-designed or semi-custom integrated circuits. This aspect is covered in more detail later. This is not to say that the designer's role is diminished in relation to reliability. Rather, the designer of an electronic system must be, more than in most other branches of engineering, a member of a team, involving people from other technologies, production, quality control, test planning, reliability engineering and others. Without such a team approach, his or her functionally correct design could be highly unreliable. It is less likely that the designer of a functionally correct hydraulic or mechanical system could be as badly let down. It is important to understand the reasons for this difference.

For the great majority of electronic components and systems the major determinant of reliability is quality control of all of the production processes. This is also true of non-electronic components (subject in both cases to the items being used within specification). However, most non-electronic equipment can be inspected and tested to an extent sufficient to assure that they will operate reliably. Electronic components cannot be easily inspected, since they are nearly always encapsulated. In fact, apart from X-ray inspection of parts for ultra-high reliability products, internal inspection is not generally possible. Since electronic components, once encapsulated, cannot be inspected, and since the size and quantity of modern components dictates that

Practical Reliability Engineering, Fifth Edition. Patrick D. T. O'Connor and Andre Kleyner.

very precise dimensions be held at very high production rates, it is inevitable that production variations will be built into any component population. Gross defects, that is, failure to function within specification, can easily be detected by automatic or manual testing. However, it is the defects that do not immediately affect performance that are the major causes of electronic component unreliability.

Consider a typical failure mechanism in an electronic component: a weak mechanical bond of a lead-in conductor to the conducting material of the device. This may be the case in a resistor, a capacitor, a transistor or an IC. Such a device might function satisfactorily on test after manufacture and in all functional tests when built into a system. No practical inspection method will detect the flaw. However, because the bond is defective it may fail at some later time, due to mechanical stress or overheating due to a high current density at the bond. Several other failure mechanisms lead to this sort of effect, for example, flaws in semiconductor material and defective hermetic sealing. Similar types of failure occur in non-electronic systems, but generally they do not predominate.

The typical 'electronic' failure mechanism is a wearout or stress induced failure of a defective item. In this context, 'good' components do not fail, since the application of specified loads during the anticipated life will not lead to failure. While every defective item will have a unique life characteristic, depending upon the nature of the defect and the load(s) applied, it is possible to generalize about the nature of the failure distributions of electronic components. Taking the case of the defective bond, its time to failure is likely to be affected by the voltage applied across it, the ambient temperature around the device and mechanical loading, for example, vibration. Other failure mechanisms, say a flaw in a silicon crystal, may be accelerated mainly by temperature variations. Defects in devices can result in high localized current densities, leading to failure when the critical value for a defective device is exceeded.

Of course, by no means all electronic system unreliability is due to defective components. Interconnects such as solder joints and wire bonds can be reliability 'weak links' especially in harsh environment applications (automotive, avionics, military, oil drilling, etc.). Other failure mechanisms will be described later in this chapter.

Also the designer still has the task of ensuring that the load applied to components in the system will not exceed rated (or derated values), under steady-state, transient, test or operating conditions. Since electronic component reliability can be affected by temperature, design to control temperatures, particularly localized 'hot spots', is necessary. Thus the designer is still subject to the reliability disciplines covered in Chapter 7.

Electronic system failures can be caused by mechanisms other than load exceeding strength. For example, parameter drifts in components, short circuits due to solder defects or inclusions in components, high resistance relay or connector contacts, tolerance mismatches and electromagnetic interference are examples of failures which may not be caused by load. We will consider these failure modes later, appropriate to the various components and processes which make up electronic systems.

9.2 Reliability of Electronic Components

Electronic components can be caused to fail by most of the same mechanisms (fatigue, creep, wear, corrosion, etc.) described in the previous chapter. Fatigue is a common cause of failure of solder joints on surface mounted components and on connections to relatively heavy or unsupported components such as transformers, switches and vertically mounted capacitors. Wear affects connectors. Corrosion can attack metal conductors on integrated circuits, connectors and other components. Electrical and thermal stresses can also cause failures that are unique to electronics. The main electrical stresses that can cause failures of electrical and electronic components and systems are current, voltage and power. For all of these failure modes there are strong interactions between the electrical and thermal stresses, since current flow generates heat.

It is important to appreciate the fact that the great majority of electronic component types do not have any mechanisms that will cause degradation or failure during storage or use, provided that they are:

- Properly selected and applied, in terms of performance, stress and protection.
- Not defective or damaged when assembled into the circuit.
- Not overstressed or damaged in use.

The quality of manufacture of modern electronic components is so high that the proportions that might be defective in any purchased quantity are typically of the order of less than ten per million for complex components like ICs, and even lower for simpler components. Therefore the potential reliability of well-designed, well-manufactured electronic systems is extremely high, and there are no practical limits to the reliability that can be achieved with reasonable care and expenditure.

9.2.1 Stress Effects

9.2.1.1 Current

Electrical currents cause the temperatures of conductors to rise. If the temperature approaches the melting point, the conductor will fuse. (Of course fuses are used as protective devices to prevent other, more serious failures from occurring.) Heat in conductors is transferred to other components and to insulation materials, primarily by conduction and convection, so thermal damage can be caused to these.

High currents can also cause component parameter values, such as resistance, to drift over time. This effect is also accelerated by high operating temperatures. Figure 9.1 shows an example of this.

Electric currents also create magnetic fields. If oscillating, they can generate acoustic noise and electromechanical vibration. We will discuss electrical interference effects later.

9.2.1.2 Voltage

Voltage stress is resisted by the dielectric strength of the material between the different potentials. The most common examples are the dielectric material between capacitor plates, and the insulation (air or other

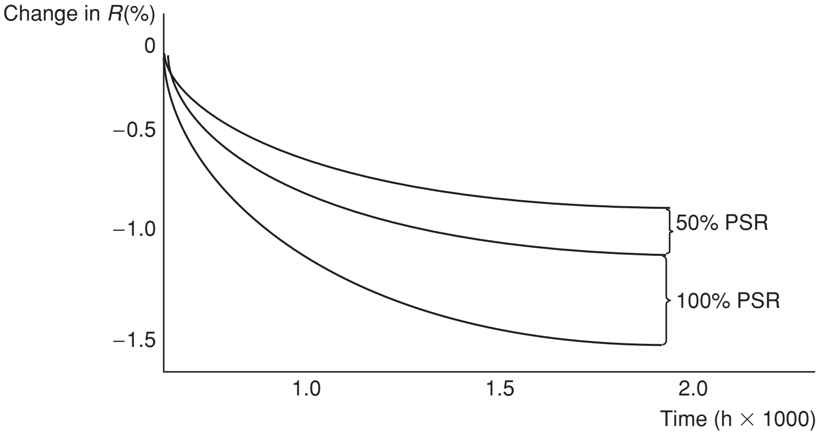


Figure 9.1 *Parameter drift.*

insulator) between conductors. Potential differences generate currents in conductors and components, and if the current carrying capacity is insufficient the conductor or component will fail, in which case the failure mechanism is current, though the cause might be that the voltage is too high. For example, an integrated circuit might fail due to current overstress if a high electrostatic voltage is accidentally applied to it, and a 110 V appliance might fail for the same reason if connected to a 240 V supply.

High voltage levels can be induced by:

- Electrostatic discharge (ESD), caused by charge accumulation on clothing, tools, and so on.
- Other electrical overstress, such as high voltage transients on power lines, unregulated power supplies, circuit faults that lead to components being overstressed, accidental connection of high voltages to low power components, and so on. This is referred to as electrical overstress (EOS).

Another effect of voltage stress is *arcing*, which can occur whenever contacts are opened, for example in switches and relays. Arcing can also occur between brushes and commutators of motors and generators. Arcing generates electromagnetic noise, and also progressively damages the contact surfaces, leading eventually to failure. Arcing can also cause damage to electric motor bearings, if they provide a current path due to inadequate design or maintenance.

Arcing can be reduced by using voltage suppression components, such as capacitors across relay or switch contacts. Arcing becomes more likely, and is more difficult to suppress, if atmospheric pressure is reduced, since the dielectric constant of air is proportional to the pressure. This is why aircraft and spacecraft electrical systems operate at relatively low voltage levels, such as 28 V DC and 115 V AC.

Corona discharge can occur at sharp points at moderate to high voltage levels. This can lead to dust or other particles collecting in the area, due to ionization.

Some components can fail due to very low or zero current or voltage application. Low-power relay contacts which pass very low DC currents for long periods can stick in the closed position due to cold welding of the contact surfaces. Electrical contacts such as integrated circuit socket connectors can become open due to build-up of a thin dielectric layer caused by oxidation or contamination, which the low voltage stress is unable to break down.

9.2.1.3 Temperature

The Arrhenius formula that relates physical and chemical process rates to temperature has been used to describe the relationship between temperature and time to failure for electronic components, and is the basis of methods for predicting the reliability of electronic systems, as described in Chapter 6. However, this is sometimes an erroneous application, since, for many modern electronic components, most failure mechanisms are not activated or accelerated by temperature increase. The materials and processes used are stable up to temperatures well in excess of those recommended in component manufacturers' application specifications. The reason why the relationship seemed to hold was probably because, in the early years of microcircuit technology, quality control standards were not as high, and therefore a fairly large proportion of components were observed to fail at higher temperatures. However, current data do not show such a relationship, except for some specific failure modes which will be described later. This has major implications for thermal design, since the erroneous impression that 'the cooler the better' is widely held.

The true relationship between temperature and failure is as shown in Figure 9.2. Most electronic components can be applied at temperatures well in excess of the figures stated in databooks. For example, databook package temperature limits for industrial grade plastic encapsulated integrated circuits and transistors are typically 85 °C, and for ceramic or metal packaged devices 125 °C. These do not, however, relate to any physical limitations, but are based more on the conventions of the industry.

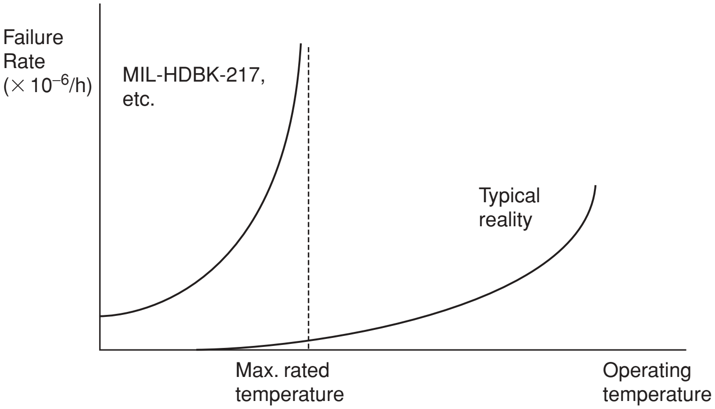


Figure 9.2 *Temperature vs. reliability for electronic components.*

Low temperatures can cause components to fail, usually due to parametric changes in electrical characteristics. Typical low temperature limits for most components are −20 °C to −60 °C. However, such failures are usually reversible, and correct function is regained if the temperature rises.

Repeated temperature changes can be more damaging than continuous operation at high temperatures. Temperature changes also cause fatigue damage and creep deformation of solder joints on surface-mounted electronic components, as described in Chapter 8.

9.2.1.4 Power

Electrical power generates heat ($W = I^2 R$). All 'active' electronic components, such as transistors, integrated circuits and amplifiers generate heat, and therefore increased temperature. So do components like coils, voltage dropping resistors, and so on. The steady-state component temperature will be the sum of the ambient temperature and the temperature generated by internal heating. The internal heat is dissipated by conduction through the device connections and mountings, and the thermal resistance between the active part of the device and the ultimate heat sink will determine the steady-state temperature. We discussed the effects of temperature on electronic device reliability earlier.

Some passive devices, such as resistors and capacitors, are also susceptible to failure due to power stress, if it causes overheating (rather than fusing due to excess current). Power stress over long periods can also cause drift of parameter values such as resistance or capacitance. Application instructions for such components generally include power stress limits.

Power stress cycling can lead to failure due to induced thermal cycling and thus fatigue, as described earlier.

9.3 Component Types and Failure Mechanisms

The main categories of electronic component types and their most common failure mechanisms are described in the sections below.

9.3.1 Integrated Circuits (ICs)

Integrated circuits (ICs) are not really 'components'. They are in fact sub-systems, containing transistors, capacitors and other discrete components within the IC structure. In the past ICs have been classified as follows: small scale integration (SSI): up to 100 logic gates; medium scale integration (MSI): up to 1000 gates; large scale integration (LSI): up to 10 000 gates; very large scale integration (VLSI): more than 10 000 gates, although those classifications are not used as often these days. Currently (2011) microprocessors can contain several billion (10^9) transistors.

Construction of ICs starts with the selective diffusion into the silicon wafer of areas of different charge level, either by ionic diffusion in a furnace or by implantation by a charged particle accelerator. The latter method is used for large-scale production. In the diffusion process, the areas for treatment are defined by masks. A sequence of treatments, with intermediate removal by chemical etch of selected areas, creates the structure of transistors and capacitors.

Different layers of the diffusion process are electrically isolated by layers of silicon dioxide (SiO_2) . This is called *passivation*. Finally, the entire die surface, apart from connector pads for the wire bonds, is protected with a further SiO_2 or Si_3N_4 layer. This process is called *glassivation*.

The connections between transistors on the chip and to the input and output pins are made via a network of conductor tracks, by depositing a thin layer of metal (*metallization*) on to the surface, through a mask. More recently, the number of interconnect levels has substantially increased due to the large number of transistors. Therefore, the timing delay in the wiring has become significant prompting a change in wiring material from aluminium to copper and from the silicon dioxides to materials with lower dielectric constant. Examples of these materials (called low-K dielectrics) include SiO_2 doped with fluorine or with carbon.

Finally, the assembly is packaged in either a plastic moulding or in a hermetic (ceramic or metal) package.

When ICs were first produced in the 1970s and 1980s they were mainly fairly simple analogue and digital circuits with defined functions (op-amps, adders, flip-flops, logic gates, etc.), and they were produced to closely defined generic specifications. For example, a 7408 is a 2-input AND gate, and it might have been manufactured by several suppliers. In parallel with the rapid growth in complexity and functionality in the years since, several different classes of IC have been developed. The classes that are available today include:

- ‘Standard’ ICs. These are the components that appear in generic specifications or in manufacturers’ catalogues. Examples are logic, memories, microprocessors, analogue to digital converters, signal processing devices, op-amps, and so on.
- Programmable logic devices (PLDs), field programmable gate arrays (FPGAs). These are standard circuits which can be ‘programmed’ by selective opening of fusible links.
- Mixed signal ICs. These have both digital and analogue circuits integrated into the same chip.
- Microwave monolithic ICs (MMICs). These are used in systems such as telecommunications, and contain microwave radio circuits.
- Complex multifunction devices (also referred as *system on a chip* (*SOC*)), which might contain mixed technologies, such as processing, memory, analogue /digital conversion, optical conversion, and so on, and a mixture of new and ‘legacy’ circuit designs.
- Micro-electro-mechanical systems (MEMS) is the technology of very small mechanical devices driven by electricity. MEMS are also referred to as micromachines in Japan, or Micro Systems Technology (MST) in Europe. Materials used for MEMS manufacturing include silicon, polymers and various metals, such as gold, nickel, aluminium, copper, and so on MEMS applications include sensors, actuators, medical devices and many others.

9.3.1.1 *Application-Specific IC*s

There is an increasing trend for ICs to be designed for specific applications. Standard ICs such as microprocessors and memories will always be used, but many circuits can be more economically implemented by using ICs which have been designed for the particular application. These are called *application-specific* ICs (ASICs).

In a *semi-custom* ASIC, all fabrication processes on the chip are previously completed, leaving arrays of transistors or cells to be interconnected by a conductor pattern designed for the particular application. In a full custom design, however, the chip is designed and manufactured entirely for the specific application.

Semi-custom ASICs are more economical than full custom ICs in relatively low quantities, but there is less flexibility of design and the utilization of chip area is less economical. Full custom ASICs are usually economical when large quantities are to be used, since the design and development costs are high. Both design approaches rely heavily on EDA, though the semi-custom method is easier to implement.

ASICs introduce important reliability aspects. The electronic system designer is no longer selecting 'black boxes', that is, standard ICs, from a catalogue, but is designing the system at the component (or functional group) level. Reliability (and testability) analysis must be performed to this level, not just to input and output pins. Since design changes are very expensive, it is necessary to ensure that the circuit is reliable and testable the first time. Particular aspects that need to be considered are:

1 Satisfactory operation under the range of operating inputs and outputs. It is not usually practicable to test a LSI or VLSI design exhaustively, due to the very large number of different operating states (analogous to the problem of testing software, see Chapter 10), but the design must be tested under the widest practicable range of conditions, particularly for critical functions.
2 The effects of failures on system functions. Different failure modes will have different effects, with different levels of criticality. For example, some failure modes might have no effect until certain specific conditions are encountered, whilst others might cause total and obvious loss of all functions.
3 The effects of system software on the total system operation. For example, the extent to which the software is designed to compensate for specified hardware failures, by providing failure indications, selecting alternate operating paths, and so on.
4 The need for and methods for providing built-in redundancy, both to increase production test yield and to improve reliability, particularly for critical applications.
5 Testability of the design. The ease with which circuits can be tested can greatly influence production costs and reliability, since untested functions present particular reliability hazards.

The EDA systems used for IC design include facilities for assessing reliability and testability. For example, failure modes can be simulated at the design stage and the effects evaluated, so stress analysis and FMECA can be integrated with the design process. Design analysis methods for electronic circuits are described in more detail later.

9.3.1.2 Microelectronics Packaging

There are two main methods of packaging IC chips. In hermetic packaging the die is attached to the package base, usually by soldering, wire bonds are connected between the die connector pads and the leadout conductors, and the package is then sealed with a lid. Packages are either ceramic or metal. Plastic encapsulated ICs (PEICs or PEDs) use an epoxy encapsulant.

PEICs are cheaper than hermetic ICs, and therefore tend to be used in domestic and much commercial and industrial equipment. However, PEICs are not usually recommended for high temperature operation (above 85 °C case). They can also suffer a life dependent (wearout) failure mode due to moisture ingress, either by absorption through the encapsulation material or along the plastic/metal boundary of the leads. The moisture provides a medium for electrolytic corrosion at the interfaces of conductor tracks and wire bonds, or of the conductor tracks themselves through any gaps or holes in the glassivation layer. No plastic encapsulant is totally impervious to moisture ingress, though modern materials and process controls have greatly reduced the problem. Therefore when PEICs are used in high temperature or moisture environments or where long life is important, for example, in military, automotive or aerospace applications, particular care should be taken to ensure their suitability.

For many years the most common package form was the dual-in-line package (DIL or DIP), with pin spacing of 0.1 inch (2.5 mm). The pins are inserted into holes in the printed circuit board (PCB) and soldered, or into a DIL socket which allows easy removal and reinsertion.

The packaging techniques used for the first 20 years or so of IC manufacture are giving way to new methods, primarily in order to enable more circuitry to be packaged in less volume. The leadless chip carrier (LCC) package and the small outline IC (SOIC) are *surface-mounted devices* (SMD). These have leadouts around the periphery, which are reflow soldered to the PCB conductor tracks (or to a ceramic substrate which is in turn soldered to the PCB conductor tracks) rather than being inserted through PCB holes as with DIP. (In reflow soldering the components are placed on the PCB or substrate, and the assembly is heated in an infra-red or vapour phase oven to melt the solder.) The leadouts are on a 0.05 inch (1.25 mm) spacing or less.

The new packaging techniques have also been developed with automation of the assembly processes in mind. The components, including the very small 'chip' packaged discrete components such as transistors, diodes, capacitors, and so on, are too small, and the solder connections too fine, to be assembled manually, and automatic placement and soldering systems are used.

More recent developments are the pin grid array (PGA) and ball grid array (BGA) packages. Leadouts are taken to an array of pins on the underside of the package, on a 0.05 inch (1.25 mm) (or less) grid. As with the LCC package, connection is made to the PCB or substrate by reflow soldering. Other packaging methods include direct mounting of the chip onto the PCB or substrate such as flip chip (solder bumps interconnects down) and chip-on-board (interconnects up), and chip scale packaging (CSP). Figure 9.3 shows examples of some of the packages mentioned above.

Also due to continued miniaturization of electronic devices IC manufacturers have been actively exploring the third or Z-height dimension in electronic packaging. Stacked die or 3-D packaging is becoming more and more common in CSPs, BGAs and other types of high density ICs. Figure 9.4 shows that stacked IC packages require large numbers of interconnects, are more difficult to manufacture and are potentially less reliable than 'traditional' IC packages.

The main reliability implication of the new IC packaging technologies is the fact that, as the volume per function is decreased, the power dissipation per unit volume increases. This can lead to difficult thermal

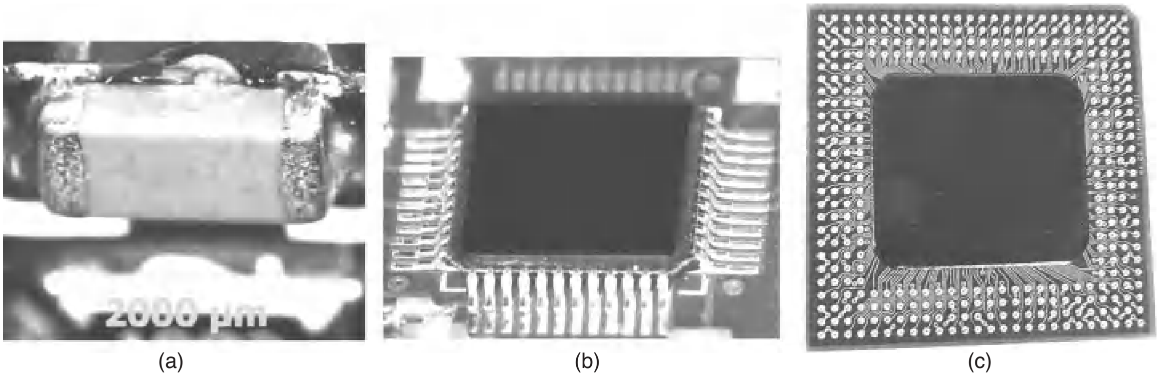


Figure 9.3 *Examples of electronic components. (a) Leadless chip capacitor (b) Quad flat pack IC package (QFP) (courtesy DfR Solutions) (c) Ball grid array (BGA) IC package.*

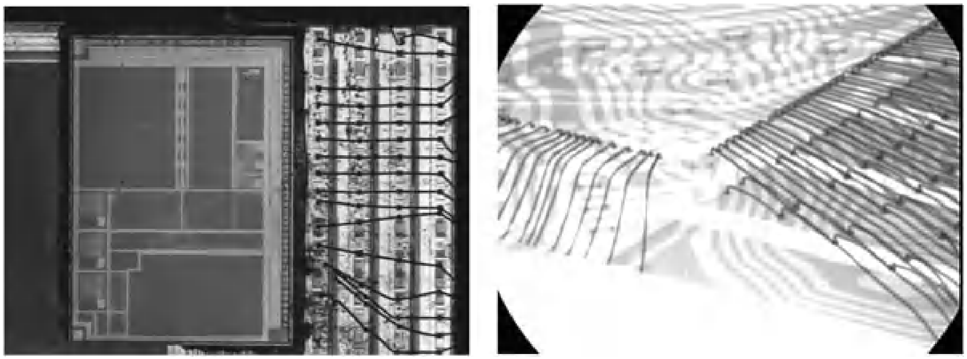


Figure 9.4 Five stacked die 4 GB flash memory (pyramid stacking with wire bond interconnects). Reproduced by permission of DfR Solutions.

management problems in order to prevent junction temperatures attaining levels above which reliability would be seriously affected. Liquid cooling of assemblies is now necessary in some applications such as in some military and high speed computing and test equipment.

Another reliability aspect of the new methods is the use of reflow soldering to the surface of the PCB or substrate. The large numbers of solder connections on the undersides of the packages cannot be inspected, except with X-rays. Also, repeated thermal cycling can cause the solder joints to fail in shear. Therefore the solder process must be very carefully controlled, and subsequent burn-in and reliability tests (see later chapters) must be designed to ensure that good joints are not damaged by the test conditions.

9.3.1.3 Hybrid /Microelectronic Packaging/Multichip Modules

Hybrid microelectronic packaging is a technique for mounting unencapsulated semiconductor and other devices on a ceramic substrate. Resistors are made by screen-printing with conductive ink and laser-trimming to obtain the desired values. Connections from the conducting tracks to the device pads are made using fine gold, copper or aluminium wire and ultrasonic bonding in the same way as within an encapsulated IC similar that shown in Figure 9.4. The complete assembly is then encased in a hermetic package (Figure 9.5).

Hybrid packaging provides certain advantages over conventional printed circuit board construction, for special applications. It is very rugged, since the complete circuit is encapsulated, and it allows higher density

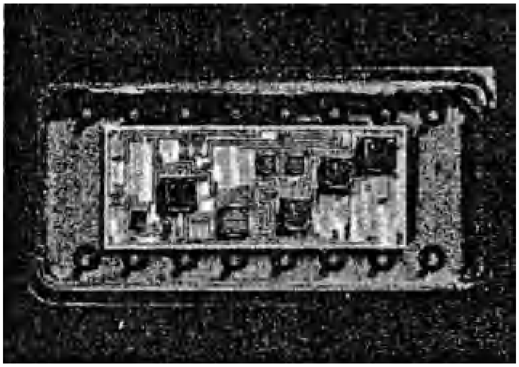


Figure 9.5 Micro-hybrid (Courtesy National Semiconductor Corporation).

packaging than PCB mounting of components. However, it is not practicable to repair hybrid circuits, since clean room conditions and special equipment are necessary. Therefore they are suited for systems where repair is not envisaged except by replacing hybrid modules and discarding defective ones. Hybrid circuits are used in missile electronics, automotive engine controls and severe environment industrial conditions, and in many other applications where compact assembly is required, such as for high frequencies. Hybrid circuits can be custom designed and manufactured or standard catalogue hybrids can be used.

Due to their relatively large size, the number of internal bonds and the long package perimeter, hybrids tend to suffer from inclusion of contamination and conducting particles, bond failure and sealing problems more than do packaged ICs on equivalent PCB circuits. Therefore, very stringent production and quality control and required if the potential reliability of hybrid circuits is to be realized. MIL-STD-883 includes the same screening techniques for hybrid microelectronics as for discrete ICs, and the European and UK specifications (CECC, BS 9450) include similar requirements, as described later.

Multichip module (MCM) packaging is a more recent development of the microhybrid approach.

See Harper (2004), Tummala (2001), Lau *et al*. (1998) and Tummala *et al*. (1997) for detailed descriptions of packaging technologies and their reliability aspects.

9.3.1.4 Microelectronic Component Attachment

Microelectronic components in DIP and LCC packages can either be soldered on the PCBs or plugged into IC sockets which are soldered in place. Plugging ICs into sockets provides three major advantages from the test and maintenance points of view:

1 Failed components can easily be replaced, with less danger of damaging the PCB or other components.
2 Testing and diagnosis is usually made much easier and more effective if complex devices such as microprocessors are not in place.
3 It is much easier to change components which are subject to upgrades or modifications, such as memories and ASICs.

On the other hand, there are some drawbacks which can override these advantages in certain circumstances. These are:

1 Heat transfer is degraded, so it might not be possible to derate junction temperatures adequately.
2 There might be electrical contact problems in high vibration, shock or contamination environments.
3 There is a risk of damage to the IC and the socket due to handling.

IC sockets are therefore used on some repairable systems, for example memory devices are often socket mounted to allow access for replacement.

9.3.1.5 Microelectronic Device Failure Modes

The main failure modes of ICs are:

– *Electrical overstress/electrostatic damage* (EOS/ESD). ICs are susceptible to damage from high voltage levels, which can be caused by transient events such as switching or electrostatic discharge from people or equipment. Most integrated circuits contain built-in EOS/ESD protection circuits, which will typically protect them against short-duration overstress conditions (typically up to 1000 V and 500 μJ).

- *Latchup* is the creation of a low resistance path between the power input and ground in an IC. CMOS ICs are prone to this failure mechanism when subjected to transient voltage overstress caused by ESD or other transient pulses from circuit operation, test equipment, and so on. The effect is permanent complete failure of the device.
- *Electromigration* (EM) is a failure mechanism that is becoming increasingly important as the metallic conductors (often referred as interconnects) inside ICs are made to extremely narrow dimensions (currently of the order of 35 nanometres and continually decreasing). Such cross-sectional areas mean that the current density, even at the very low current and voltage levels within such circuits, can be very high. EM is the bulk movement of conductor material, at the level of individual metal crystals, due to momentum interchange with the current-carrying electrons. This can result in local narrowing of the conductor track, and thus increased local current density and eventual fusing. Also, the displaced material can form conducting whiskers, which can cause a short circuit to a neighbouring track. The EM process can be quantified using *Black's Law* (9.1):

$$t_{\mathrm{EM}} = A(W)J^{-N} \exp[E_{\mathrm{A}}/kT] \tag{9.1}$$

where: t_{EM} = time to failure due to EM process.
J = current density (A/m^2).
N = empirical constant, between 1 and 3.
$A(W)$ = material constant, a function of line width.
E_{A} = activation energy (see Arrhenius law, Chapter 8).
k = Boltzmann's constant (8.6173 × 10-5 eV/K).

EM is an important failure mode in electronic systems which must operate for long times, particularly if operating temperatures are high, such as in engine controls, spacecraft, and telecommunications systems (repeaters, switches, etc.). Electromigration is becoming a more serious problem as IC miniaturization continues.

- *Time-dependent dielectric breakdown* (TDDB) is a failure mode of the capacitors within ICs caused by whiskers of conductive material growing through the dielectric (silicon dioxide), and eventually short-circuiting the device. The effect is accelerated by voltage stress and by temperature and therefore becomes worse as electronic devices decrease in size.
- *Slow trapping* is the retention of electrons in the interstitial boundaries between Si and SiO_2 layers in ICs. These cause incorrect switching levels in digital logic and memory applications. Susceptibility to slow trapping is primarily dependent on device manufacturing processes. Again, continued decrease in size of electronic devices and consequent increase in electric fields causes more charge trapping in ICs.
- *Hot carriers* are electrons (or holes) that have sufficient energy to overcome the energy barrier of the Si–Si and SiO_2 boundary, and become injected into the SiO_2. It occurs in sub-micron ICs in which the electric field strengths can be sufficiently high. The effects are to increase switching times in digital devices and to degrade the characteristics of analogue devices. Hot carrier effects can be reduced by process design techniques and by circuit design, both to reduce voltage stress at critical locations.
- *Soft errors* are the incorrect switching of a memory cell caused by the passage of cosmic ray particles or alpha particles. Cosmic rays create such effects in circuits in terrestrial as well as space applications. Alpha particles are generated by trace heavy metal impurities in device packaging materials. The errors can be corrected by refreshing the memory.

- Processing problems in manufacture (diffusion, metallization, wire bonding, packaging, testing, etc.) can cause a variety of other failure mechanisms. Most will result in performance degradation (timing, data loss, etc.) or complete failure.

Ohring (1998), Bajenescu and Bazu (1999), and Amerasekera and Najm (1997) describe microelectronic component reliability physics.

It is important to appreciate that, despite the many ways by which microelectronic devices can fail, and their great complexity, modern manufacturing processes provide very high quality levels, with defective proportions being typically 0–100 per million. Also, appropriate care in system design, manufacture and use can ensure adequate protection against externally-induced failures. As a result, only a small proportion of modern electronic system failures are due to failures of microelectronic devices.

9.3.1.6 Microelectronic Device Specifications

In order to control the quality and reliability of microelectronic devices for military purposes, US Military Specification M-38510 was developed. This describes general controls, and separate sections ('slash sheets') give detailed specifications of particular device types. Similar international (International Electrotechnical Commission – IEC), European (CECC) and British (British Standards Institution – BS 9400) specifications have since been generated. These specifications are generally 'harmonized', so that there is little if any difference between them for a particular device type. Components produced to these specifications are referred to as 'approved' components. Due to the rapid evolution of electronic components, M-38510 is not frequently used these days.

Military system specifications usually require that electronic components are manufactured and tested to these standard specifications, in order to provide assurance of reliability and interchangeability. However, with the rapid growth in variety of device types, the specification systems have not kept pace, so that many of the latest device types available on the market do not have such specifications. In order to cope with this problem, an approach called *capability approval or qualified manufacturers list* (QML) provides generic approval for a device manufacturer's processes, covering all similar devices from that line. US MIL-STD-PRF 38535C describes the system for US military application. Capability approval is also appropriate for ASIC manufacture.

The general improvements in process quality have also resulted in removing the quality gap between 'approved' and industrial and commercial-grade components, so the justification for the specification systems are not always applied, and more flexibility of application is allowed, depending on factors such as application and cost. Most manufacturers of high-reliability non-military electronic systems use commercial grade components, relying on the manufacturers' specifications and quality control.

9.3.1.7 Microelectronic Device Screening

Screening is the name given to the process of finding by test which of a batch of components or assemblies is defective, without weakening or causing failure of good items. It is justified when:

1 The expected proportion defective is sufficiently high that early removal will improve yield in later tests and reliability in service.
2 The cost of screening is lower than the consequential costs of not screening.

The assumption that no weakening of good items will occur implies that the hazard rate will be decreasing.

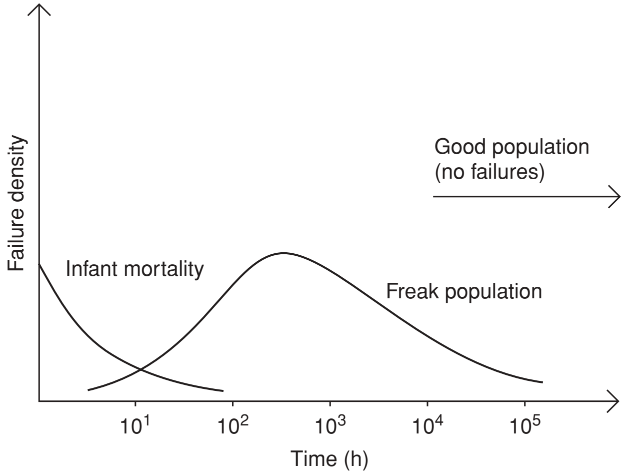


Figure 9.6 *Typical failure density functions of electronic components when no component burn-in has been carried out.*

Figure 9.6 shows the three categories of component that can be manufactured in a typical process. Most are ‘good’, and are produced to specification. These should not fail during the life of the equipment. Some are initially defective and fail when first tested, and are removed. They therefore do not cause equipment failures. However, a proportion might be defective, but nevertheless pass the tests. The defects will be potential causes of failure at some future time. Typical defects of this type are weak wire bond connections, silicon, oxide and conductor imperfections, impurities, inclusions and non-hermetic packages. These components are called *freaks*.

Screening techniques have been developed specifically for microcircuit devices. The original standard for these is US MIL-STD-883G: *Test Methods and Procedures for Microelectronic Devices*. The other national and international standards mentioned above include very similar methods. There are three basic screen levels, as summarized in Table 9.1. The ‘A’ level screen (also referred to as ‘S’, for spacecraft application) is the most severe, and the most expensive. ‘B’ level screening is typically applied to microcircuits to be used in military, avionic and other severe-environment, high-integrity systems, particularly if a long operating life is required. ‘C’ level is a more relaxed specification, which does not include burn-in, as described below.

Burn-in is a test in which the components are subjected to high temperature operation for a long period, to stimulate failure of defective components by accelerating the stresses that will cause failure due to these defects, without damaging the good ones. In MIL-STD-883 the temperature to be used is 125 °C (package temperature), for 168 hours duration. The electrical test conditions are also specified.

Component manufacturers and users have developed variations of the standard screens and burn-in methods. The main changes are in relation to the burn-in duration, since 168 hours has been shown to be longer than necessary to remove the great majority of defectives (the only justification for 168 hours is merely that it is the number of hours in a week). Also, more intensive electrical tests are sometimes applied, beyond the simple reverse-bias static tests specified. Dynamic tests, in which gates and conductors are exercised, and full functional tests with monitoring, are applied to memory devices, other VLSI devices, and ASICs, when the level of maturity of the process or design and the criticality of the application justify the additional costs.

Table 9.1 *Microelectronic device screening requirements*[a].

		Screen level applicability		
		MIL-STD-883/BS		9400
Screen	Defects effective against	A	B	C
Pre-encapsulation visual inspection (30–200 × magnification)	Contamination, chip surface defects, wire bond positioning	100 %	100 %	100 %
Stabilization bake	Bulk silicon defects, metallization defects (stabilizes electrical parameters)	100 %	100 %	100 %
Temperature cycling	Package seal defects, weak bonds, cracked substrate	100 %	100 %	100 %
Constant acceleration (20 000 *g*)	Chip adhesion, weak bonds, cracked substrate	100 %	100 %	100 %
Leak tests	Package seal	100 %	100 %	100 %
Electrical parameter tests (pre-burn-in)	Surface and metallization defects, bond failure, contamination/ particles	100 %	100 %	100 %
Burn-in test (168 h, 125 °C with applied a.c. voltage stress)	Surface and metallization defects	100 %	100 %	—
	Weak bonds	240 h	168 h	
Electrical test (post-burn-in)	Parameter drift	100 %	100 %	—
X-ray	Particles, wire bond position	100 %	—	—

[a] This table is not comprehensive, and the reader should refer to the appropriate standard to obtain full details of applicable tests.

Most plastic encapsulated components cannot be burnt-in at 125 °C, so lower temperatures are used. Also, in place of leak tests they are tested for moisture resistance, typically for 1000 hours in an 85 °C/85 % relative humidity (RH) chamber. This is not, however a 100 % screen, but a sample test to qualify the batch.

A more severe test, using a non-saturating autoclave at 100 °C and 100 % RH, is also used, as the 85 °C/85 % RH test is not severe enough for the latest encapsulating processes.

The recent trends in microelectronic device quality have to a large extent eliminated the justification for burn-in by component users. Most component manufacturers burn-in components as part of their production processes, particularly for VLSI components and ASICs, using variations of the standard methods. Also, the new packaging technologies are not suitable for handling other than by automatic component placement machines, so user burn-in is generally inadvisable, as the handling involved can lead to damage and can degrade the solderability of the contacts.

Burn-in methods are described in Kuo *et al*. (1998).

9.3.2 Other Electronic Components

Other electronic component types are primarily ‘active’ devices, such as transistors and diodes, and ‘passive’ devices such as resistors, capacitors, inductors, PCBs, connectors, and so on. In general these discrete components are very reliable and most have no inherent degradation mechanisms (exceptions are light-emitting diodes, relays, some vacuum components, electrolytic capacitors, etc.). Factors that can affect reliability include thermal and electrical stress and quality control of manufacture and assembly processes. Standard specifications exist, as for microcircuits, but screening is not normally applied, apart from the manufacturers' functional tests.

Application guidelines for reliability are given in Ohring (1998) and US MIL-HDBK-338 and in component manufacturers' databooks. The main component types are discussed in more detail below.

9.3.2.1 Discrete Semiconductors

– Processing problems in manufacture (diffusion, surface condition, metallization, wire bonding, packaging, testing, etc.) can cause a variety of failure mechanisms. Most will result in performance degradation or complete failure.
– For power devices the uniformity and integrity of the thermal bond between the chip and the package is an important feature to ensure reliability at high power and thermal stress.

9.3.2.2 'Passive' Components

– Resistors, capacitors, inductors and other components can fail due to fabrication problems, damage on assembly into circuits, ESD and other causes. These usually cause the component to become open circuit or high resistance.
– Component parameter values can be out of tolerance initially, or parameter values can drift over time due to applied stresses, as described earlier.
– Components can be electrically 'noisy', due to intermittent internal contacts or impurities in materials.

9.3.2.3 Capacitors

– High voltage overstress generally causes the capacitor to become open circuit. High voltage/high power capacitors might even explode if they are short-circuited.
– Low voltage, low power capacitors, such as those built into integrated circuits as memory storage devices, can suffer long-term failures due to the mechanism of dendritic metal whisker growth through the dielectric.
– Capacitors that use a liquid or paste dielectric (electrolytic capacitors) degrade over time if no voltage stress is applied, and then fail short-circuit when used. They must be 're-formed' at intervals to prevent this. Capacitors kept in storage, or units such as power supplies which contain such components and which are stored or kept idle for long periods, must be appropriately maintained and checked.
– Electrolytic capacitors are damaged by reverse or alternating voltage, and so must be correctly connected and if necessary protected by diodes. Miniature tantalum capacitors are also degraded by ripple on the applied voltage, so they should not be used with unsmoothed voltage levels.
– Multilayer ceramic chip capacitors (MLCC) are usually small in size and large in capacitance. They are designed for surface mount applications and usually have a long life due to complete sealing of inner electrodes.

9.3.2.4 Electro-Optical Components

Many modern systems use optical fibres, connectors and electro-optical (EO) components for data transmission. Optical frequencies permit very high data rates. A major reliability advantage of EO systems is that they do not create EMI and they are immune to it. EO components are also used to provide over-voltage protection on data lines, by converting electrical signals to optical signals and back again to electrical. Failure modes of EO components are:

- Breakage of optical fibres.
- Misalignment of optical fibres at connections within connectors and to components. The connecting ends must be accurately cut across the length and aligned with one another.
- Degradation of light output from light-emitting diodes (LEDs).

9.3.2.5 Cables and Connectors

Electrical power and signals must be conducted within and between circuits. Cables and connectors are not usually perceived as high-technology or high-risk components, but they can be major contributors to unreliability of many systems if they are not carefully selected and applied.

The most common cable systems are copper wires in individual or multiconductor cables. Multiconductor cables can be round or flat (ribbon). Cable failure mechanisms include damage during manufacture, use or maintenance, and fatigue due to vibration or movement. Failures occur mainly at terminations such as connections to terminals and connectors, but also at points where damage is applied, such as by repeated bending around hinges. The failure modes are either permanent or intermittent open circuit. Cable runs should be carefully supported to restrict movement and to provide protection. Testing should address aspects such as the possibility of damage during assembly and maintenance, chafing of insulation due to vibration, fatigue due to vibration or other movement, and so on.

The main types of connectors are circular multipin, for connecting round cables, and flat connectors for connecting ribbon cables and circuit boards. Individual wires are often connected by soldering or by using screw-down or push-on terminals. Low-cost connectors may not be sufficiently robust for severe environments (vibration, moisture, frequent disconnection/reconnection, etc.) or long-life applications. Connectors for important or critical applications are designed to be rugged, to protect the connector pin surfaces from moisture and contamination, and the connecting surfaces are gold plated.

In many modern electrical and electronic systems connectors contribute a high proportion, and often the majority of failures in service. The most common failure modes are permanent or intermittent open circuit due to damage or buildup of insulation on the mating surfaces due to oxidation, contamination or corrosion. Therefore it is important that they are carefully selected for the application, protected from vibration, abuse, corrosion and other stresses, and that their failure modes are taken into account in the test programme.

Data signals are also transmitted as pulses of light through optical fibre conductors and connectors. Light-emitting diodes operating in the infra-red part of the spectrum act as transmitting and receiving devices at the ends of the fibres. Special optical connectors are used to connect the ends of fibres, which must be accurately aligned and mated to ensure transmission. Optical fibres can break as a result of bending stresses.

9.3.2.6 Insulation

Insulation is as important in electrical and electronic systems as is conduction. All conductors, in cables and connectors and on circuit cards, must be insulated from one another. Windings in coils (solenoids, motors, generators, etc.) also require insulation. Insulation also provides protection against injury or death from human contact with high voltages.

Insulators can degrade and fail due to the following main causes:

- Mechanical damage, being trapped, chafed, cut, and so on.
- Excessive temperature, causing charring and hence loss of dielectric strength. We are all familiar with the smell of burning shellac when an electrical appliance such as a drill or a microwave oven suffers a short circuit in a motor or transformer coil.

- Embrittlement and then fracture, caused by exposure to high temperatures, UV radiation, or chemical contamination. (Oil contamination can cause some cable insulator materials to swell and soften.)
- Rodent attack. Some insulator materials are enjoyed by mice and rats which take care of agricultural machinery during the winter.

Degradation of insulation is nearly always a long-term phenomenon (10 or more years typically).

9.3.3 Solder

Solder and soldering process are critical to the reliability of electronic systems. Large numbers of failures are attributed to solder joint fatigue and cracking, especially in harsh environment industries, such as automotive, avionics and military.

9.3.3.1 Tin-Lead Solder

Tin–lead solder (typically Sn63Pb37, melting point 183 °C) has for many years been by far the most common material used for attaching electronic components to circuit boards. Solder joints can be made manually on relatively simple circuits which do not utilize components with fine pitches (less than 2.5 mm) between connections. However, for the vast majority of modern electronic systems solder connections are made automatically. The main techniques used are:

- *Through-hole mounting and manual or wave solder*. Components are mounted onto circuit boards, with their connections inserted through holes. DIP ICs are mounted this way, as are many types of discrete components. The hole spacing is typically 2.5 mm. In wave soldering, the boards are passed over a standing wave in a bath of liquid solder, so that each connection on the underside is immersed briefly in the wave. These methods are now used mainly for components and circuits which do not need to utilize the most compact packaging technologies, such as power circuits.
- *Surface mount and infra-red or vapour phase solder* (sometimes referred as reflow). The circuit boards are printed with solder paste at the component connection positions, and the SMT components are mounted on the surface by automatic placement machines. BGA solder connections are made by solder balls (typically 1 mm diameter) being accurately positioned on the solder paste on the board, then the BGA package is positioned on top. The ‘loaded’ boards are passed through ‘reflow’ ovens that melt the solder for just long enough to wet the solder so that intermetallics are formed, before the solder solidifies. The ovens are heated either with infra-red radiation, or, in the more common vapour phase or convection ovens, with gas heated to above the solder melting point. In the latter, the latent heat of condensation of the gas is transferred to the solder, and heating is very even.
- Laser soldering is also used to a limited extent, but further developments are likely.

The different types of component and solder methods are sometimes used in combination on circuit boards.

A reliable solder joint must provide good mechanical and electrical connection. This is created by the formation of intermetallic alloys at the interfaces between the solder and the surfaces being joined. The most common reasons for solder joint failure are:

- Inadequate solder wetting of the surfaces to be joined, due to surface contamination or oxidation. Components must be carefully stored and protected before placement. Components should not be stored for long periods before assembly, and fine pitch components should be handled only by placement machines.

- Insufficient heat (time or temperature) applied. The solder might be melted sufficiently to bond to one or both surfaces, but not enough to form intermetallics. Such joints will conduct, but will be mechanically weak.
- Fatigue due to thermally induced cyclic stress or vibration, as described in Chapter 8.
- Creep due to thermally induced cyclic stress, as described in Chapter 8.

All of these can lead to operating failures that show up on production test or in service. Failures can be permanent or intermittent. Modern electronic circuits can contain tens of thousands of solder connections, all of which must be correctly made. Control of the solder processes is a major factor in ensuring quality and reliability, and inspection and test of joint quality is an important feature of modern production test systems.

Pecht (1993) and Brindley and Judd (1999) describe soldering methods and problems.

9.3.3.2 Lead-Free Solder

The introduction and implementation of the Restriction of Hazardous Substances Directive (RoHS) in Europe has seriously impacted the electronics industry as a whole. This directive restricts the use of several hazardous materials in electronic equipment; most notably, it forces manufacturers to remove lead from the soldering process. This transition has seriously affected manufacturing processes, validation procedures, failure mechanisms and many engineering practices associated with lead-free electronics.

The electronics industry continues experimenting with various lead-free alloys of tin with silver, copper, bismuth, indium and traces of other metals. Currently (2011) the most commonly used lead-free solder alloy consists of: Sn-96.5 %, Ag-3.0 %, Cu-0.5 % and is often referred as SAC305. The popularity of this alloy is partially based on the reduced melting point of 217 °C, which is higher for other alloys. In general, lead-free solder is a less compliant material than tin-lead, which causes concerns with pad cratering and cracking due to vibration. Pad cratering is a phenomenon where due to lower ductility a solder joint creates higher pulling force during the cooling process separating pads from a circuit board.

A large amount of literature has been published on reliability of lead-free solder including CALCE centre at the University of Maryland (CALCE, 2011), DfR Solutions (DfR, 2010), J.-P. Clech of EPSI Inc. (Clech *et al*., 2009), works by W. Engelmaier (Evans, 2010) and many others.

The thermal fatigue properties of tin-lead and lead free solder are different and depending on component types, geometry and soldering process one can have higher reliability than the other. Also, compared to the thermal cycling acceleration factors, the lead-free solder fatigue models are more complicated and more strongly influenced by parameters like maximum and minimum cycling temperature and dwell times (see more in Chapter 13).

Reducing lead in electronics caused another problem in terms of reliability – tin whiskers. Pure or almost pure tin tends to grow whiskers – crystalline filaments, which have not been observed in tin-lead solder. Tin whiskers can grow up to several millimetres long and bridge adjacent terminals causing a short. They can also break loose and cause a short some other place in the circuit or impede the movement of mechanical parts.

Tin whiskers are more likely to grow on tin plated terminals and mechanical stress, temperature and humidity contribute to their growth. Also, whiskers can have a long dormancy period (up to 3000 hours) making it difficult to create an effective burn-in process.

Additional reliability concerns for lead-free solder include formation of *Kirkendall voids* at the interfaces of tin and copper. Kirkendall voids can be caused by thermal ageing especially at elevated temperatures. The formation of a string of these voids can produce a perforated tear line that represents a significant weakness relative to mechanical shock.

As the electronics industry transition from tin-lead to lead-free solder continues, reliability is expected to remain one of the major concerns.

9.4 Summary of Device Failure Modes

For electronic devices used in a system, the most likely failure modes must be considered during the design so that their effects can be minimized. Circuit FMECAs and system reliability block diagrams should also take account of likely failure modes. Table 9.2 summarizes the main failure modes for the most common device types. The failure modes listed are not exhaustive and the device types listed are only a summary of the range. The failure mode proportions can vary considerably depending on type within the generic headings,

Table 9.2 *Device failure modes.*

Type	Main failure modes	Typical approximate proportions (per cent)
Microcircuits		
Digital logic	Output stuck at high or low	80
	No function	20
Linear	Parameter drift	20
	No output	70
	Hard over output	10
Transistors	Low gain	20
	Open-circuit	30
	Short-circuit	20
	High leakage collector-base	30
Diodes		
Rectifier, general purpose	Short-circuit	10
	Open-circuit	20
	High reverse current	70
Resistors		
Film, fixed	Open-circuit	30
	Parameter change	70
Composition, fixed	Open-circuit	10
	Parameter change	90
Variables	Open-circuit	30
	Intermittent	10
	Noisy	10
	Parameter change	50
Relays		
	No transfer	20
	Intermittent	70
	Short-circuit	10
Capacitors		
Fixed	Short-circuit	60
	Open-circuit	20
	Excessive leakage	10
	Parameter change	10
Solder, connectors	Open circuit	50
	Short circuit	20
	Intermittent	30

application, rating and source. Devices used within a particular design should be individually assessed, for example, a resistor rated very conservatively is likely to have a reduced relative chance of failing open-circuit.

Circuit design should take account of the likely failure modes whenever practicable. Capacitors in series will provide protection against failure of one causing a short-circuit and resistors in parallel will provide redundancy against one failing open. Blocking diodes are often arranged in series to protect against shorts.

Bajenescu and Bazu (1999) and Ohring (1998) describe reliability aspects of most types of electronic component.

9.5 Circuit and System Aspects

9.5.1 Distortion and Jitter

Distortion is any change in the shape of a waveform from the ideal. Distortion can be caused by several factors, including mismatched input and output impedances, crossover distortion in transistors, optical devices and op-amps, transistor saturation, interference (see below), thermal effects, and so on. All waveforms (power, audio, HF to microwave signals, digital signals, etc.) can be affected, and the problems grow as frequencies increase. Circuit designs should minimize distortion, but it can cause failures, or it can be the symptom of component failures or parameter variations.

Jitter is a form of distortion that results in an intermittent variation of a waveform from its ideal position, such as timing, period or phase instabilities. It can affect the operation of high-speed circuits, particularly the timing of digital pulses, and can therefore cause corruption of signals and data.

9.5.2 Timing

Timing is an important aspect of most digital electronic circuit design. To function correctly, the input and output voltage pulses to and from circuit elements must appear at the correct times in the logic sequences. This is relatively easy to arrange by design for simple circuits which operate at relatively low speed (clock rate or frequency). However, as speeds and complexity have increased, and continue to do so, it becomes increasingly difficult to ensure that every pulse occurs at the correct time and sequence. The integrity of the pulse waveform also becomes more difficult to assure at higher frequencies. Any digital circuit will have a speed above which it will begin to perform incorrectly. At higher assembly levels, such as telecommunications or control systems, further limitations on speed of operation can be caused by inductive and capacitive effects within the circuit and by propagation delays along conductors.

9.5.3 Electromagnetic Interference and Compatibility

Electromagnetic interference (EMI) is the disturbance of correct circuit operation caused by changing electromagnetic fields or other electrical stimuli, which are then received by signal lines and so generate spurious signals. EMI is also called 'noise'? Electromagnetic compatibility (EMC) is the ability of circuits and systems to withstand these effects. EMC is sometimes referred to as electromagnetic immunity.

EMI can be generated by many sources, such as:

- High frequency radiation within the system, from switching within components, transmissions from data lines, and oscillators. Every component and conductor is a potential transmitter and a potential receiver. This is called 'cross-coupling'.

- Transient differences between voltage potentials on different parts of the circuit ground plane, due to inductive or capacitive effects.
- Electromagnetic emissions from RF components or subsystems when these are part of the system, or from other systems, such as radios, radars, engine ignition systems, electric motors, arcing across relays, lightning strikes, and so on.
- Switching of inductive or capacitive loads such as motors on the same power circuit.
- The operating frequencies of modern digital systems are in the radio frequency range (500 MHz to over 3 GHz), and harmonics are also generated. The design objective is to ensure that all signals travel through the circuit conductors, but at such high frequencies there are inevitably radiated emissions, which can then be received by other conductors and components. Circuit design to prevent EMI is a difficult and challenging aspect of all modern system designs. Methods are described later.

9.5.4 Intermittent Failures

A large proportion of the failures of modern electronic systems are in fact of an intermittent nature. That is, the system performs incorrectly only under certain conditions, but not others. Such failures are most often caused by connectors that fail to connect at some time such as under vibration or at certain temperatures, broken circuit card tracks that are intermittently open circuit, tolerance buildup effects between component parameters, and so on. It is not uncommon for more than 50 % of reported failures of systems to be diagnosed on investigation as ‘*no fault found*’ (NFF) or ‘*retest OK*’ (RTOK), mainly due to the effects of intermittent failures. Worse, since the causes of the failures are mostly not detected, the faulty units are not repaired, and can cause the same system failure when reinstalled. These can therefore generate high costs of system downtime, repair work, provision of spare units, and so on.

9.5.5 Other Failure Causes

There are many other causes of failure of electrical/electronic components and systems. It is impracticable to attempt to try to provide a comprehensive list, but examples include:

- Failure of vacuum devices (CRTs, light bulbs and tubes, etc.) due to seal failures.
- Mechanical damage caused by assembly operations or maintenance.
- Failures due to non-operating environments, such as storage or standby conditions. Pecht and Pecht (1995) covers these aspects.

9.6 Reliability in Electronic System Design

9.6.1 Introduction

The designer of an electronic system must consider the following main aspects in order to create an inherently reliable design:

1 Electrical and other stresses, particularly thermal, on components, to ensure that no component can be overstressed during operation or testing.
2 Variation and tolerances of component parameter values, to ensure that circuits will function correctly within the range of likely parameter values.

3 The effects of non-stress factors, such as electrical interference, timing and parasitic parameters. These are particularly important in high frequency and high gain circuits.
4 Ease of manufacture and maintenance, including design for test.

In addition to these primary considerations, there are other aspects of circuit and system design which can be applied to improve reliability. By reducing the number of different part types, the parts selection effort can be reduced and designs become easier to check. This also generates cost savings in production and use. Redundancy can also be designed into circuits. Whenever practicable the need for adjustments or fine tolerances should be avoided.

Not all of the means of achieving reliable electronic design are complementary. For example, redundancy and the inclusion of additional protective devices or circuits are not compatible with reducing complexity and the number of part types. The various design options relevant to reliability must be considered in relation to their effectiveness, cost and the consequences of failure.

The sections which follow outline the most important methods available to ensure high reliability. They are by no means comprehensive: circuit designers should consult their own organizations' design rules, component application notes, the Bibliography to this chapter and other relevant sources. However, what follows is intended as a guide to the points which reliability engineers and circuit designers need to consider.

9.6.2 Transient Voltage Protection

Electronic components are prone to damage by short duration high voltage transients, caused by switching of loads, capacitive or inductive effects, electrostatic discharge (ESD), incorrect testing, and so on. Small semiconductor components such as ICs and low power transistors are particularly vulnerable, owing to their very low thermal inertias. MOS devices are very vulnerable to ESD, and require special protection, both externally and on-chip.

Logic devices which interface with inductive or capacitive loads, or which 'see' test connections, require transient voltage protection. This can be provided by: a *capacitor* between the voltage line to be protected and ground, to absorb high frequency transients (buffering), *diode protection*, to prevent voltages from rising beyond a fixed value (clamping), and *series resistances*, to limit current values. Figure 9.7 and Figure 9.8 show typical arrangements for the protection of a logic device and a transistor. IC protection is also provided by transmitting logic signals via a light-emitting diode (LED) and optical transducer combination, called an opto-isolator or opto-coupler.

The transient voltage levels which can cause failure of semiconductor devices are referred to as VZAP. VZAP values depend upon transient duration. Maximum safe transient voltages are stated in manufacturers' databooks, and standard tests have been developed, for example in MIL-STD-883.

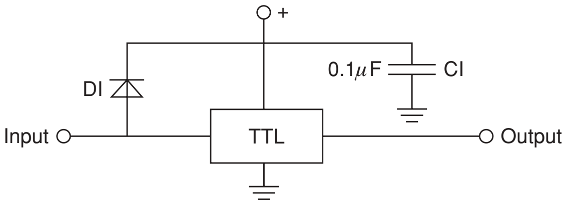


Figure 9.7 *Logic device protection. Diode D1 prevents the input voltage from rising above the power supply voltage. Capacitor C1 absorbs high frequency power supply transients. Reproduced by permission of Reliability Analysis Center.*

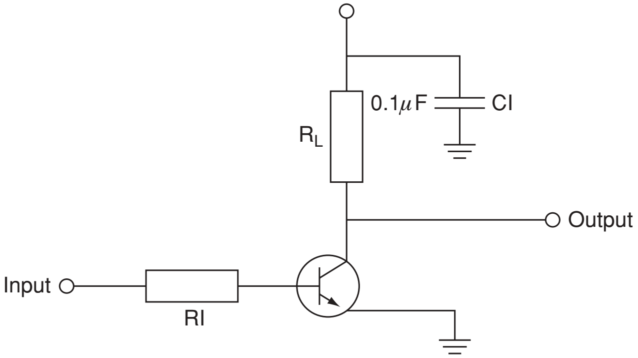


Figure 9.8 Transistor protection. Resistor R1 limits the base current I_B and capacitor C1 absorbs power supply high frequency transients. Reproduced by permission of Reliability Analysis Center.

Passive devices can also be damaged by transient voltages, but the energy levels required are much higher than for small semiconductor devices. Therefore passive devices do not normally need individual protection.

Very high electrostatic potentials, up to 5000 V, can be generated by triboelectric effects on clothing, packaging material, automatic handling and assembly equipment, and so on. If these are discharged into ESD sensitive components, either directly by contact with their pins, or via conductors in the system, damage or destruction is likely. Therefore it is essential that components are handled with adequate ESD precautions at all stages, and that protection is designed into circuits to safeguard components after assembly. Thereafter, care must be taken during test and maintenance, though the components will no longer be as vulnerable.

ESD can damage or destroy components even when they are unpowered, so precautions are necessary during all operations involving handling. Warning labels should be fixed to packages and equipments, and workbenches, tools and personnel must all be electrically grounded during assembly, repair and test.

Ohring (1998) is a good source for information on ESD.

9.6.3 Thermal Design

It is important to control the thermal design of electronic systems, so that maximum rated operating temperatures are not exceeded under worst cases of environment and load, and so that temperature variations within the system are not severe.

The reasons are that high temperatures can accelerate some failure modes in marginally defective components, and temperature cycling between ambient and high values can cause thermal fatigue of bonds and component structures, particularly if there are high local temperature gradients.

The maximum temperature generated within a device depends on the electrical load and the local ambient temperature, as well as the thermal resistance between the active part of the device and the external environment. Temperature at the active area of a device, for example the junction of a power transistor, or averaged over the surface of an IC, can be calculated using the formula

$$T_J = T_A + \theta W \tag{9.2}$$

where T_J is the local temperature at the active region referred as *junction temperature*, T_A is the ambient temperature around the component, W is the power dissipation, and θ is the thermal resistance between the active region and the ambient, measured in °C per Watt.

For devices that consume significant power levels in relation to their heat dissipation capacity, it is necessary to provide additional thermal protection. This can be achieved by mounting the device on a heat sink, typically a metal block, with fins to aid convective and radiant heat dissipation. Some devices, such as power transistors, have integral heat sinks. Small fans are also used, for example to cool microprocessors. Further protection can be provided by temperature-sensitive cut-off switches or relays; power supply and conversion units often include such features.

It is sometimes necessary to consider not only the thermal path from the component's active area to the local ambient, but to design to allow heat to escape from assemblies. This is essential in densely packaged systems such as avionics, military electronics and computers. In such systems a copper heat plane is usually incorporated into the PCB, to enable heat to flow from the components to the case of the equipment. Good thermal contact must be provided between the edge of the heat plane and the case. In turn the case can be designed to dissipate heat effectively, by the use of fins. In extreme cases liquid cooling systems are used, fluid being pumped through channels in the walls of the case or over heat sinks.

Temperature control can be greatly influenced by the layout and orientation of components and sub-assemblies such as PCBs. Hot components should be positioned downstream in the heat flow path (heat plane or air flow), and PCBs should be aligned vertically to allow convective air flow. Fans are often used to circulate air in electronic systems, to assist heat removal.

When additional thermal control measures are employed, their effects must be considered in evaluating component operating temperatures. The various thermal resistances, from the component active area to the external environment, must all be taken into account, as well as heat inputs from all heat-generating components, external sources such as solar radiation, and the effects of convection or forced cooling measures. Such a detailed thermal evaluation can best be performed with thermal modelling software, using finite element methods. Such software can be used to produce thermal maps of PCBs, taking into account each component's power load and all thermal resistances.

Thermal evaluation is important for any electronic design in which component operating temperatures might approach maximum rated values. Good detailed guidelines on thermal design for electronic systems are given in Harper (2004), McCluskey *et al.* (1997) and Sergent and Krum (1998).

9.6.4 Stress Derating

Derating is the practice of limiting the stresses which may be applied to a component, to levels below the specified maxima, in order to enhance reliability. Derating values of electrical stress are expressed as ratios of applied stress to rated maximum stress. The applied stress is taken as the maximum likely to be applied during worst case operating conditions.

Derating enhances reliability by:

1 Reducing the likelihood that marginal components will fail during the life of the system.
2 Reducing the effects of parameter variations.
3 Reducing long-term drift in parameter values.
4 Providing allowance for uncertainty in stress calculations.
5 Providing some protection against transient stresses, such as voltage spikes.

Typical derating guidelines are shown in Table 9.3, which gives electrical and thermal derating figures appropriate to normal and for critical (Hi-rel) applications such as spacecraft, or for critical functions within other systems. Such guidelines should usually be taken as advisory, since other factors such as cost or volume might be overriding. However, if stress values near to rated maxima must be used it is important that the component is carefully selected and purchased, and that stress calculations are doubled-checked.

Table 9.3 *Device derating guidelines.*

		Max. rating	
Device type	Parameter	Normal	Hi-rel
Microelectronics			
Digital	Power supply, input voltages	Derated but within performance spec.	
	Output current (load, fan-out)	0.8	0.8
	Junction temp. (hermetic): TTL	130 °C	100 °C
	(hermetic): CMOS	110 °C	90 °C
	(plastic)	100 °C	70 °C
	Speed	0.8	0.8
Linear	Power supply Derated but within performance spec.		
	Voltage		
	Input voltage	0.8	0.7
	Junction temp. (hermetic)	110 °C	90 °C
	(plastic)	100 °C	70 °C
Transistors			
Silicon (general purpose)	Collector current	0.8	0.5
	Voltage V_{cc}	0.8	0.6
	Junction temp. (hermetic)	120 °C	100 °C
	(plastic)	100 °C	80 °C
Silicon (power)	Collector current	0.8	0.6
	Voltage V_{cc}	0.8	0.6
	Voltage (reverse bias)	0.9	0.8
	Junction temp. (hermetic)	130 °C	110 °C
	(plastic)	110 °C	90 °C
Diodes			
Silicon (general purpose)	Forward current, voltages	0.8	0.5
Zener	Junction temp.	120 °C	100 °C
Resistors			
	Power dissipation	0.8	0.5
	Operating temp.	Rated−20 °C	Rated−40 °C
Capacitors			
	Voltage	0.8	0.5
	Operating temp.	Rated−20 °C	Rated−40 °C
Relays and switches			
Resistive or capacitive load	Current	0.8	0.5
Inductive load	Current	0.5	0.3
Motor	Current	0.3	0.2
Filament	Current	0.2	0.1

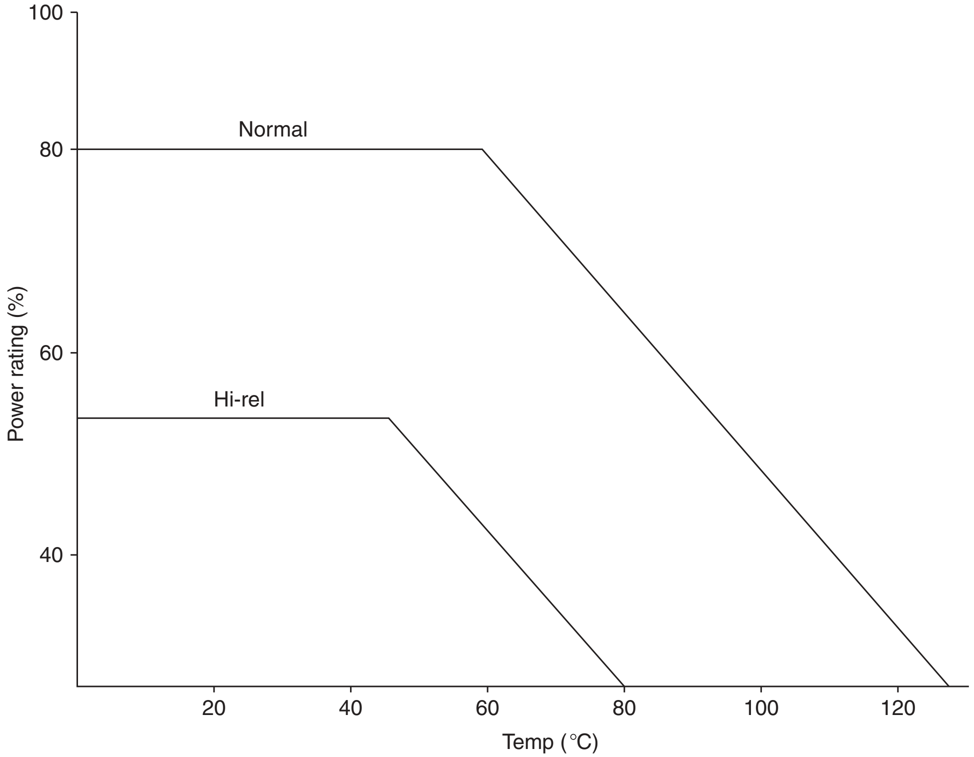


Figure 9.9 *Temperature–power derating for transistors and diodes (typical).*

Since thermal stress is a function of the surrounding temperature and power dissipation, combined temperature–power stress derating, as shown in Figure 9.9, is often advised for components such as power transistors. The manufacturers' databooks should be consulted for specific guidelines, and Pecht (1995) and US MIL-HDBK-338 also provide information.

The following are the most commonly used derating standards:

NAVSEA-TE000-AB-GTP-010: Parts Derating Requirements and Application Manual for Navy Electronic Equipment issued by the Naval Sea Systems Command. This standard provides derating curves for ten electrical and electronic parts.

MIL-STD-975M: NASA Standard Electrical, Electronic, and Electromechanical (EEE) Parts List, issued by the U.S. National Aeronautics and Space Administration in 1994. This standard provides part selection information and derating curves for electronics parts, materials and processes for space and launch vehicles.

MIL-STD-1547A: Electronic Parts, Materials and Processes for Space and Launch Vehicles, issued by the U.S. Department of Defense in November 1998. This standard provides part selection information and derating curves for electrical, electronic and electromechanical parts used in the design and construction of space flight hardware in space missions as well as essential ground support equipment (GSE).

ECSS-Q-30-11-A: Space Product Assurance, issued by the European Cooperation for Space Standardization in April, 2006. This standard provides derating requirements for electronic, electrical and electromechanical components used for space projects and applications.

9.6.5 Component Uprating

Component uprating is in a way opposite to derating and is intended to assess the ability of a part (typically electronic component) to meet the functionality and performance requirements in the applications in which it is used beyond the manufacturer's specifications. For example, large numbers of semiconductor parts in commercial applications are specified (or rated) to a maximum temperature of 70 °C and to a lesser extent to 85 °C (Das *et al.*, 2001). However, there is a need for parts requiring higher operating temperatures, especially in harsh environment applications such as automotive, avionics, military, and so on. Those industries do not generate large enough demand to stimulate semiconductor manufacturers to produce parts rated at higher temperatures forcing manufacturers in those industries to use the existing parts. Clearly this approach presents a risk of early failure and lower reliability, therefore uprating is designed to assess the part's ability to function in this environment and to assess the risk. Part uprating usually involves testing at the temperatures outside the specification and may be done in several different ways. The most common methods are: parameter conformance, parameter re-characterization, and stress balancing (Das *et al.*, 2001). The CALCE centre at the University of Maryland contributed to the development of uprating methods and test procedures. IEC TR 62240 and ANSI/EIA-4900-2002 are the two commonly used standards on the use of semiconductor devices outside the manufacturer's specified temperature range.

Another common reason for component uprating is cost reduction. In some non-critical applications designers might replace an electronic component with a cheaper part rated at the lower temperature, though within the application specifications limits. For example replacing an existing part rated at 85 °C with a similar part rated at 70 °C for applications where the maximum expected temperature is 70 °C. This type of uprating does not require any special test procedure, but reduces the reliability of the system due to lower temperature specifications of the new part. A method of estimating failure rates of the replacement parts based on the existing test data for the original part was discussed in Kleyner and Boyle (2003) both for derating and uprating.

9.6.6 Electromagnetic Interference and Compatibility (EMI/EMC)

Circuit design to prevent EMI is a difficult and challenging aspect of all modern electronic system design.

The main design techniques are:

1 The use of filter circuits to decouple noise and transients from or to the power supply.

2 Circuits and conductors can be shielded by enclosing them in grounded, conductive boxes (Faraday shields), or, in the case of cables, grounded conductive screens. Cables can also be made less susceptible to picking up noise by using a twisted pair arrangement.

3 Circuit impedances should be balanced, for example, between power supplies and loads, so that any noise pickup will be the same in each conductor, and will thus be self-cancelling.

4 All circuit grounds must be at the same electrical potential during circuit operation, and therefore all ground connections must provide a low impedance path back to the current source. This is particularly important in high frequency digital systems.

5 Contacts which make or break during circuit operation, for example, microswitches and relays, must be selected to minimize EMI, and if necessary filter circuits must be designed around them.

6 Digital systems must include noise filters at the PCB power input and near to each IC. The normal approach is to use decoupling capacitors. The capacitance value must be selected in relation to the circuit frequency, so that the resonant frequency of the local *L–C* circuit (see Figure 9.10) is well above the circuit operating frequency (to prevent resonance), but with a large enough capacitance to supply the transient current needed by the IC for its switching function.

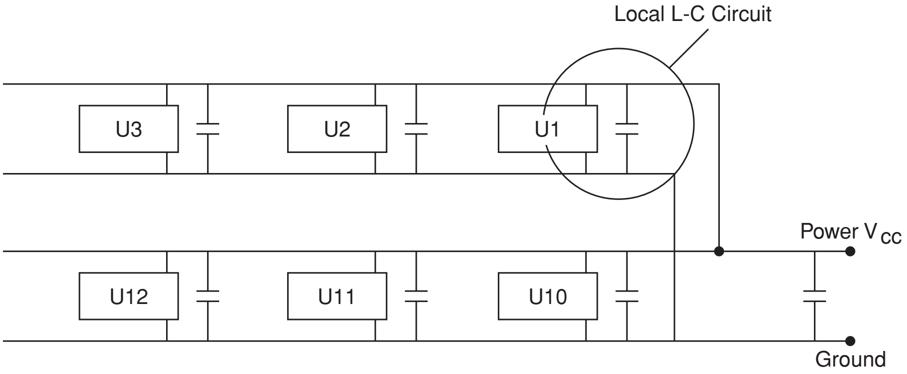


Figure 9.10 *Digital circuit noise decoupling.*

7 Optical fibres used for data transmission are immune to electromagnetic effects.
8 In software-driven systems, coding methods can be used to provide EMI protection, see Chapter 10.

Software is available for EMI/EMC analysis.

National and international regulations exist to set standards for electromagnetic and power line emissions and associated test methods.

Most of the protection techniques against transient high voltages, described earlier, are useful in relation to EMI, and vice versa. Therefore the topics are often combined in specialist texts and training. Ott (2009) and Schmitt (2002) are good introductions to the whole field of EMI and EMC.

9.6.7 Redundancy

Redundancy techniques were covered in Chapter 6. In electronic circuit and system design it is possible to apply redundancy at any level from individual components to subsystems. Decisions on when and how to design-in redundancy depend upon the criticality of the system or function and must always be balanced against the need to minimize complexity and cost. However, it is often possible to provide worthwhile reliability improvements by using redundant circuit elements, at relatively little cost, owing to the low cost of most modern devices. The most likely component failure modes must be considered; for example, resistors in parallel will provide redundant, though possibly degraded operation if one becomes open circuit, and short circuit is an unlikely failure mode. Opposite considerations apply to capacitors.

9.6.8 Design Simplification

Like all good engineering, electronic system designs must be kept as simple as practicable. The motto often quoted is KISS – ‘Keep it simple, stupid’. In electronics, design simplification is mainly a matter of minimizing the number of components to perform a required function. Reducing the number of components and their connections should improve reliability as well as reduce production costs. However, the need to provide adequate circuit protection and component derating, and where necessary redundancy, should normally take priority over component count reduction.

Minimizing the number of component types is also an important aspect of design simplification. It is inevitable that when a number of designers contribute to a system, different solutions to similar design problems will be used, resulting in a larger number of component types and values being specified than is necessary. This leads to higher production costs, and higher costs of maintenance, since more part types must be bought and stocked. It can also reduce reliability, since quality control of bought-in parts is made more difficult if an unnecessarily large number of part types must be controlled.

Design rules can be written to assist in minimizing component types, by constraining designers to preferred standard approaches. Component type reduction should also be made an objective of design review, particularly of initial designs, before prototypes are made or drawings frozen for production.

9.6.9 Sneak Analysis

A sneak circuit is an unwanted connection in an electrical or electronic circuit, not caused by component failure, which leads to an undesirable circuit condition or which can inhibit a desired condition. Sneak circuits can be inadvertently designed into systems when interfaces are not fully specified or understood, or when designers make mistakes in the design of complex circuitry. Sneak analysis is a technique developed to identify such conditions in electrical and electronic circuits, and in operating software.

It is based on the identification within the system of ‘patterns’ which can lead to sneak conditions. The five basic patterns are shown in Figure 9.11.

Any circuit can be considered as being made of combinations of these patterns. Each pattern is analysed to detect if conditions could arise, either during normal operation or due to a fault in another part of the system, that will cause a sneak. For example, in the power dome or the combination dome the power sources could be reversed if S1 and S2 are closed.

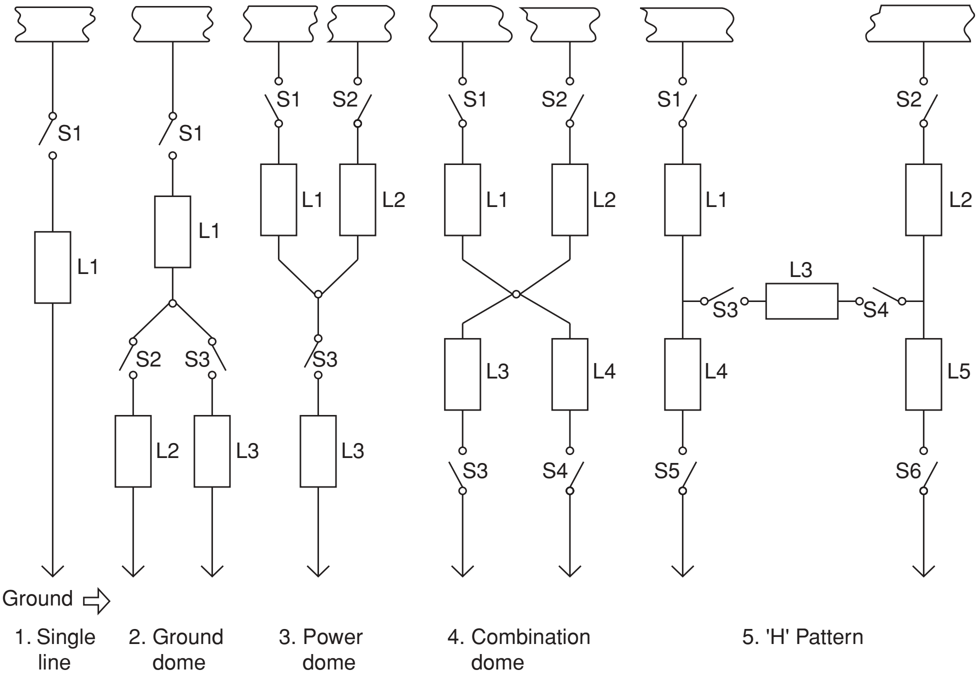


Figure 9.11 *Sneak analysis basic patterns (hardware).*

Sneak circuits are of five types:

1 *Sneak paths*. Current flows along an unexpected route.
2 *Sneak opens*. Current does not flow along an expected route.
3 *Sneak timing*. Current flows at the incorrect time or does not flow at the correct time.
4 *Sneak indications*. False or ambiguous indications.
5 *Sneak labels*. False, ambiguous or incomplete labels on controls or indicators.

When potential sneak conditions are identified, they must be validated by test or detailed investigation, reported and corrective action must be considered.

Sneak analysis is a tedious task when performed on relatively large systems. However, it has proved to be very beneficial in safety analysis and in assessing the integrity of controls in aircraft and industrial systems. Most of today's circuit simulation software packages have the capability of running sneak circuit analysis. Software applications are described in Chapter 10.

Sneak circuits can be avoided by careful design, or detected by adequate testing. The formal analysis technique is appropriate for critical systems, particularly when there are complex interfaces between sub-systems.

9.7 Parameter Variation and Tolerances

9.7.1 Introduction

All electrical parameters of electronic components are subject both to initial component-to-component variation, and sometimes to long-term drift. Parameter values can also vary as a result of other factors, particularly temperature. Whether these variations are important or not in a particular design depends upon the requirements for accuracy of the parameters in that application. For example, the resistance value of a resistor in a feedback circuit of a high gain amplifier might be critical for correct operation, but would not need to be as closely controlled in a resistor used as a current limiter.

Initial variation is an inevitable consequence of the component production processes. Most controlled parameters are measured at the end of production, and the components are assigned to tolerance bands, or rejected if they fall outside the limits. For example, typical resistors are provided in tolerance ranges of 1, 5, 10 and 20 % about the nominal resistance. Since the selection is often from the same batch, which may have had a parameter distribution as shown in Figure 9.12, the parameter distributions of the selected tolerance ranges would be as shown, assuming only two tolerance bands had been selected. Depending upon the application, knowledge of the shape of the parameter distribution might be important.

For many component parameters, for example transistor characteristics, maximum and minimum values are stated. It is also important to note that not all parameter values are controlled in manufacture and selection. Some parameters are given only as 'typical' values. It is never a good idea to design critical circuit operation around such parameters, since they are not usually measured, and therefore are not guaranteed.

Since conductance, both of conductor and of semiconductor materials, varies with temperature, all associated parameters will also vary. Therefore resistance of resistors, and gain and switching time of transistors, are typical of temperature-dependent parameters. High temperature can also increase noise outputs. Other parameters can interact, for example the capacitance values between transistor connections is affected by bias voltage.

Parameter drift with age is also usually associated with changes in conductance, as well as in dielectric performance, so that resistors and capacitors are subject to drift, at rates which depend upon the type of materials and construction used, operating temperature and time.

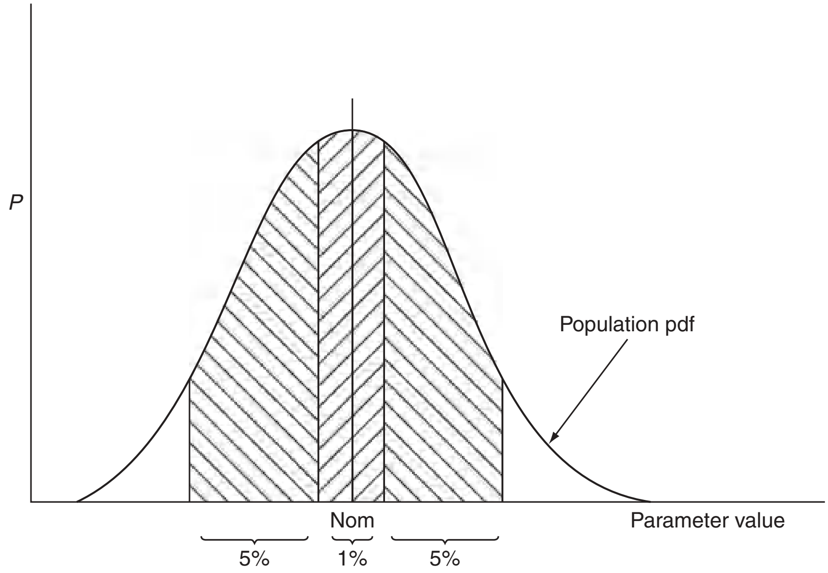


Figure 9.12 *Parameter distributions after selection.*

Another source of circuit parameter variation arises from what are referred to as *parasitic* parameters. These are electrical parameters that are not intrinsic to the theoretical design of the component or the circuit, but which are due to construction and layout features. For example, wire wound resistors are inductive, the inductance depending upon the type of construction, PCB conductor tracks have mutual inductance and capacitance, and integrated circuit lead frames are inductive. Parasitic effects can be very important, and difficult to control, in high gain and high frequency systems.

Parameter variation can affect circuits in two ways. For circuits which are required to be manufactured in quantity, a proportion might not meet the required operating specification, and production yield will then be less than 100 %, adding to production cost. Variation can also cause circuits to fail to work correctly in service. Initial component-to-component variation mainly affects yield, and stress or time related variation mainly affects reliability.

9.7.2 Tolerance Design

Every electronic circuit design (for that matter, not only electronic circuits, but any system) must be based on the nominal parameter values of all the components that will contribute to correct performance. This is called *parameter design*. It encompasses the qualities of knowledge and inventiveness necessary for the solution of design problems. However, having created the correct functional design, it is necessary to evaluate the effects of parameter variation on yield, stability and reliability. This is called *tolerance design*.

The first step in tolerance design is to determine which parameter values are likely to be most sensitive in affecting yield and performance. This can be performed initially on the basis of experience and system calculations made during the parameter design stage. However, a more systematic approach, using the techniques described below, should be used for serious design.

The next step is to determine the extent of variation of all of the important parameter values, by reference to the detailed component specifications. This step is sometimes omitted by designers using lists of ‘preferred parts', which list only primary nominal values and tolerances, without giving full details of other data relevant to the application. All relevant parasitic parameters should also be evaluated at this stage.

It is possible to design-in compensating features for some variation, for example by using temperature compensation components such as thermistors or adjustable components such as variable resistors. However, these add complexity and usually degrade reliability, since they are themselves prone to drift, and adjustable components can be degraded by wear, vibration and contamination. Wherever possible the design should aim for minimum performance variation by the careful selection of parameter values and tolerances.

9.7.3 Analysis Methods

The analysis of tolerance effects in general design situations will be covered in Chapter 11. This section introduces further methods available to analyse the effects of variation and tolerances in electronic circuits.

9.7.3.1 Worst Case Analysis

Worst case analysis (WCA) involves evaluating the circuit performance when the most important component parameter values are at their highest and lowest tolerance values. It is a straightforward extension of the parameter design calculations. However, it is only realistic for simple circuits. Also, purely digital circuits are relatively easy to analyse in this respect, so long as frequency and timing requirements are not too severe. However, if there are several parameters whose variation might be important, particularly if there are interactions or uncertainty, then more powerful methods of analysis should be used.

9.7.3.2 The Transpose Circuit

The sensitivity of the output of a circuit to parameter changes can be analysed using the *transpose circuit* method. A transpose circuit is a circuit which is topologically identical to that under investigation, except that the forward and reverse transmission coefficients of each 3-terminal device (e.g. transistor) are interchanged, the input is replaced by an open circuit, and the output by a current source of 1A. Figure 9.13 shows a circuit and a possible transpose of it. Here we are looking only at those components whose parameter variations we consider might most affect the output, in this case V_0, when the input is I_i. Note that these are instantaneous values of terms in the frequency domain (or DC values).

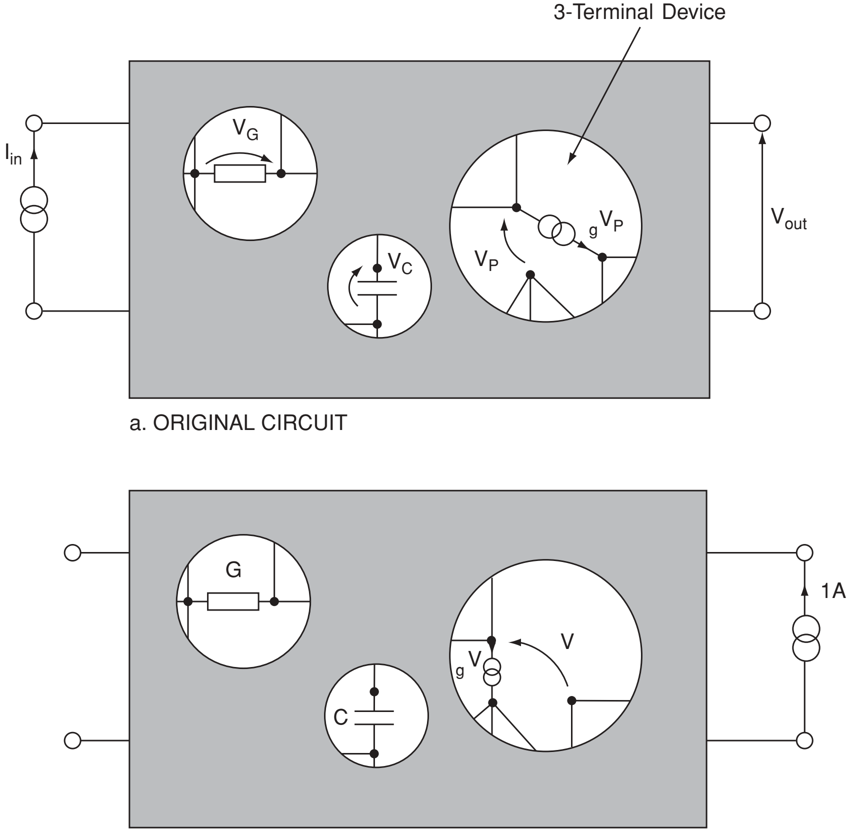
The sensitivity of V_0 to small changes in the conductance G of the resistor is given by

$$\frac{\partial V_0}{\partial G} = -V_\mathrm{G} V_\mathrm{G/T} \tag{9.3}$$

where $\mathrm{V_G}$ is the voltage across the conductance in the actual circuit and $V_\mathrm{G/T}$ is the voltage across the conductance in the transpose circuit. Similar relations hold for other component parameters, if the two circuits are analysed at the same frequency. This is *Tellegen's theorem*. Using these relationships, the sensitivities to all the critical parameters can be evaluated, by analysing only two different circuits, using for example circuit simulation software. This is an extremely efficient technique for analysing the effects of small, single variations. It is described in Spence and Soin (1988).

9.7.3.3 Simulation

Another method for analysing the effects of tolerances and variation is simulation, using the Monte Carlo method. The principles of Monte Carlo simulation were described in Chapter 4. Most modern circuit



a. ORIGINAL CIRCUIT

b. TRANSPOSE CIRCUIT

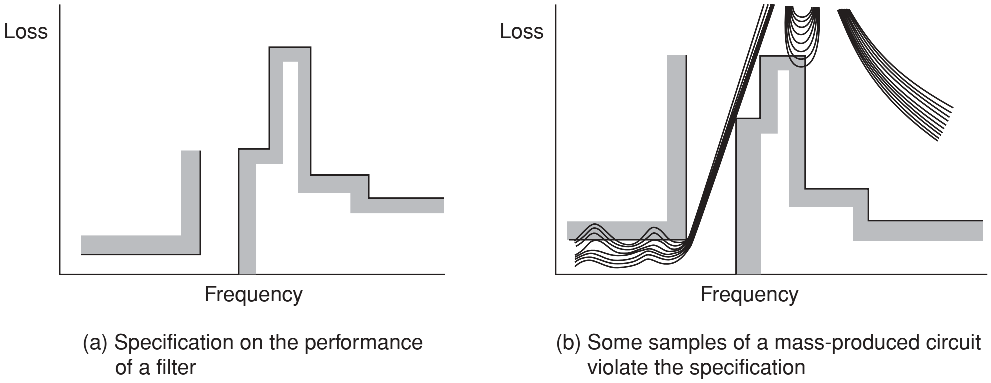
$$\frac{\partial V_o}{\partial G} = -V_G V_{GT}, \text{ etc}$$

Figure 9.13 *Transpose circuit (from Spence and Soin (1988)).*

simulation software packages include Monte Carlo techniques, which allow parameter variations to be set, either as a range or as a defined distribution shape (e.g. normal, bimodal), and the program will then randomly ‘build’ circuits using parameter values from the distributions, and will analyse these. Monte Carlo circuit simulation evaluates the effects of multiple simultaneous variations. Refinements allow the most critical parameters to be identified, and statistically designed experiments (see Chapter 11) can be run. Circuits can be analysed in the time and frequency domains, and there are no practical limitations, apart from the time required to run each analysis, on the number of parameters, input conditions and simulations.

Figure 9.14 shows the results of a number of Monte Carlo simulations of a filter circuit, in relation to the specification. This shows that some parameter combinations give performance outside the specification. Carrying out a number of runs provides an estimate of production yield, and the particular parameter value combinations that caused circuits to be outside the specification can be identified.

See Singhal and Vlach (2010) and Spence and Soin (1988) for descriptions of the topics covered in this section.



(a) Specification on the performance of a filter

(b) Some samples of a mass-produced circuit violate the specification

Figure 9.14 *Monte Carlo analysis of filter circuit (from Spence and Soin (1988)).*

9.8 Design for Production, Test and Maintenance

Electronic circuits should be designed to be testable, using the test methods that will be applied in production. These methods depend upon the manufacturing test policy and economics, as discussed in Chapter 15. Testability is an important design feature, which can make a significant impact on production costs. Testability also affects reliability, since production defects which are not detected by the tests can lead to failure in service, and circuits which are difficult to diagnose are more likely to be inadequately or incorrectly repaired.

It is important that the circuit designer is aware of the test methods that will be used, and the requirements that they impose on the design to ensure effective and economic test coverage. Test methods for electronics are described in O'Connor (2001).

Circuit designs must allow the ATE to initialize the operating states of components, control the circuit operation, observe and measure output states and values, and partition the circuit to reduce test program complexity. It is good practice to conduct a careful review, with the test engineers, of the testability of the design before it is finalized. Modern EDA software includes facilities for performing testability analysis. Turino and Binnendyk (1991) cover the subject of design for test in detail.

Design of electronic systems for production, test and maintainability should be included in design rules and design review. Aspects of good electronic design which contribute to these are listed below:

1 Avoid the necessity for adjustments, for example, potentiometers, whenever possible. Adjustable components are less reliable than fixed-valued components and are more subject to drift.
2 Avoid 'select on test' situations, where components must be selected on the basis of measured parameter values. Specify components which can be used anywhere within the tolerance range of the applicable parameters and do not rely on parameters which are listed as typical rather than guaranteed. Where 'typical' tolerances must be used, ensure that the appropriate component screening is performed before assembly.
3 Ensure that adjustments are easily accessible at the appropriate assembly level.
4 Partition circuits so that subassemblies can be tested and diagnosed separately. For example, if several measured values require amplification, analogue-to-digital conversion, logic treatment and drivers for displays, it might be better to include all functions for each measured value on one PCB rather than use a PCB for each function, since fault diagnosis is made easier and repair of one channel does not affect the performance or calibration of others. However, other factors such as cost and space must also be considered.

Questions

1. How does increasing temperature affect the reliability of electronic components? Illustrate your answer with three examples of specific failure mechanisms and component types.
2. Briefly describe three failure modes that can occur in modern integrated circuits. In each case, explain how they are influenced by temperature (high/low), electrical stresses, and manufacturing quality.
3. Describe the design, manufacturing and application factors that influence the following failure modes of integrated circuits: (i) electromigration; (ii) latch-up; (iii) electrostatic damage.
4. Screening is a process used to improve the quality and reliability of integrated circuits. Explain the engineering justification for IC screening, and describe briefly the tests that are typically applied in the screening process.
5. How have the recent developments in microcircuit device packaging affected reliability? How can the risks of failure be minimized?
6. What are the main factors to consider, from the reliability point of view, when selecting the type of packaging for the integrated circuits to be used in a circuit designed for mass production? Consider hermetic versus plastic packaging, and through-hole versus surface mounting.
7. Suggest ways in which the capability approval process is a more useful approach to ASIC line approval than the approved components system.
8. Use the equation from MIL-HDBK-217 from Chapter 6 $\lambda_p = \pi_Q\pi_L[C_1\pi_T + C_2\pi_E]/10^6\text{h}$ to calculate the device *hazard rate* for a device screened to level B-1 (π_Q = 2.0) from the following data: $\pi_L = 1$, $C_1 = 0.12$, $\pi_T = 3.7$, $\pi_V = 1$, $C_2 = 0.01$, $\pi_E = 4.2$. What is the use of the hazard rate calculated in this way?
9. Discuss the reliability of electronic components as an overall contributor to system reliability in modern systems.
10. Describe the ways in which solder connections can fail. How can circuit designers minimize solder failures?
11. Describe the ways in which electrical cables and connectors can fail. How can circuit designers minimize these failures?
12. What are ‘no fault found’ failures? Describe the main causes of such failures. Why are they important?
13. Why are the thermal aspects of electronic system design important for reliability? What methods can designers use to reduce the operating temperatures of electronic components?
14. A supplier has presented you with a derating policy based on deriving the derating curves from power stress versus base hazard rate curves from MIL-HDBK-217 data. State your reservations concerning this approach.
15. a for a small plastic transistor operating at 120 mW, estimate T_J if $\theta = 0.4\,^\circ\text{C}\,\text{mW}^{-1}$ above 25 °C, if the ambient temperature is 50 °C.
    b If the maximum junction temperature is 150 °C, estimate what power the transistor will dissipate at an ambient temperature of 60 °C.
16. Using Figure 9.9 and Table 9.3, estimate the percentage power allowed for a general-purpose silicon plastic-sealed transistor in a Hi-rel application.
17. State some of the advantages of employing thermal derating techniques in an electronic design.
18. What are the main sources of electromagnetic interference (EMI) that can affect electronic systems? Describe three methods that can be used to protect circuits from EMI.
19. In a normal domestic kitchen containing a fluorescent light fitting and a washing machine, list the EMI sources you may find and how as a designer you may mitigate these effects.
20. Describe three methods for analysing the effects of component parameter variations on the performance of an electronic circuit. For each, describe how the variations and their effects can be minimized by the designer.

21. Discuss and compare reliability characteristics of tin-lead vs. lead free solder. Describe pros and cons of the transition to lead-free solder. Make examples of where you would prefer tin-lead and where you would prefer lead-free solder.
22. You conducted a series of experiments with electromigration at two different temperatures $T_1 = 25$ °C and $T_2 = 60$ °C. The current density in the first case was $2\text{mA}/\mu\text{m}^2$ and in the second $1\text{mA}/\mu\text{m}^2$; the activation energy is $E_A = 0.6\,\text{eV}$. As a result of the experiments $\text{MTTF}_1 = 1457$ hours and $\text{MTTF}_2 =$ 500 hours. Determine both experimental constants A and N in Black's equation (9.1).
23. A resistor is advertised as being rated at 2 Watt. You have a choice of using this resistor in a 1 Watt application and in a 3 Watt application. Which case would constitute derating and which is uprating? Explain your answer.

Bibliography

General

CALCE (2011) *Lead Free and Green Electronics Forum*, CALCE, University of Maryland. Available at: http://www.calce.umd.edu/lead-free/index.html.

Clech, J-P. (2004) *Lead-Free and Mixed Assembly Solder Joint Reliability Trends*. Proceedings (CD-ROM), IPC / SMEMA Council APEX 2004 Conference, Anaheim, CA, Feb. 23-26, 2004, pp. S28-3-1 through S28-3-14. Available at http://www.jpclech.com/EPSI_Publications.html.

Clech, J-P., Henshall, G. and Miremadi, J. (2009) Closed-Form, Strain-Energy Based Acceleration Factors for Thermal Cycling of Lead-Free Assemblies, Proceedings of SMTA International Conference (SMTAI 2009), Oct. 4-8, 2009, San Diego, CA.

DfR Solutions (2010) *Pb-Free Solder Joints*. Available at http://www.dfrsolutions.com/pb-free-solder-joints/.

Evans, J. (2010) *A Guide to Lead-free Solders. Physical Metallurgy and Reliability* (ed. W. Engelmaier), Springer.

Garrou, P. and Gedney, R. (2004) *Tin Whiskers: An Industry Perspective*. Advanced Packaging, December 2004 issue.

Horowitz P. and Hill, W. (1989) *The Art of Electronics*, 2nd edn, Cambridge University Press.

JEDEC (2006) JESD201 standard, *Environmental Acceptance Requirements for Tin Whisker Susceptibility of Tin and Tin Alloy Surface Finishes*. Available at http://www.jedec.org/standards-documents.

Pascoe, N. (2011) *Reliability Technology: Principles and Practice of Failure Prevention in Electronic Systems*, Wiley.

Pecht, M. (ed.) (1995) *Product Reliability, Maintainability and Supportability Handbook*. ARINC Research Corp.

US MIL-HDBK-338. *Electronic Reliability Design Handbook*. Available from the National Technical Information Service, Springfield, Virginia.

Components

Amerasekera, E. and Najm, F. (1997) *Failure Mechanisms in Semiconductor Devices*, 2nd edn, Wiley.

Bajenescu, T. and Bazu, M. (2011) *Failure Analysis: A Practical Guide for Manufacturers of Electronic Components and Systems*, J. Wiley.

Brindley, K. and Judd, M. (1999) *Soldering in Electronics Assembly*, Newnes.

Bajenescu, T. and Bazu, M. (1999) *Reliability of Electronic Components*, Springer-Verlag.

British Standard, BS 9000. *Components of Assessed Quality*. British Standards Institution, London. (And equivalent US MIL, European (CECC) and International (IEC) standards.)

Das, D., Pendse, N., Wilkinson, C. and Pecht, M. (2001) Parameter Recharacterization: A Method of Thermal Uprating. *IEEE Transactions on Components and Packaging Technologies*, **24**(4), December 2001

Hannemann, R., Kraus, A. and Pecht, M. (1997) *Semiconductor Packaging: A Multidisciplinary Approach*, Wiley.

Harper, C. (2004) *Electronic Packaging and Interconnection Handbook*, 4th edn, McGraw-Hill Professional.

IEC TR 62240 (2005) *Process management for avionics –Use of semiconductor devices outside manufacturers' specified temperature range*.

Jensen, F. (1995) *Electronic Component Reliability*, Wiley.

Kleyner, A. and Boyle, J. (2003) Reliability Prediction of Substitute Parts Based on Component Temperature Rating and Limited Accelerated Test Data. *Proceedings of Annual Reliability and Maintainability Symposium*, Tampa, Florida, pp. 518–522.

Kuo, W., Chien, W. and Kim, T. (1998) *Reliability, Yield, and Stress Burn-In: a Unified Approach for Microelectronics Systems Manufacturing & Software Development*, Springer.

Lau, J., Wong, C., Prince J. and Nakayama, W. (1998) *Electronic Packaging: Design, Materials, Process, and Reliability*, McGraw-Hill.

Ohring, M. (1998) *Reliability and Failure of Electronic Materials and Devices*, Academic Press.

Pecht, M. (ed.) (1993) *Soldering Processes and Equipment*, Wiley.

Pecht, J. and Pecht, M. (eds) (1995) *Long-Term Non-Operating Reliability of Electronic Products*, CRC Press.

Pecht, M., Radojcic, R. and Rao, G. (1999) *Guidebook for Managing Silicon Chip Reliability*, CRC Press.

Tummala, R. (2001) *Fundamentals of Microsystems Packaging*, McGraw-Hill.

Tummala, R., Rymaszewski, E. and Klopfenstein, A. (1997) *Microelectronics Packaging Handbook*, Kluwer Academic Publishers.

US MIL-STD-883. *Test Methods and Procedures for Microelectronic Devices*. Available from the National Technical Information Service, Springfield, Virginia.

Woodgate, R. (1996) *The Handbook of Machine Soldering: SMT and TH*, 3rd edn, Wiley-Interscience.

Mechanical and thermal effects and design

Lall, P., Pecht, M. and Hakim, E. (1997) *Influence of Temperature on Microelectronics and System Reliability*, CRC Press.

McCluskey, P., Grzybowski R. and Podlesak, T. (eds) (1997) *High Temperature Electronics*, CRC Press.

Sergent, J. and Krum, A. (1998) *Thermal Management Handbook: For Electronic Assemblies*, McGraw-Hill.

Steinberg, D. (2000) *Vibration Analysis for Electronic Equipment*, 3rd edn, Wiley.

Thermal Guide for Reliability Engineers. Rome Air Development Center Report TR-82-172. Available from the National Technical Information Service, Springfield, Virginia.

EMI/EMC/ESD

Chatterton, P. and Houlden, M. (1991) *EMC: Electromagnetic Theory to Practical Design*, Wiley.

Ott, H. (2009) *Electromagnetic Compatibility Engineering*, Wiley-Interscience.

Schmitt, R. (2002), *Electromagnetics Explained: A Handbook for Wireless/RF, EMC, and High-Speed Electronics*. Elsevier.

US MIL-E-6051D, Electromagnetic Compatibility Requirements, Systems. NTIS, Springfield, VI.

US MIL-STD-461D, Requirements for the Control of Electromagnetic Interference Emisions and Susceptibility. NTIS, Springfield VI.

Tolerance design and electronics testing

O'Connor, P.D.T. (2001) *Test Engineering*. Wiley.

Singhal, K. and Vlach, J. (2010) *Computer Methods for Circuit Analysis and Design*, Kluwer Academic Publishers.

Spence, R. and Soin, R. (1988) *Tolerance Design in Electronic Circuits*, Addison-Wesley.

Turino, J. and Binnendyk, H. (1991) Design to Test, 2nd edn, Logical Solutions Inc., Campbell, CA (Also *Testability Advisor* software).

10

Software Reliability

10.1 Introduction

Software is now part of the operating system of a very wide range of products and systems, and this trend continues to accelerate with the opportunities presented by low cost microcontroller devices. Software is relatively inexpensive to develop, costs very little to copy, weighs nothing, and does not fail in the ways that hardware does. Software also enables greater functionality to be provided than would otherwise be feasible or economic. Performing functions with software leads to less complex, cheaper, lighter and more robust systems. Therefore software is used increasingly to perform functions that otherwise would be performed by hardware, and even by humans. Recent examples are aircraft flight control systems, robotic welders, engine control systems, domestic bread-making machines, and so on.

The software 'technology' used today is the same basic sequential digital logic first applied in the earliest computers. The only significant changes have been in the speed and word length capability of processors and the amount of memory available, which in turn have enabled the development of high-level computer languages and modern operating systems. Some attempts have been made to develop radically different approaches such as parallel processing and fuzzy logic, but these remain fringe applications. Therefore, the basic principles of software development, to ensure that programs are correct, safe and reliable, remain largely unchanged since they were first described in the 1970s (e.g. Myers, 1976).

Every copy of a computer program is identical to the original, so failures due to variability cannot occur. Also, software does not degrade, except in a few special senses,[1] and when it does it is easy to restore it to its original standard. Therefore, a correct program will run indefinitely without failure, and so will all copies of it. However, software can fail to perform the function intended, due to undetected errors. When a software error ('bug') does exist, it exists in all copies of the program, and if it is such as to cause failure in certain circumstances, the program will always fail when those circumstances occur.

Software failures can also occur as a function of the machine environment, for example, machines can be restarted and the software 'fixed' by clearing queues, removing memory leaks, and refreshing the state of the machine. So identical copies can behave differently depending on their 'age' since rebooting.

[1] Data or programs stored in some media can degrade. Magnetic media such as discs are susceptible to electromagnetic fields, or even to being closely packed for long periods. VLSI semiconductor devices can suffer changes in the voltage state of individual memory cells due to naturally occurring alpha-particle bombardment. In each case a refresh cycle will restore the program.

Practical Reliability Engineering, Fifth Edition. Patrick D. T. O'Connor and Andre Kleyner.

Software errors can cause system failure effects that can range from trivial to catastrophic. Therefore, software reliability and safety effort must be directed at the consequences of errors, not just at the prevention or removal of most of them.

Since most programs consist of very many individual statements and logical paths, all created by the efforts of humans, there is ample scope for errors. Therefore we must try to prevent the creation of errors, and maximize the likelihood of detecting and correcting those that are created, by imposing programming disciplines, by checking and by testing.

When software is an integral part of a hardware-software system, system failures might be caused by hardware failures or by software errors. When humans are also part of the system they can also cause failures (e.g. the Airbus crash during a low flying display, which some 'experts' immediately blamed on the new flight control software, but which the investigation concluded was caused by the pilot putting the aircraft into a situation from which the system could not prevent the crash). In some cases, it might be difficult to distinguish between hardware, software and human causes.

There are several ways by which hardware and software reliability differ. Some have already been mentioned. Table 10.1 lists the differences.

10.2 Software in Engineering Systems

The software that forms an integral part or sub-system of an engineering system is in some important ways different from software in other applications, such as banking, airline booking, logistics, CAE, PC operating systems and applications, and so on. The differences are:

- Engineering programs are 'real time': they must operate in the system timescale, as determined by the system clock and as constrained by signal propagation and other delays (switches, actuators, etc.). A chess program or a circuit simulation program, for example, will run when executed, and it is not critical exactly how long it takes to complete the run. However, in an operational system such as a process controller or an autopilot, it is essential that the software is ready to accept inputs and completes tasks at the right times. The software must be designed so that functions are correctly timed in relation to the system clock pulses, task execution times, interrupts, and so on. Timing errors are a common cause of failure in real-time systems, particularly during development. They are often difficult to detect, particularly by inspection of code. Timing errors can be caused by hardware faults or by interface problems. However, logic test instruments (logic analysers) can be used to show exactly when and under what conditions system timing errors occur, so that the causes can be pinpointed.
- Engineering programs share a wider range of interfaces with the system hardware. In addition to basic items such as processors, memory, displays and keyboards, other engineering interfaces can include measurement sensors, A/D and D/A converters, signal analysers, switches, connectors, and so on.
- Engineering programs might be 'embedded' at different levels within a system: for example the main operating program might be loaded and run from disc or accessible PROM devices, but other software might be embedded in components which are less accessible, such as ASICs, programmable gate arrays, signal processing ICs and flash memory devices. The BIOS chip in a PC is also an example of software embedded in this way.
- There is often scope for alternative solutions to design problems, involving decisions on which tasks should be performed by hardware (or humans) and which by software.
- Engineering software must sometimes work in electrically 'noisy' environments, so that data might be corrupted.
- Engineering programs are generally, though not always, rather smaller and simpler than most other applications.

Table 10.1 Comparison of Hardware and Software Reliability Characteristics.

	Hardware	Software
1	Failures can be caused by deficiencies in design, production, use and maintenance.	Failures are primarily due to design faults. Repairs are made by modifying the design to make it robust against the condition that triggered the failure.
2	Failures can be due to wear or other energy-related phenomena. Sometimes warning is available before failure occurs (e.g. system can become noisy indicating degradation and impending failure).	There are no wearout phenomena. Software failures occur without warning, although very old code can exhibit an increasing failure rate as a function of errors introduced into the code while making functional code upgrades.
3	No two items are identical. Failures can be caused by variation.	There is no variation: all copies of a program are identical.
4	Repairs can be made to make equipment more reliable. This would be the case with preventive maintenance where a component is restored to an as new condition.	There is no repair. The only solution is redesign (reprogramming), which, if it removes the error and introduces no others, will result in higher reliability.
5	Reliability can depend on burn-in or wearout phenomena; that is, failure rates can be decreasing, constant or increasing with respect to time.	Reliability is not so time-dependent. Reliability improvement over time may be affected, but this is not an operational time relationship. Rather, it is a function of reliability growth of the code through detecting and correcting errors.
6	Reliability may be time-related, with failures occurring as a function of operating (or storage) time, cycles, etc.	Reliability is not time related. Failures occur when a specific program step or path is executed or a specific input condition is encountered, which triggers a failure.
7	Reliability may be related to environmental factors (temperature, vibration, humidity, etc.)	The external environment does not affect reliability except insofar as it might affect program inputs. However, the program reliability is a function of the internal machine environment (queues, memory leakage, etc.)
8	Reliability can be predicted, in principle but mostly with large uncertainty, from knowledge of design, parts, usage, and environmental stress factors.	Reliability cannot be predicted from any physical bases, since it entirely depends on human factors in design. Some approaches exist based on the development process used and the extent of the code, but these are controversial.
9	Reliability can be improved by redundancy. The successful use of redundancy presumes ready detection, isolation, and switching of assets.	Reliability cannot be improved by redundancy if the parallel paths are identical, since if one path fails, the other will have the error. It is possible to provide redundancy by having diverse parallel paths with different programs written by different teams.
10	Failures can occur in components of a system in a pattern that is, to some extent, predictable from the stresses on the components and other factors. Reliability critical lists are useful to identify high risk items.	Failures are rarely predictable from analyses of separate statements. Errors are likely to exist randomly throughout the program, and any statement may be in error. Most errors lie on the boundary of the program or in its exception handling. Reliability critical lists are not appropriate.
11	Hardware interfaces are visual; one can see a Ten-pin connector.	Software interfaces are conceptual rather than visual.
12	Computer-aided design systems exist that can be used to create and analyse designs.	There are no computerized methods for software design and analysis. Software design is more of an 'art form' lacking the provability of hardware, except to a limited extent through formal methods (see later).
13	Hardware products use standard components as basic building blocks.	There are no standard parts in software, although there are standardised logic structures. Software reuse is being deployed, but on a limited basis.

Therefore, it is very important that engineering software is developed (specified, designed, programmed, tested, managed) in close integration with the hardware and overall system work. The not uncommon practice of writing a software specification then 'outsourcing' the program development work should not be an option for important engineering software.

10.3 Software Errors

Software errors ('bugs') can arise from the specification, the software system design and from the coding process.

10.3.1 Specification Errors

Typically more than half the errors recorded during software development originate in the specification. Since software is not perceivable in a physical sense, there is little scope for common sense interpretation of ambiguities, inconsistencies or incomplete statements. Therefore, software specification must be very carefully developed and reviewed. The software specification must describe fully and accurately the requirements of the program. The program must reflect the requirements exactly. There are no safety margins in software design as in hardware design. For example, if the requirement is to measure 9 V ± 0.5 V and to indicate if the voltage is outside these tolerances, the program will do precisely that. If the specification was incorrectly formulated, for example, if the tolerances were not stated, the out-of-tolerance voltage would be indicated at this point every time the measured voltage varied by a detectable amount from 9 V, whether or not the tolerances were exceeded. Depending upon the circumstances this might be an easily detectable error, or it might lead to unnecessary checks and adjustments because the out-of-tolerance indication is believed. This is a relatively simple example. Much more serious errors, such as a misunderstanding or omission of the logical requirement of the program, can be written into the specification. This type of error can be much harder to correct, involving considerable reprogramming, and can be much more serious in effect.

The Eurospace Ariane 5 spacecraft launcher failure was caused by such an error: the guidance computer and the inertial measurement unit used different bit formats for numerical data, but, even though this fact was known, no compensation was made because it had not resulted in failures on previous Ariane launchers. The new launcher's greater rocket thrust led to an overflow when the inertial unit measured velocities higher than experienced before. The NASA Mars Polar Orbiter spacecraft collided with the planet because part of the system was designed using measurements in miles while an interfacing subsystem used kilometres.

The specification must be logically complete. Consider the statement: 'Sample inputs A, B and C. If any one exceeds by $> \pm 10$ units the average of the other two, feed forward the average of these two. Indicate failure of the out-of-tolerance input. If the out-of-tolerance condition does not exist, feed forward the average of the three inputs.'

This is an example of a two-out-of-three majority voting redundant system. The logic is shown in the flow diagram, Figure 10.1. Consider the situation when the values of A, B and C are 100, 90 and 120. The values of the derived parameters, and route taken by the program, are shown in Figure 10.1. In this case, two fault conditions exist, since both B and C exceed the average of the other two inputs. The program will indicate a B failure, as the algorithm compares B before C. The specification has not stated what should happen in the event of more than one input being out of tolerance. The program will work as shown in the algorithm, in the sense that an input will always be available, but the system may not be safe. The flowchart complies with the specification, but it probably does not reflect the real wishes of the specification writer. A software specification must cover all the possible input conditions and output requirements, and this usually requires much more detailed consideration than for a hardware specification.

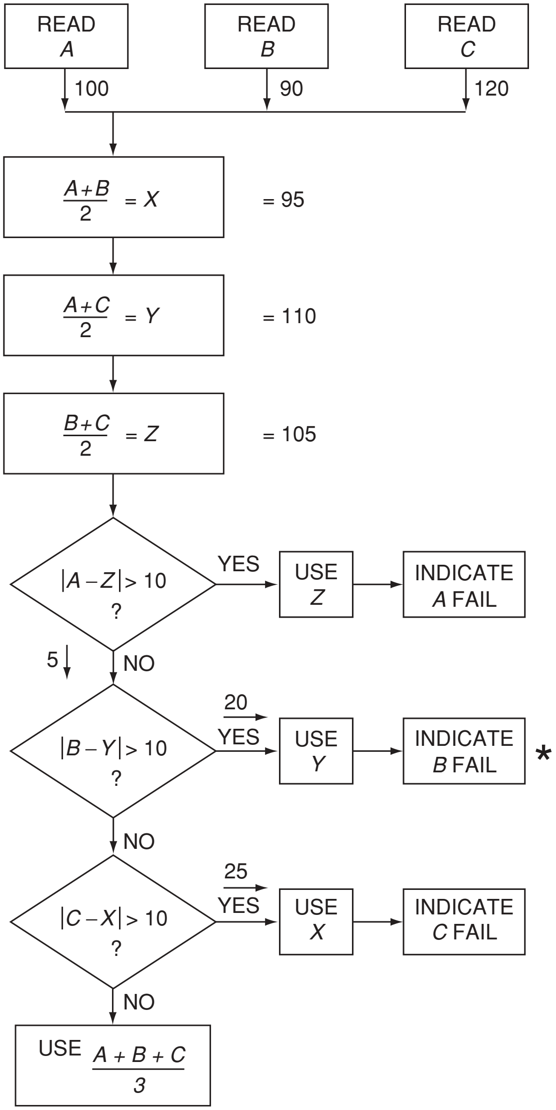


Figure 10.1 *Voting redundant system.*

The specification must be consistent. It must not give conflicting information or use different conventions in different sections (e.g. miles and kilometres).

The specification must not include requirements that are not testable, for example, accuracy or speed requirements that are beyond the capability of the hardware.

The specification should be more than just a description of program requirements. It should describe the structure to be used, the program test requirements and documentation needed during development and test, as well as basic requirements such as the programming language, and inputs and outputs. (Program structure, test and documentation will be covered later.)

10.3.2 Software System Design

The software system design follows from the specification. The system design may be a flowchart and would define the program structure, test points, limits, and so on. Errors can occur as a result of incorrect interpretation of the specification, or incomplete or incorrect logic. Errors can also occur if the software cannot handle data inputs that are incorrect but possible, such as missing or incorrect bits.

An important reliability feature of software system design is *robustness*, the term used to describe the capability of a program to withstand error conditions without serious effect, such as becoming locked in a loop or 'crashing'. The robustness of the program will depend upon the design, since it is at this stage that the paths to be taken by the program under error conditions are determined.

10.3.3 Software Code Generation

Code generation is a prime source of errors, since a typical program involves a large number of code statements. Typical errors can be:

- Typographical errors. (sic).
- Incorrect numerical values, for example, 0.1 for 0.01.
- Omission of symbols, for example, parentheses.
- Inclusion of variables which are not declared, or not initialized at the start of program run.
- Inclusion of expressions which can become indeterminate, such as division by a value which can become zero.
- Accidental shared use of memory locations.

Changes to code can have dire consequences. The likelihood of injecting new faults can run as high as 50 %, and tends to be highest for small code changes. The injected faults tend to be more obscure and harder to detect and remove. Changes can be in conflict with the original architecture and increase code complexity.

We will briefly describe the methods that can be used to minimize the creation of errors, and to detect errors that might have been created.

10.4 Preventing Errors

10.4.1 Specification

The overall system specification and the software specification must be prepared in harmony. Both should allow flexibility in relation to allocation of functions and should encourage integrated engineering.

Software specifications should be more than just descriptions of requirements. They must describe the functions to be performed, in full and unambiguous detail, and the operating environment (hardware, memory allocation, timing, etc.). They should also describe explicitly all of the conditions that must NOT be allowed to occur. They should describe the program structure to be used, the program test requirements and documentation needed during development, as well as basic requirements such as the programming language, memory allocations, inputs and outputs. By adequately specifying these aspects, a framework for program generation will be created which minimizes the possibilities for creating errors, and which ensures that errors will be found and corrected.

The specifications must be carefully reviewed, to ensure that they meet all of the requirements described above, and contain no ambiguities. Specification review must be performed by the project team, including the programmers and engineers whose work will be driven by the specifications.

10.5 Software Structure and Modularity

10.5.1 Structure

Structured programming is an approach that constrains the programmer to using certain clear, well-defined approaches to program design, rather than allowing total freedom to design ‘clever’ programs which might be complex, difficult to understand or inspect, and prone to error. A major source of error in programs is the use of the GOTO statement for constructs such as loops and branches (decisions). The structured programming approach therefore discourages the use of GOTOs, requiring the use of control structures which have a single entry and a single exit. For example, the simple branch instruction in Figure 10.2 can be programmed (using BASIC) in either an unstructured or a structured way as shown. The unstructured approach can lead to errors if the wrong line number is given (e.g. if line numbers are changed as a result of program changes), and it is difficult to trace the subroutines (A, B) back to the decision point.

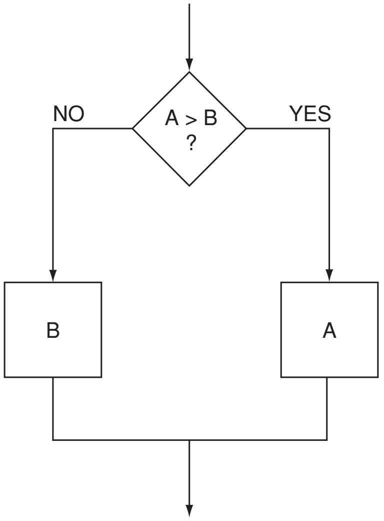
On the other hand, the structured approach eliminates the possibility of line number errors, and is much easier to understand and to inspect.

Structured programming leads to fewer errors, and to clearer, more easily maintained software. On the other hand, structured programs might be less efficient in terms of speed or memory requirements.

10.5.2 Modularity

Modular programming breaks the program requirement down into separate, smaller program requirements, or modules, each of which can be separately specified, written and tested. The overall problem is thus made easier to understand and this is a very important factor in reducing the scope for error and for easing the task of checking. The separate modules can be written and tested in a shorter time, thus reducing the chances of changes of programmer in mid-stream.

Each module specification must state how the module is to interface with other parts of the program. Thus, all the inputs and outputs must be specified. Structured programming might involve more preparatory work



Unstructured:

```
If A > B goto A [line number]
else goto B [line number]
A [line number]
B [line number]
```

Structured:

```
If A > B then A [subroutine]
else B [subroutine]
```

Figure 10.2 Structured versus unstructured programming.

in determining the program structure, and in writing module specifications and test requirements. However, like good groundwork in any development programme, this effort is likely to be more than repaid later by the reduced overall time spent on program writing and debugging, and it will also result in a program which is easier to understand and to change. The capability of a program to be modified fairly easily can be compared to the maintainability of hardware, and it is often a very important feature. Program changes are necessary when logical corrections have to be made, or when the requirements change, and there are not many software development projects in which these conditions do not arise.

The optimum size of a module depends upon the function of the module and is not solely determined by the number of program elements. The size will usually be determined to some extent by where convenient interfaces can be introduced. As a rule of thumb, modules should not normally exceed 100 separate statements or lines of code in a high level language, and less in assembler code.

10.5.3 Requirements for Structured and Modular Programming

Major software customers specify the need for programs to be structured and modular, to ensure reliability and maintainability. These disciplined approaches can greatly reduce software development and life cycle costs. ISO/IEC90003 covers structured and modular programming in more detail.

10.5.4 Software Re-Use

Sometimes existing software, for example from a different or previous application, can be used, rather than having to write a new program or module. This approach can lead to savings in development costs and time, as well as reducing the possibility of creating new errors. However, be careful! Remember Ariane 5 and Mars Polar Orbiter!

Computer-aided design systems, for example Labview® and Simulink®, include embedded software for components in their databases.

10.6 Programming Style

Programming style is an expression used to cover the general approach to program design and coding. Structured and modular programming are aspects of style. Other aspects are, for example, the use of ‘remark’ statements in the listing to explain the program, ‘defensive’ programming in which routines are included to check for errors, and the use of simple constructs whenever practicable. Obviously, a disciplined programming style can have a great influence on software reliability and maintainability, and it is therefore important that style is covered in software design guides and design reviews, and in programmer training.

10.7 Fault Tolerance

Programs can be written so that errors do not cause serious problems or complete failure of the program. We have mentioned ‘robustness’ in connection with program design, and this is an aspect of fault tolerance. A program should be able to find its way gracefully out of an error condition and indicate the error source. This can be achieved by programming internal tests, or checks of cycle time, with a reset and error indication if the set conditions are not met. Where safety is a factor, it is important that the program sets up safe conditions when an error occurs. For example, a process controller could be programmed to set up known safe conditions and indicate a problem, if no output is generated in two successive program cycle times or if the output value changes by more than a predetermined amount.

These software techniques can also be used to protect against hardware failures, such as failure of a sensor which provides a program input. Examples of this approach are:

– checks of cycle time for a process (e.g. time to fill a tank), and automatic shutdown if the correct time is exceeded by a set amount. This might be caused by failure of a sensor or a pump, or by a leak.
– failure of a thermostat to switch off a heating supply can be protected against by ensuring that the supply will not remain on for more than a set period, regardless of the thermostat output.
– checks for rates of change of input values. If a value changes by more than a predetermined amount take corrective action as above. For example, a pressure measurement might abruptly change to zero because of a transducer or connector failure, but such an actual pressure change might be impossible. The system should not be capable of inappropriate response to a spurious input.
– allow two or more program cycles for receipt of input data, to allow for possible data loss, interruption or corruption.

Features such as these can be provided much more easily with software than with hardware, at no extra material cost or weight, and therefore, the possibility of increasing the reliability and safety of software controlled systems should always be analysed in the specification and design stages. Their provision and optimization is much more likely when the software development is managed as part of an integrated, system approach.

10.8 Redundancy/Diversity

Fault tolerance can also be provided by program redundancy. For high integrity systems separately coded programs can be arranged to run simultaneously on separate but connected controllers, or in a time-sharing mode on one controller. A voting or selection routine can be used to select the output to be used. This approach is also called *program diversity*. The effectiveness of this approach is based on the premise that two separately coded programs are very unlikely to contain the same coding errors, but of course this would not provide protection against a specification error. Redundancy can also be provided within a program by arranging that critical outputs are checked by one routine, and if the correct conditions are not present then they are checked by a different routine (Figure 10.3).

10.9 Languages

The selection of the computer language to be used can affect the reliability of software. There are three main approaches which can be used:

1 Machine code programming.
2 Assembly level programming.
3 High level (or high order) language (HLL or HOL) programming.

Machine code programming is the creation of the microcode that the processor runs. However, programming at this level should not be used, since it confers no advantages in speed or memory, is very prone to creation of errors, is extremely difficult to check, and has no error trap capabilities.

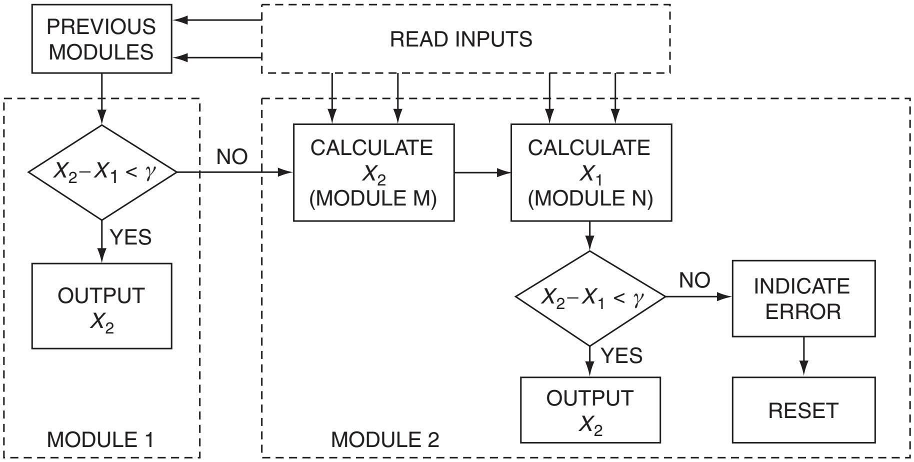


Figure 10.3 *Fault tolerant algorithm.*

Assembly level programs are faster to run and require less memory than HLLs. Therefore they can be attractive for real-time systems. However, assembly level programming is much more difficult and is much harder to check and to modify than HLLs. Several types of error which can be made in assembly level programming cannot be made, or are much less likely to be made, in a HLL. Therefore, assembly level programming is not favoured for relatively large programs, though it might be used for modules in order to increase speed and to reduce memory requirements. *Symbolic assemblers*, however, have some of the error-reduction features of HLLs.

Machine code and assembly programming are specific to a particular processor, since they are aimed directly at the architecture and operating system.

HLLs are processor-independent, working through a *compiler* which converts the HLL to that processor's operating system. Therefore, HLLs require more memory (the compiler itself is a large program) and they run more slowly. However, it is much easier to program in HLLs, and the programs are much easier to inspect and correct. The older HLLs (FORTRAN, BASIC) do not encourage structured programming, but the more recently developed ones (PASCAL, Ada, C, C++) do.

Since HLLs must work through a compiler, the reliability of the compiler can affect system reliability. Compilers for new HLLs and for new processors sometimes cause problems for the first few years until all errors are found and corrected. Generally speaking, though, compilers are reliable once fully developed, since they are so universally used. Modern compilers contain error detection, so that many logical, syntactical or other errors in the HLL program are displayed to the programmer, allowing them to be corrected before an attempt is made to load or run it. Automatic error correction is also possible in some cases, but this is limited to certain specific types of error.

Fuzzy logic is used to a limited extent in some modern systems. The ways in which fuzzy logic programs can fail are basically the same as for conventional logic.

Programmable logic controllers (PLCs) are often used in place of processors, for systems such as machine tools, factory automation, train door controls, and so on. Programming of PLCs is much easier than for microprocessors, since only basic logic commands need to be created. PLC-based systems also avoid the need for the other requirements of processor-based systems, such as operating system software, memory, and so on, so they can be simpler and more robust, and easier to test.

10.10 Data Reliability

Data reliability (or information integrity) is an important aspect of the reliability of software-based systems. When digitally coded data are transmitted, there are two sources of degradation:

1 The data might not be processed in time, so that processing errors are generated. This can arise, for example, if data arrive at a processing point (a 'server', e.g. a microprocessor or a memory address decoder) at a higher rate than the server can process.
2 The data might be corrupted in transmission or in memory by digital bits being lost or inverted, or by spurious bits being added. This can happen if there is noise in the transmission system, for example, from electromagnetic interference or defects in memory.

System design to eliminate or reduce the incidence of failures due to processing time errors involves the use of queueing theory, applied to the expected rate and pattern of information input, the number and speed of the 'servers', and the queueing disciplines (e.g. first-in-first-out (FIFO), last-in-first-out (LIFO), etc.). Also, a form of redundancy is used, in which processed data are accepted as being valid only if they are repeated identically at least twice, say, in three cycles. This might involve some reduction in system processing or operating speed.

Data corruption due to transmission or memory defects is checked for and corrected using error detection and correction codes. The simplest and probably best known is the parity bit. An extra bit is added to each data word, so that there will always be an even (or odd) number of ones (even (or odd) parity). If an odd number of ones occurs in a word, the word will be rejected or ignored. More complex error detection codes, which provide coverage over a larger proportion of possible errors and which also correct errors, are also used. Examples of these are Hamming codes and BCH codes.

Ensuring reliable data transmission involves trade-offs in memory allocation and operating speed.

10.11 Software Checking

To confirm that the specification is satisfied, the program must be checked against each item of the specification. For example, if a test specification calls for an impedance measurement of $15 \pm 1\ \Omega$, only a line-by-line check of the program listing is likely to discover an error that calls for a measurement tolerance of $+1\ \Omega$, $-\ 0\ \Omega$. Program checking can be a tedious process, but it is made much easier if the program is structured into well-specified and understandable modules, so that an independent check can be performed quickly and comprehensively. Like hardware design review procedures, the cost of program checking is usually amply repaid by savings in development time at later stages. The program should be checked in accordance with a prepared plan, which stipulates the tests required to demonstrate specification compliance.

Formal program checking, involving the design team and independent people, is called a *structured walkthrough*, or a *code review*.

10.11.1 FMECA

It is not practicable to perform a failure modes, effects and criticality analysis (FMECA) (Chapter 7) on software, since software 'components' do not fail. The nearest equivalent to an FMECA is a code review, but whenever an error is detected it is corrected so the error source is eliminated. With hardware, however, we cannot eliminate the possibility of, say, a transistor failure. Attempts have been made to develop FMECA methods tailored for application to software, but these have not been generally adopted or standardized.

In performing a FMECA of an engineering system that combines hardware and software it is necessary to consider the failure effects in the context of the operating software, since system behaviour in the event of a hardware failure might be affected by the software, as described above. This is particularly the case in systems utilizing built-in-test software, or when the software is involved in functions such as switching redundancy, displays, warnings and shut-down.

10.11.2 Software Sneak Analysis

The sneak analysis (SA) method described in Chapter 9 for evaluating circuit conditions that can lead to system failure is also applicable to software. Since a section of code does not fail but performs the programmed functions whether or not they are the intended ones, there is an analogy with an erroneous circuit design.

The program must be reduced to a set of topological patterns, as for hardware SA. Since a program of reasonable size is very difficult to reduce in this way, this step is usually computerized.

Six basic sneak patterns exist, as shown in Figure 10.4. Note that most software sneak patterns are related to branching instructions, such as GOTO or IF THEN/ELSE statements. The conditions leading to and deriving from such statements, as well as the statements themselves, are important clues in the SA.

Software sneak conditions are:

1 *Sneak output*. The wrong output is generated.
2 *Sneak inhibit*. Undesired inhibit of an input or output.
3 *Sneak timing*. The wrong output is generated because of its timing or incorrect input timing.
4 *Sneak message*. A program message incorrectly reports the state of the system.

Figure 10.1 illustrates a potential sneak message condition, since the program will not indicate that *C* has failed if *A* and *B* have failed. This failure is brought about by an incorrect line pattern. The program

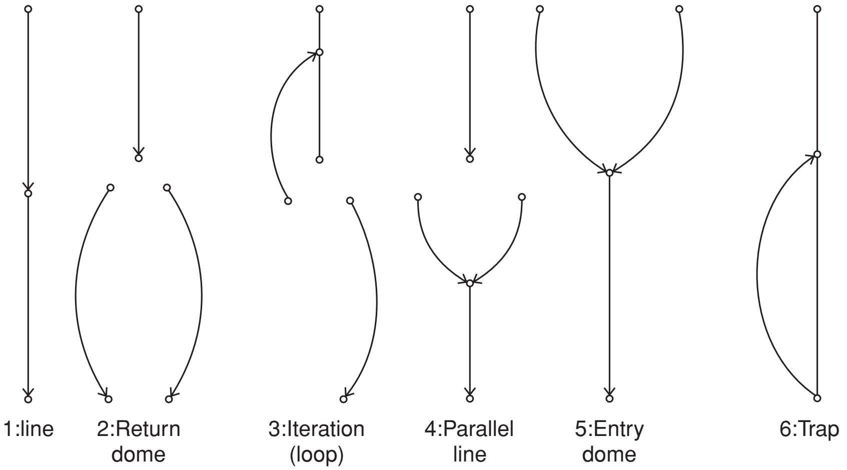


Figure 10.4 *Software sneak patterns.*

correctly identifies the correct A value and the incorrect B value, and proceeds to the output with no chance of testing C.

10.12 Software Testing

The objectives of software testing are to ensure that the system complies with the requirements and to detect remaining errors. Testing that a program will operate correctly over the range of system conditions is an essential part of the software and system development process. Software testing must be planned and executed in a disciplined way since, even with the most careful design effort, errors are likely to remain in any reasonably large program, due to the impracticability of finding all errors by checking, as described above. Some features, such as timing, overflow conditions and module interfacing, are not easy to check.

Few programs run perfectly the first time they are tested. The scope for error is so large, due to the difficulty that the human mind has in setting up perfectly logical structures, that it is normal for some time to be spent debugging a new program until all of the basic errors are eliminated.

There are limitations to software testing. It is not practicable to test exhaustively a reasonably complex program. The total number of possible paths through a program with n branches and loops is 2^n, analogous to the digital circuit testing problem discussed in Chapter 9. However, there is no 'ATE' for software, so all tests must be set up, run and monitored manually. It is not normally practicable to plan a test strategy which will provide high theoretical error coverage, and the test time would be exorbitant. Therefore, the tests to be performed must be selected carefully to verify correct operation under the likely range of operating and input conditions, whilst being economical.

The software test process should be iterative, whilst code is being produced. Code should be tested as soon as it is written, to ensure that errors can be corrected quickly by the programmer who wrote it, and it is easier to devise effective tests for smaller, well-specified sections of code than for large programs. The earliest testable code is usually at the module level. The detection and correction of errors is also much less expensive early in the development programme. As errors are corrected the software must be re-tested to confirm that the redesign has been effective and has not introduced any other errors. Later, when most or all modules have been written and tested, the complete program must be tested, errors corrected, and retested. Thus, design and test proceed in steps, with test results being fed back to the programmers.

It is usual for programmers to test modules or small programs themselves. Given the specification and suitable instructions for conducting and reporting tests, they are usually in the best position to test their own work. Alternatively, or additionally, programmers might test one another's programs, so that an independent approach is taken. However, testing of larger sections or the whole program, involving the work of several programmers, should be managed by a person with system responsibility, though members of the programming team should be closely involved. This is called *integration testing*. Integration testing covers module interfaces, and should demonstrate compliance with the system specification.

The software tests must include:

- All requirements defined in the specification ('must do' and 'must not do' conditions).
- Operation at extreme conditions (timing, input parameter values and rates of change, memory utilization).
- Ranges of possible input sequences.
- Fault tolerance (error recovery).

Since it may not be practicable to test for the complete range of input conditions it is important to test for the most critical ones and for combinations of these. Random input conditions, possibly developed from system simulation, should also be used when appropriate to provide further assurance that a wide range of inputs is covered.

Software can be tested at different levels:

- *White box* testing involves testing at the detailed structural level, for aspects such as data and control flow, memory allocation, look-ups, and so on. It is performed in relation to modules or small system elements, to demonstrate correctness at these levels.
- *Verification* is the term sometimes used to cover all testing in a development or simulated environment, for example, using a host or lab computer. Verification can include module and integration testing.
- *Validation* or *black box* testing covers testing in the real environment, including running on the target computer, using the operational input and output devices, other components and connections. Validation is applicable only to integration testing, and it covers the hardware/software interface aspects, as described earlier.

These terms are not defined absolutely, and other interpretations are also applied.

10.12.1 Managing Software Testing

Software testing must be managed as an integral part of the overall system test plan. It is essential to plan the software tests with full understanding of how software can fail and in relation to the interfaces with the system hardware. In the system test plan, the software should be treated as a separate subsystem (verification) and as part of the overall system (validation).

The test specifications (for modules, integration/verification, validation) must state every test condition to be applied, and the test reports must indicate the result of each test.

Formal 100 % error reporting should also be started at this stage, if it is not already in operation. Obviously, all errors must be corrected, and the action taken must also be reported (changes to specification, design, code, hardware, as appropriate). The relevant test must be repeated to ensure that the correction works, and that no other errors have been created.

Formal configuration control should be started when integration testing commences, to ensure that all changes are documented and that all program copies at the current version number are identical.

For validation and other system tests, failures caused by software should be reported and actioned as part of an integrated engineering approach, since the most appropriate solutions to problems could involve hardware or software changes.

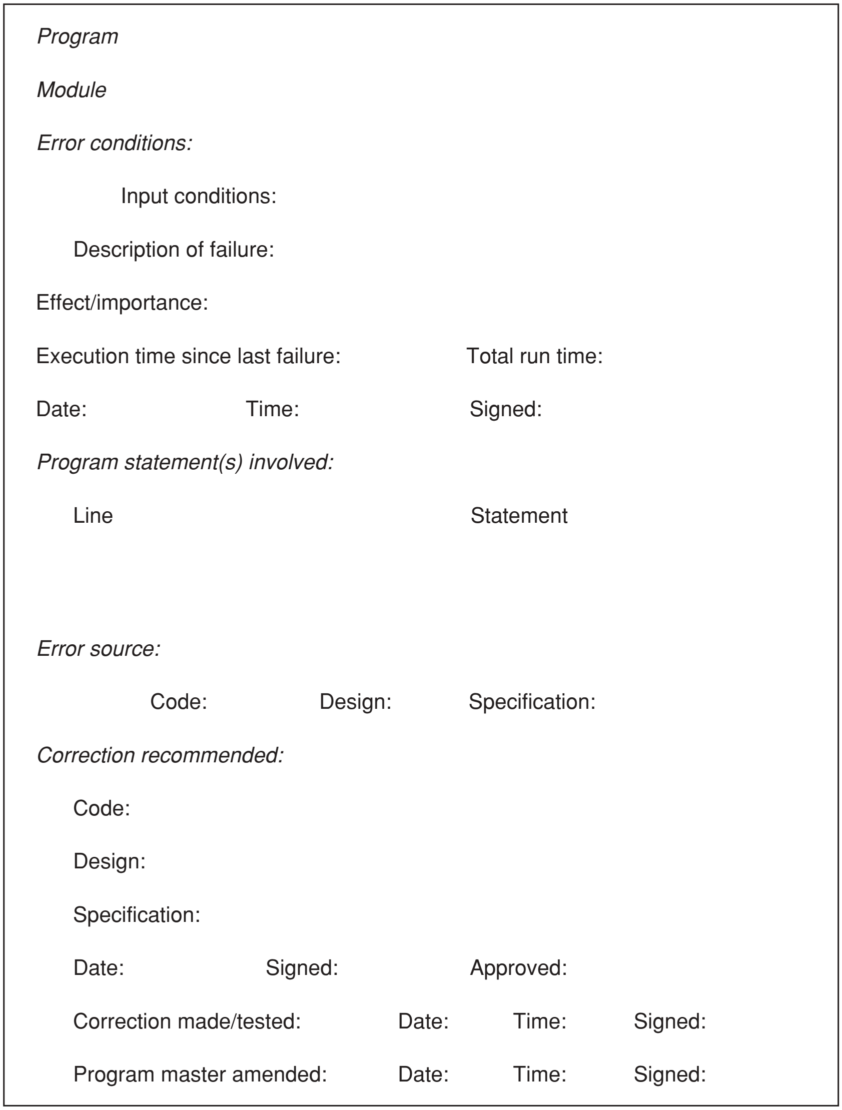
Each software test needs to be performed only once during development, unless there have been changes to the program or to related hardware. Of course there is no need to test software as part of the system production test process, since the software cannot vary from copy to copy or change over time.

There are some who argue that software testing (and checking) should be performed by people who are entirely independent of the programming team. This is called the '*cleanroom*' approach to software development. The approach is controversial, and is not consistent with the philosophy on which this book is based.

Ould and Unwin (1986), Beizer (1995), Patton (2006), and Kaner *et al*. (1999) provide more information on software testing.

10.13 Error Reporting

Reporting of software errors is an important part of the overall program documentation. The person who discovers an error may not be the programmer or system designer, and therefore all errors, whether discovered during checking, testing or use, need to be written up with full details of program operating conditions at the time. The corrective action report should state the source of the error (specification, design, coding) and describe the changes made. Figure 10.5 shows an example of a software error reporting form. A software



Program

Module

Error conditions:

Input conditions:

Description of failure:

Effect/importance:

Execution time since last failure: Total run time:

Date: Time: Signed:

Program statement(s) involved:

Line Statement

Error source:

Code: Design: Specification:

Correction recommended:

Code:

Design:

Specification:

Date: Signed: Approved:

Correction made/tested: Date: Time: Signed:

Program master amended: Date: Time: Signed:

Figure 10.5 *Software error reporting form.*

error reporting and corrective action procedure is just as important as a failure reporting system for hardware. The error reports and corrective action details should be retained with the module or program folder as part of the development record.

10.14 Software Reliability Prediction and Measurement

10.14.1 Introduction

Efforts to quantify software reliability usually relate to predicting or measuring the probability of, or quantity of, errors existing in a program. Whilst this is a convenient starting point, there are practical difficulties.

The reliability of a program depends not only upon whether or not errors exist but upon the probability that an existing error will affect the output, and the nature of the effect. Errors which are very likely to manifest themselves, for example, those which cause a failure most times the program is run, are likely to be discovered and corrected during the development phase. An error which only causes a failure under very rare or unimportant conditions may not be a reliability problem, but the coding error that caused the total loss of a spacecraft, for example, was a disaster, despite all the previous exhaustive checking and testing.

Error generation, and the discovery and correction of errors, is a function of human capabilities and organization. Therefore, whilst theoretical models based upon program size might be postulated, the derivation of reliability values is likely to be contentious. For example, a well-structured modular program is much easier to check and test, and is less prone to error in the first place, than an unstructured program designed for the same function. A skilled and experienced programming team is less likely to generate errors than one which is less well endowed. A further difficulty in software reliability modelling is the fact that errors can originate in the specification, the design and the coding. With hardware, failure is usually a function of load, strength and time, and whether a weakness is due to the specification, the design or the production process, the physics of failure remain the same. With software, however, specification, design and coding errors are often different in nature, and the probability of their existence depends upon different factors. For example, the number of coding errors might be related to the number of code statements, but the number of specification errors might not have the same relationship. One specification error might lead to a number of separate program errors.

The following sections briefly outline some of the statistical models which have been proposed for software reliability. See Musa *et al.* (1987) for a detailed discussion of software reliability prediction and measurement. However, it is important to appreciate that the utility of any statistical software reliability model depends upon acceptable values for the distribution parameters being available. Unlike hardware failure statistics, there is no physical basis for parameter estimation and the data that have been analysed to date are very limited compared with the wealth of available hardware failure data. Since software reliability is so dependent upon human performance and other non-physical factors, data obtained on one program or group of programs are unlikely to be accepted as being generally applicable, in the way that data on material properties are.

The logical limitations inherent in the prediction of reliability as described in Chapter 6 apply equally to software. Indeed they are even more severe, since there are no physical or logical connections between past data and future expectation, as there are with many hardware failure modes. Therefore the methods described in this section are of mainly academic interest, and they have not been generally accepted or standardized by the software engineering community.

10.14.2 The Poisson Model (Time-Related)

It is assumed that errors can exist randomly in a code structure and that their appearance is a function of the time the program is run. The number of errors occurring in time t is $N(t)$. If the following conditions exist:

1 $N(0) = 0$,
2 not more than one error can occur in the time interval $(t, t + \mathrm{d}t)$,
3 the occurrence of an error is independent of previous errors.

then the occurrence of errors is described by the non-homogeneous Poisson distribution:

$$P[N(t) = n] = \frac{[m(t)]^n}{n!} \exp[-m(t)] \quad (n \geq 0) \tag{10.1}$$

where

$$m(t) = \int_0^t \lambda(s)\mathrm{d}s$$

$m(t)$ is the mean (expected) number of errors occurring in the interval (0, t):

$$m(t) = a[1 - \exp(-bt)]$$

where a is the total number of errors and b is a constant. The number of errors remaining after time t, assuming that each error which occurs is corrected without the introduction of others, is

$$\bar{N}(t) = a\exp(-bt) \tag{10.2}$$

The reliability function, after the most recent error occurs and is corrected at time s, is

$$R(t) = \exp[-a\{\exp(-bs) - \exp[-b(s+t)]\}] \tag{10.3}$$

In using a time-related model, the question arises as to what units of time should be used. The Poisson model has been tested against software error data using calendar time during which errors were detected and corrected and values for the parameters a and b derived. However, since software errors are not time-related in the way that physical (hardware) failure processes are, the use of time-related models for software errors is problematical.

10.14.3 The Musa Model

The Musa model uses program execution time as the independent variable. A simplified version of the Musa model is

$$n = N_0\left[1 - \exp\left(\frac{-Ct}{N_0T_0}\right)\right] \tag{10.4}$$

where N_0 is the inherent number of errors, T_0 the MTTF at the start of testing (MTTF is mean time to failure) and C the ‘testing compression factor’ equal to the ratio of equivalent operating time to testing time.

The present MTTF:

$$T = T_0\exp\left(\frac{Ct}{N_0T_0}\right)$$

gives

$$\mathrm{R}(t) = \exp\left(\frac{-t}{T}\right) \tag{10.5}$$

From these relationships we can derive the number of failures which must be found and corrected, or the program execution time necessary, to improve from T_1 to T_2:

$$\Delta_n = N_0 T_0 \left(\frac{1}{T_1} - \frac{1}{T_2} \right) \tag{10.6}$$

$$\Delta_t = \left(\frac{N_0 T_0}{C} \right) \ln \left(\frac{T_2}{T_1} \right) \tag{10.7}$$

Example 10.1

A large program is believed to contain about 300 errors and the recorded MTTF at the start of testing is 1.5 h. The testing compression factor is assumed to be 4. How much testing is required to reduce the remaining number of errors to ten? What will then be the reliability over 50 h of running?

From Eqs. (10.6) and (10.7),

$$(300 - 10) = 300 \times 1.5 \left(\frac{1}{1.5} - \frac{1}{T_2} \right)$$

$$t = \left(\frac{300 \times 1.5}{4} \right) \ln \left(\frac{T_2}{1.5} \right)$$

Therefore

$$T_2 = 45\,\text{h}$$

and

$$\Delta_t = 382.6\,\text{h}$$

giving

$$\text{R}_{50} = \exp \left(\frac{-50}{45} \right) = 0.33$$

10.14.4 The Jelinski–Moranda and Schick–Wolverton Models

Two other exponential-type models which have been suggested are the Jelinski–Moranda (JM) model and the Schick–Wolverton (SW) model. In the JM and SW models, the hazard function h(t) is given respectively by:

$$\text{h}(t_i) = \phi[N_0 - n_{i-1}] \tag{10.8}$$

$$\text{h}(t_i) = \phi[N_0 - n_{i-1}]t_i \tag{10.9}$$

where t_i is the length of the ith debugging interval, that is the time between the $(i - 1)$th and the ith errors, and ϕ is a constant.

10.14.5 Littlewood Models

Littlewood attempts to take account of the fact that different program errors have different probabilities of causing failure. If $\phi_1, \phi_2, \ldots, \phi_N$ are the rates of occurrence of errors $1, 2, \ldots, N$, the pdf for the program time to failure, after the ith error has been fixed, is

$$\mathrm{f}(t) = \lambda \exp(-\lambda t) \tag{10.10}$$

where λ is the program failure rate

$$\lambda = \phi_1 + \phi_2 + \cdots \phi_{N-i}$$

φ is assumed to be gamma-distributed, that is errors do not have constant rates of occurrence but rates which are dependent upon program usage. If the gamma distribution parameters are (α, β) (equivalent to $(a, 1/\lambda)$ in Eq. 2.28) then it can be shown, using a Bayes approach, that

$$\mathrm{f}(t) = \frac{(N-i)\alpha(\beta + t')^{(N-i)\alpha}}{(\beta + t' + t)^{1+(N-i)\alpha}} \tag{10.11}$$

where t' is the time taken to detect and correct i errors. From this

$$\mathrm{R}(t) = \left(\frac{\beta + t'}{\beta + t' + t}\right)^{(N-i)\alpha} \tag{10.12}$$

and

$$\lambda(t) = \frac{(N-i)\alpha}{\beta + t' + t} \tag{10.13}$$

At each error occurrence and correction, $\lambda(t)$ falls by an amount $\alpha/(\beta + t')$. It is assumed that all detected errors are corrected, without further errors being introduced.

Example 10.2

A large program is assumed to include a total of 300 errors, of which 250 have been detected and corrected in 20 h of execution time. Assuming the Littlewood model holds and the distribution parameters are $\alpha = 0.005$, $\beta = 4$, what is the expected reliability over a further 20 h?

From Eq. (10.12),

$$\begin{aligned}\mathrm{R}(20) &= \left(\frac{4+20}{4+20+20}\right)^{(300-250)0.005} \\ &= 0.86\end{aligned}$$

10.14.6 Point Process Analysis

Since a program can be viewed as a repairable system, with errors being detected and corrected in a time continuum, the method of point process analysis described in Chapter 2 can be applied to software reliability measurement and analysis.

10.15 Hardware/Software Interfaces

In software controlled systems, failures can occur which are difficult to diagnose to hardware or software causes, due to interactions between the two. We have already covered examples of such situations, where hardware elements provide program inputs. The software design can minimize these possibilities, as well as provide automatic diagnosis and fault indication. However, there are other types of failure which are more difficult, particularly when the hardware/software interface is less clearly defined.

Hardware meets software most closely within electronic devices such as processors and memories. A failure of a memory device, say of an individual memory cell which always indicates a logic state 1 (i.e. stuck at 1) regardless of the input, can cause failures which appear to be due to software errors. If the program is known to work under the input conditions, electronic fault-finding techniques can be used to trace the faulty device. There are times, particularly during development, when the diagnosis is not clear-cut, and the software and hardware both need to be checked. Timing errors, either due to device faults or software errors can also lead to this situation.

Memory devices of all types, whether optical or magnetic media or semiconductor memory devices, can cause system failures. Memory media and devices belong to the class of equipment sometimes called ‘firmware’, to indicate their interface status. Since many memory media are dynamic, that is the same data are handled in different locations at different times during program execution, firmware failures can lead to system failures which occur only under certain operating conditions, thus appearing to be due to software errors. Such failures can also be intermittent. Software and memory or microprocessor devices can be designed to protect the system against such failures. For example, the redundancy techniques described above could provide protection against some types of dynamic memory failure (other than a catastrophic failure of, say, a complete memory device). Redundancy can be provided to data or logic held in memory by arranging for redundant memory within the operating store or by providing independent, parallel memory devices. The program logic then has to be designed to store and access the redundant memory correctly, so the program becomes more complex.

10.16 Conclusions

The versatility and economy offered by software control can lead to an under-estimation of the difficulty and cost of program generation. It is relatively easy to write a program to perform a simple defined function. To ensure that the program will operate satisfactorily under all conditions that might exist, and which will be capable of being changed or corrected easily when necessary, requires an effort greater than that required for the basic design and first-program preparation. Careful groundwork of checking the specification, planning the program structure and assessing the design against the specification is essential, or the resulting program will contain many errors and will be difficult to correct. The cost and effort of debugging a large, unstructured program containing many errors can be so high that it is cheaper to scrap the whole program and start again.

Software that is reliable from the beginning will be cheaper and quicker to develop, so the emphasis must always be to minimize the possibilities of early errors and to eliminate errors before proceeding to the next phase. The essential elements of a software development project to ensure a reliable product are:

1 Specify the requirements completely and in detail (system, software).
2 Make sure that all project staff understand the requirements.
3 Check the specifications thoroughly. Keep asking ‘what if . . .?’
4 Design a structured program and specify each module fully.
5 Check the design and the module specifications thoroughly against the system specifications.

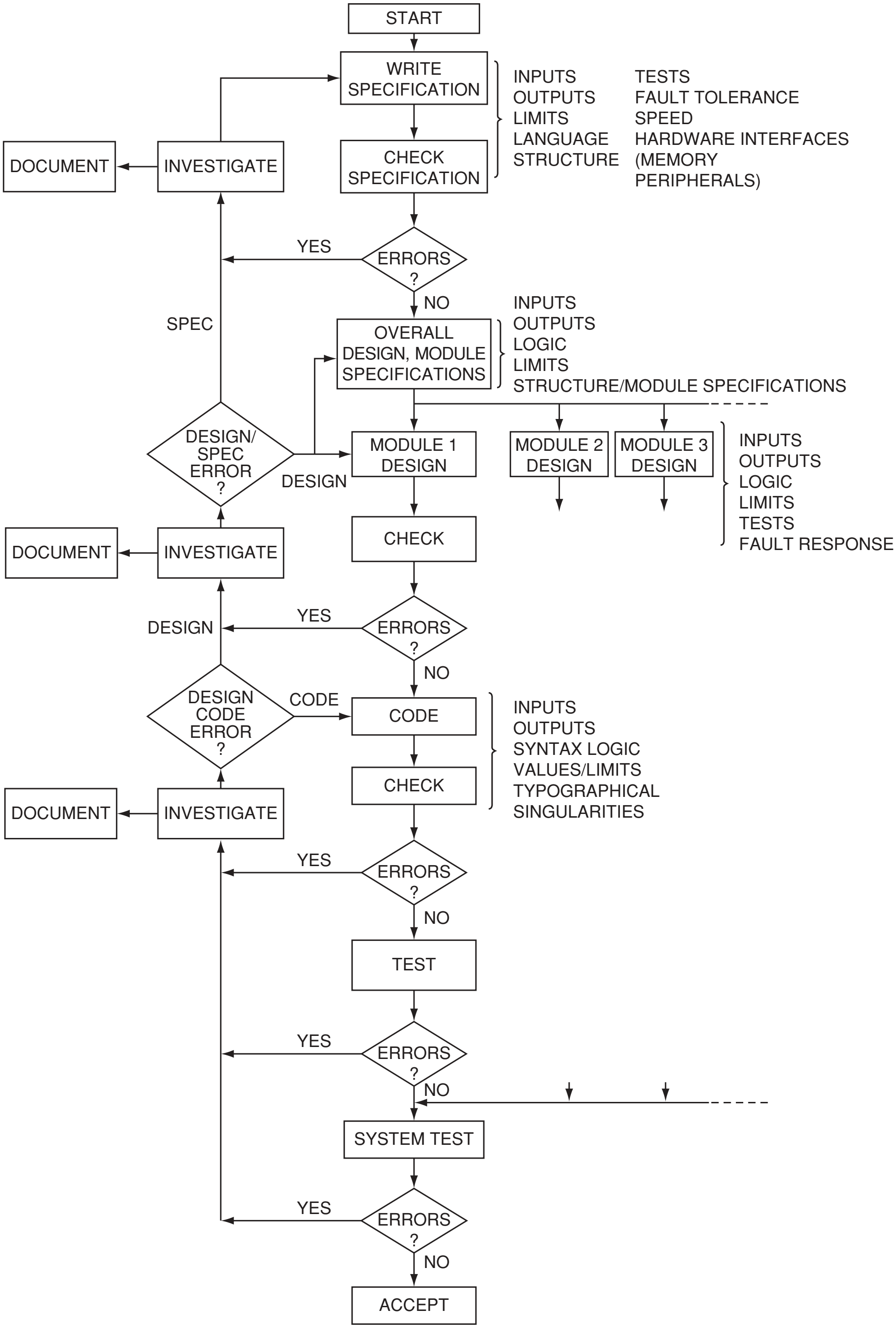


Figure 10.6 *Software development for reliability.*

6 Check written programs for errors, line by line.
7 Plan module and system tests to cover important input combinations, particularly at extreme values.
8 Ensure full recording of all development notes, tests, checks, errors and program changes.

Figure 10.6 shows the sequence of development activities for a software project, with the emphasis on reliability. Musa (2004) and Leveson (1995) provide excellent overviews of software reliability and safety engineering. ISO/IEC90003.2004 provides guidelines for the generation of software within the ISO9001 quality system. US DoD Standard 2168 describes quality requirements for the development of software for military projects.

Questions

1. Discuss the main differences between the ways in which software and hardware can fail to perform as required. Give four examples to illustrate these differences.
2. Explain the ways in which software in engineering systems can be different to other kinds of system.
3. What are the three principal stages in software development that can lead to errors in programs? Give one example of the type of software error that can be created in each stage.
4. What is structured and modular design in the context of software? Describe the main advantages and disadvantages of these approaches.
5. How can software be used to protect against hardware failures in systems that embody both? Give two examples of how software can be used to provide such protection.
6. Describe the essential points to be considered in setting up a test programme for newly developed software. Include the distinction between verification and validation.
7. Several methods have been postulated for predicting and measuring the reliability of software. What are the two main categories of software reliability model? Briefly describe one model in each category, and discuss their main assumptions in relation to predicting the reliability of a new program.

Bibliography

Beizer, B. (1995) *Black-Box Testing*, J. Wiley and Sons.

Kaner, C., Falk, J. and Nguyen, H. (1999) *Testing Computer Software*, 2nd edn, Wiley.

Leveson, N. (1995) *Safeware – System Safety & Computers*, Addison Wesley.

Musa, J. (2004) *Software Reliability Engineering*: *more reliable software faster and cheaper (2nd edn.) Authorhouse*.

Musa, J., Iannino, A. and Okumoto, K. (1987) *Software Reliability Prediction and Measurement*, McGraw-Hill.

Ould, M.A. and Unwin, C. (1986) *Testing in Software Development*, Cambridge University Press.

Patton, R (2006) *Software Testing*, 2nd edn, SAMS Publishing.

US DoD Standard 2168 *Defence System Software Quality Program.* Available from the National Technical Information Service, Springfield, Virginia.

ISO/IEC 90003. (2004) *Guidelines for the Application of ISO 9001 to Computer Software*.

11

Design of Experiments and Analysis of Variance

11.1 Introduction

Product testing is a common part of reliability practitioner's work, which may also involve experimenting intended to improve the product design or some of its characteristics. This chapter deals with the problem of assessing the combined effects of multiple variables on a measurable output or other characteristic of a product, by means of experiments. When designs have to be optimized in relation to variations in parameter values, processes, and environmental conditions, particularly if these variations can have combined effects, we should use methods that can evaluate the effects of the simultaneous variations. For example, it might be necessary to maximize the power output from a generator, and minimize the variation of its output, in relationship to rotational speed, several dimensions, coil geometry, and load conditions. All of these could have single or combined effects which cannot all be easily or accurately computed using theoretical calculations.

Statistical methods of experimentation have been developed which enable the effects of variation to be evaluated in these types of situations. They are applicable whenever the effects cannot be easily theoretically evaluated, particularly when there is a large component of random variation or interactions between variables. For situations when multiple variables might affect an output, the methods are much more economical than performing separate experiments to evaluate the effect of one variable at a time. This 'traditional' approach also does not enable interactions to be analysed, when these are not known empirically. The rest of this chapter describes the statistical experimental methods, and how they can be adapted and applied to optimization and problem-solving in engineering.

11.2 Statistical Design of Experiments and Analysis of Variance

The statistical approach to design of experiments (DOE) and the analysis of variance (ANOVA) technique was developed by R. A. Fisher, and is a very elegant, economical and powerful method for determining the significant effects and interactions in multivariable situations. Analysis of variance is used widely in such

Practical Reliability Engineering, Fifth Edition. Patrick D. T. O'Connor and Andre Kleyner.

fields as market research, optimization of chemical and metallurgical processes, agriculture and medical research. It can provide the insights necessary for optimizing product designs and for preventing and solving quality and reliability problems. Whilst the methods described below might appear tedious, there is a variety of commercially available software packages designed to analyse the results of statistical experiments. Minitab® is a widely utilized package, which will be used for illustration in this chapter.[1]

11.2.1 Analysis of Single Variables

The variance of a set of data (sample) is equal to

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

(Eq. 2.15), where n is the sample size and $\bar{x}$ is the mean value. The population variance estimate is derived by dividing the sum of squares, $\Sigma(x_i - \bar{x})^2$, not by n but by $(n - 1)$, where $(n - 1)$ denotes the number of degrees of freedom (DF). Then $\hat{\sigma} = \Sigma(x_i - \bar{x})^2/(n - 1)$ (Eq. 2.16).

Example 11.1

To show how the variance of a group of samples can be analysed, consider a simple experiment in which 20 bearings, five each from four different suppliers, are run to failure. Table 11.1 shows the results.

We need to know if the observed variation between the samples is statistically significant or is only a reflection of the variations of the populations from which the samples were drawn. Within each sample of five there are quite large variations. We must therefore analyse the difference between the 'between sample' (BS) and the 'within sample' (WS) variance and relate this to the populations.

The next step is to calculate the sample totals and sample means, as shown in Table 11.1. The overall average value

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{120}{20} = 6$$

Since n is 20, the total number of degrees of freedom (DF) is 19. We then calculate the values of $(x_i - \bar{x})^2$. These are shown in Table 11.2.

Table 11.1 *Times to failure of 20 bearings.*

Sample	Times to failure, $^a x_i$ (h)					Sample totals	Sample means, x_i'
1	4	1	3	5	7	20	4
2	6	6	5	10	3	30	6
3	3	2	5	7	8	25	5
4	7	8	8	12	10	45	9

$^a\Sigma x_i = 120$.

[1] Portions of the input and output contained in this publication/book are printed with permission of Minitab Inc. All material remains the exclusive property and copyright of Minitab Inc. All rights reserved.

Table 11.2 Values of $(x_i - \bar{x})^2$ for the data of Table 11.1.

Sample	$(x_i - \bar{x})^2$					$\Sigma(x_i - \bar{x})^{2a}$
1	4	25	9	1	1	40
2	0	0	1	16	9	26
3	9	16	1	1	4	31
4	1	4	4	36	16	61

[a] Overall $\Sigma(x_i - \bar{x})^2 = 158$.

Table 11.3 Values of $(x_i - x_i')$ for the data of Table 11.1.

Sample	x_i''					WS$\Sigma x_i''$
1	0	−3	−1	1	3	0
2	0	0	−1	4	−3	0
3	−2	−3	0	2	3	0
4	−2	−1	−1	3	1	0

Table 11.4 Values of WS $(x_i'' - \bar{x}'')^2$ for the data of Table 11.1.

Sample	WS $(x_i'' - \bar{x}'')^2$					WS $\Sigma(x_i'' - \bar{x}'')^{2\ a}$
1	0	9	1	1	9	20
2	0	0	1	16	9	26
3	4	9	0	4	9	26
4	4	1	1	9	1	16

[a] Overall $\Sigma(x_i'' - \bar{x})^2 = 88$.

Having derived the total sum of squares and the DF, we must derive the WS and BS values. To derive the BS effect, we assume that each item value is equal to its sample mean (x_i'). The sample means for each item in samples 1 to 4 are 4, 6, 5 and 9, respectively. The BS sums of squares $(x_i' - \bar{x})^2$ are then, for each item in samples 1 to 4, 4, 0, 1 and 9, respectively, giving sample totals $\Sigma(x_i' - \bar{x})^2$ of 20, 0, 5 and 45 and an overall BS $\Sigma(x_i' - \bar{x})$ of 70. The BS DF is 4 – 1 = 3.

Now we derive the equivalent values for the WS variance, by removing the BS effect. We achieve this by subtracting x_i' from each item in the original table to give a value x_i''. The result is shown in Table 11.3.

Table 11.4 gives the values of the WS sums of squares, derived by squaring the values as they stand, since now $\bar{x}'' = 0$. The number of WS DF is $(5 - 1) \times 4 = 16$ (4 DF within each sample for a total of four samples). We can now tabulate the analysis of variance (Table 11.5).

Table 11.5 Sources of variance for the data in Table 11.1.

Source of variance	$\Sigma(.)$	DF	σ^2
BS	70	3	23.33
WS (residual)	88	16	5.50
Total	158	19	8.32

Table 11.6 *Example 11.1 Minitab® solution.*

One-way ANOVA: Time versus Supplier

Source	DF	SS	MS	F	P
Supplier	3	70.00	23.33	4.24	0.022
Error	16	88.00	5.50		
Total	19	158.00			

If we can assume that the variables are normally distributed and that all the variances are equal, we can use the F-test (variance ratio test) (Section 2.12.4) to test the null hypothesis that the two variance estimates (BS and WS) are estimates of the same common (population) variance. The WS variance represents the experimental error, or *residual* variance. F is the ratio of the variances, and F-values for various significance levels and degrees of freedom are well tabulated and can easily be found on the Internet (see, e.g. NIST, 2011). F-values can also be calculated using Excel function FINV.

In this case

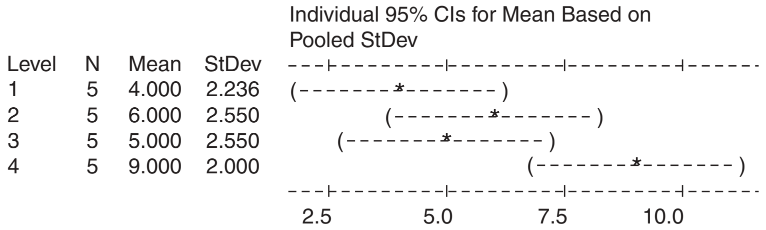
$$F = \frac{23.33}{5.50} = 4.24$$

For 3 DF in the greater variance estimate and 16 DF in the smaller variance estimate, the 5 % significance level of F is 3.239 = FINV(0.05, 3, 16). Since our value of F is greater than this, we conclude that the variance ratio is significant at the 5 % level, and the null hypothesis that there is no difference between the samples is therefore rejected at this level.

A solution of Example 11.1 using the Minitab® software would provide additional analysis options, such as that shown in Table 11.6 and Figure 11.1.

In addition to the degrees of freedom, sums of squares, and variances already shown in Table 11.5 it calculates the F-value and its significance level (P-column). In this case it is 1-P = 1 – 0.022 = 0.978 which is greater than the sought significance of 95 %.

Figure 11.1 gives a graphical representation of the data showing the mean values and their 95 % confidence intervals. For example, it shows that 95 % confidence bounds for sample 1 and sample 4 (column 1) do not overlap, graphically illustrating the statistical significance of that difference.



```
                               Individual 95% CIs for Mean Based on
                               Pooled StDev
Level  N   Mean   StDev  ---+---------+---------+---------+------
1      5  4.000   2.236  (-------*-------)
2      5  6.000   2.550          (-------*-------)
3      5  5.000   2.550      (-------*-------)
4      5  9.000   2.000                     (-------*-------)
                         ---+---------+---------+---------+------
                           2.5       5.0       7.5      10.0
```

Figure 11.1 *Minitab® chart showing statistically significant difference between samples 1 and 4.*

11.2.2 Analysis of Multiple Variables (Factorial Experiments)

The method described above can be extended to analyse more than one source of variance. When there is more than one source of variance, interactions may occur between them, and the interactions may be more significant than the individual sources of variance. The following example will illustrate this for a three-factor situation.

Example 11.2

On a hydraulic system using ‘O’ ring seals, leaks occur apparently randomly. Three manufacturers’ seals are used, and hydraulic oil pressure and temperature vary through the system. Argument rages as to whether the high temperature locations are the source of the problem, or the manufacturer, or even pressure. Life test data show that seal life may be assumed to be normally distributed at these stress levels, with variances which do not change with stress. A test rig is therefore designed to determine the effect on seal performance of oil temperature, oil pressure and type of seal. We set up the experiment in which we apply two values of pressure, 15 and 18 MN m^{-2} (denoted p_1 and p_2, in increasing order), and three values of temperature, 80, 100 and 120 °C (t_1, t_2 and t_3, also in increasing order). We then select seals for application to the different test conditions, with two tests at each test combination. The results of the test are shown in Table 11.7, showing the elapsed time (h) to a detectable leak.

Table 11.7 Results of experiments on ‘O’ ring seals.

	Type (T) 1		T2		T3	
Temperature	p_1	p_2	p_1	p_2	p_1	p_2
	104	209	181	172	157	178
t_1	196	132	129	151	187	211
	136	140	162	133	141	164
t_2	97	122	108	114	174	128
	108	96	99	112	122	135
t_3	121	110	123	109	130	118

To simplify Table 11.7, subtract 100 from each value. The coded data then appear as in Table 11.8.

The sums of squares for the three sources of variation are then derived as follows:

1 The correction factor:

$$\mathrm{CF} = \frac{(\Sigma x_i)^2}{n} = \frac{(1409)^2}{36} = 55\,147$$

Table 11.8 The data of Table 11.7 after subtracting 100 from each datum.

	Type (T) 1		T2		T3		
Temperature	p_1	p_2	p_1	p_2	p_1	p_2	
	4	109	81	72	57	78	
t_1	96	32	29	51	87	111	
	36	40	62	33	41	64	$\Sigma x_i(t)$
t_2	−3	22	8	14	74	28	
	8	−4	−1	12	22	35	
T_3	21	10	23	9	30	18	

2 The between types sum of squares:

$$\begin{aligned}\mathrm{SS_T} &= \frac{\Sigma x_i^2(\mathrm{T})}{2 \times 3 \times 2} - \mathrm{CF}\\ &= \frac{371^2 + 393^2 + 645^2}{12} - 55\,147 = 3863\end{aligned}$$

Between three types there are two DF.

3 The between pressures sum of squares:

$$\begin{aligned}\mathrm{SS}_p &= \frac{\Sigma x_i^2(p)}{3 \times 3 \times 2} - \mathrm{CF}\\ &= \frac{675^2 + 734^2}{18} - 55\,147 = 96\end{aligned}$$

Between two pressures there is one DF.

4 The between temperatures sum of squares:

$$\begin{aligned}\mathrm{SS}_t &= \frac{\Sigma x_i^2(t)}{2 \times 3 \times 2} - \mathrm{CF}\\ &= \frac{807^2 + 419^2 + 183^2}{12} - 55\,147 = 16\,544\end{aligned}$$

Between three temperatures there are two DF.

5 The types–temperature interaction sum of squares:

$$\begin{aligned}\mathrm{SS}_{\mathrm{T}t} &= \frac{\Sigma x_i^2(\mathrm{T}t)}{2 \times 2} - \mathrm{CF} - \mathrm{SS_T} - \mathrm{SS}_t\\ &= \frac{241^2 + 95^2 + 35^2 + 233^2 + 117^2 + 43^2 + 333^2 + 207^2 + 105^2}{4}\\ &- 55\,147 - 3863 - 16\,544\\ &= 176\end{aligned}$$

The T and the t effects each have 2 DF, so the interaction has 2 × 2 = 4 DF.

6 The types–pressure interaction sum of squares:

$$\begin{aligned}\mathrm{SS}_{\mathrm{T}p} &= \frac{\Sigma x_i^2(\mathrm{T}_p)}{6} - \mathrm{CF} - \mathrm{SS_T} - \mathrm{SS}_p\\ &= \frac{162^2 + 209^2 + 202^2 + 191^2 + 311^2 + 334^2}{2 \times 3} - 55\,147 - 3863 - 96\\ &= 142\end{aligned}$$

The T effect has 2 DF and the p effect has 1 DF, so the Tp interaction has 2 × 1 = 2 DF.

7 The temperature–pressure interaction sum of squares:

$$\begin{aligned}\mathrm{SS}_{tp} &= \frac{\Sigma x_i^2(tp)}{2\times 3} - \mathrm{CF} - \mathrm{SS}_t - \mathrm{SS}_p\\ &= \frac{354^2+453^2+218^2+201^2+103^2+80^2}{6} - 55\,147 - 16\,544 - 96\\ &= 789\end{aligned}$$

The t effect has 2 DF and the p effect 1 DF. Therefore the tp interaction has 2 × 1 = 2 DF.

8 The type–temperature–pressure interaction sum of squares:

$$\begin{aligned}\mathrm{SS}_{\mathrm{T}tp} &= \frac{\Sigma x_i^2(\mathrm{T}tp)}{2} - \mathrm{CF} - \mathrm{SS}_\mathrm{T} - \mathrm{SS}_t - \mathrm{SS}_p - \mathrm{SS}_{\mathrm{T}t} - \mathrm{SS}_{tp} - \mathrm{SS}_{\mathrm{T}p}\\ &= \frac{154\,613}{2} - 55\,147 - 3863 - 16\,544 - 96 - 176 - 789 - 142\\ &= 550\end{aligned}$$

There are 2 × 1 × 2 = 4 DF.

9 The total sum of squares:

$$\begin{aligned}\mathrm{SS}_{\mathrm{tot}} &= \Sigma x_i^2 - \mathrm{CF}\\ &= 91\,475 - 55\,147 = 36\,328\end{aligned}$$

Total DF = 36 – 1 = 35 DF.

10 Residual (experimental error) sum of squares is

$$\begin{aligned}\mathrm{SS}_{\mathrm{tot}} - (\text{all other SS})\\ &= 36\,328 - 3863 - 96 - 16\,544 - 376 - 142 - 789 - 550\\ &= 13\,988\end{aligned}$$

The residual DF:

Total DF – (all other DF) = 35 -2 -1 -2 -4 -2 - 2 -4 = 18 DF

Examination of the analysis of variance table (Table 11.9) shows that all the interactions show variance estimates much less than the residual variance, and therefore they are clearly not statistically significant.

Having determined that none of the interactions are significant, we can assume that these variations are also due to the residual or experimental variance. We can therefore combine these sums of squares and degrees of freedom to provide a better estimate of the residual variance. The revised residual variance is thus:

$$\begin{aligned}\frac{176+142+789+550+13988}{4+2+2+4+18} &= \frac{15\,645}{30}\\ &= 521 \text{ with } 30\,\mathrm{DF}\end{aligned}$$

Table 11.9 *Analysis of variance table.*

				Interactions				
	Main factors			First order			Second order	
Effect	Type (T)	Pressure (*p*)	Temperature(*t*)	T*t*	T*p*	*tp*	T*tp*	Residual
SS	3863	96	16 544	176	142	789	550	13 920
DF	2	1	2	4	2	2	4	18
SS/DF[a]	1932	96	8272	44	71	395	138	773
SD	44	10	91	10	8	20	9	28

[a] Variance estimate.

We can now test the significance of the main factors. Clearly the effect of pressure is not statistically significant. For the ‘type’ (T) main effect, the variance ratio F =1932/521 = 3.71 with 2 and 30 DF. Using Excel® function for F-distribution FDIST(3.71, 2, 30) = 0.0363, which shows that this is significant at the 5 % level (3.63 %). For the temperature (t) main effect, F = 8272/521 = 15.88 with 2 and 30 DF. This is significant, even at the 1 % level (FDIST(15.88, 2, 30) = 1.98×10^{-5}).

Therefore the experiment shows that the life of the seals is significantly dependent upon operating temperature and upon type. Pressure and interactions show effects which, if important, are not discernible within the experimental error; the effects are ‘lost in the noise’. In other words, no type is significantly better or worse at higher pressures. Referring back to the original table of results, we can calculate the mean lives of the three types of seal, under the range of test conditions: type 1, 130 h; type 2, 132 h; and type 3, 154 h. Assuming that no other aspects such as cost predominate, we should therefore select type 3, which we should attempt to operate at low oil temperatures.

11.2.3 Non-Normally Distributed Variables

In Example 11.2 above it is important to note that the method described is statistically correct only for normally distributed variables. As shown in Chapter 2, in accordance with the central limit theorem many parameters in engineering are normally distributed. However, it is prudent to test variables for normality before performing an analysis of variance. If any of the key variables are substantially non-normally distributed, non-parametric analysis methods can be used, the data can be converted into normally distributed values, or other statistical methods can be applied. For more details on dealing with non-normal data see Deshpande (1995).

11.2.4 Two-Level Factorial Experiments

It is possible to simplify the analysis of variance method if we adopt a two-level factorial design for the experiment. In this approach, we take only two values for each main effect, high and low, denoted by + and −. Example 11.3 below shows the results of a three-factor non-replicated experiment (such an experiment is called a 2^3 factorial design, i.e. three factors, each at two levels).

Example 11.3

From the results of Table 11.10, the first three columns of Table 11.11 represent the *design matrix* of the factorial experiment. The response value is the mean of the values at each test combination or in this case, as there is no replication, the single test value.

Table 11.10 Results of a three-factor non-replicated experiment.

A			B	
	−	−	+	+
			C	
	−	+	−	+
−	12	11	12	13
+	14	14	13	15

The main effects can be simply calculated by averaging the difference between the response values for each high and low factor setting, using the appropriate signs in the ‘factor’ columns. Thus:

$$\text{Main effect A} = [(15-13)+(13-12)+(14-11)+(14-12)]/4 = 2$$
$$\text{This is the same as: } (15+13+14+14-13-12-11-12)/4 = 2$$

Likewise,

$$\text{Main effect B} = (15+13-14-14+13+12-11-12)/4 = 0.5$$

and,

$$\text{Main effect C} = (15-13+14-14+13-12+11-12)/4 = 0.5$$

The sum of squares of each effect is then calculated using the formula

$$\text{SS} = 2^{k-2} \times (\text{effect estimate})^2$$

where k is the number of factors. Therefore,

$$\text{SS}_\text{A} = 2 \times 2^2 = 8$$
$$\text{SS}_\text{B} = 2 \times 0.5^2 = 0.5$$
$$\text{SS}_\text{C} = 2 \times 0.5^2 = 0.5$$

Table 11.11 Response table and interaction of effects A, B, C.

Factor				Interactions				
A	B	C		AB	AC	BC	ABC	Response
+	+	+	+	+	+	+		15
+	+	−	+	−	−	−		13
+	−	+		−	+	−	−	14
+	−	−	−	−	+	+		14
−	+	+		−	−	+	−	13
−	+	−	−	+	−	+		12
−	−	+	+	−	−	+		11
−	−	−		+	+	+	−	12

Table 11.12 *Analysis of variance table.*

	Effect							
	A	B	C	AB	AC	BC	ABC	*Total*
SS	8	0.5	0.5	0.5	0.5	2	0	12
DF	1	1	1	1	1	1	1	7
SS/DF[a]	8	0.5	0.5	0.5	0.5	2	0	

[a] Variance estimate.

We can derive the interaction effects by expanding Table 11.11. Additional columns are added, one for each interaction. The signs under each interaction column are derived by algebraic multiplication of the signs of the constituent main effects. The ABC interaction signs are derived from AB × C or AC × B or BC × A.

The AB interaction is then

$$(15 + 13 - 14 - 14 - 13 - 12 + 11 + 12)/4 = -0.5$$

and

$$\mathrm{SS_{AB}} = 2 \times (-0.5)^2 = 0.5$$

The other interaction sums of squares can be calculated in the same way and the analysis of variance table (Table 11.12) constructed.

To illustrate the additional capability of the software analysis, the quick analysis of the data in Example 11.3 with Minitab® would quickly produce the main effects diagram (Figure 11.2). Judging by the slopes of the lines we can easily notice the largest A main effect and lowest effect of ABC interaction.

Whilst the experiment in Example 11.3 indicates a large A main effect, and possibly an important BC interaction effect, it is not possible to test these statistically since no residual variance is available. To obtain a value of residual variance further replication would be necessary, or a value of experimental error might be available from other experiments. Alternatively, if the interactions, particularly high order interactions, are insignificant, they can be combined to give a residual value. Also, with so few degrees of freedom, the F test requires a very large value of the variance ratio in order to give high confidence that the effect is significant. Therefore, an unreplicated 2^3 factorial experiment may not always be sufficiently sensitive. However, the example will serve as an introduction to the next section, dealing with situations where several variables need to be considered.

11.2.5 Fractional Factorial Experiments

So far we have considered experiments in which all combinations of factors were tested, that is *full factorial* experiments. These can be expensive and time-consuming, since if the number of factors to be tested is f, the number of levels is L and the number of replications is r, then the number of tests to be performed is rL^f, that is in the hydraulic seal example $2 \times 3^2 = 18$, or in a three-level four-factor experiment with two replications $2 \times 3^4 = 162$. In a *fractional factorial* experiment we economize by eliminating some test combinations. Obviously we then lose information, but if the experiment is planned so that only those effects

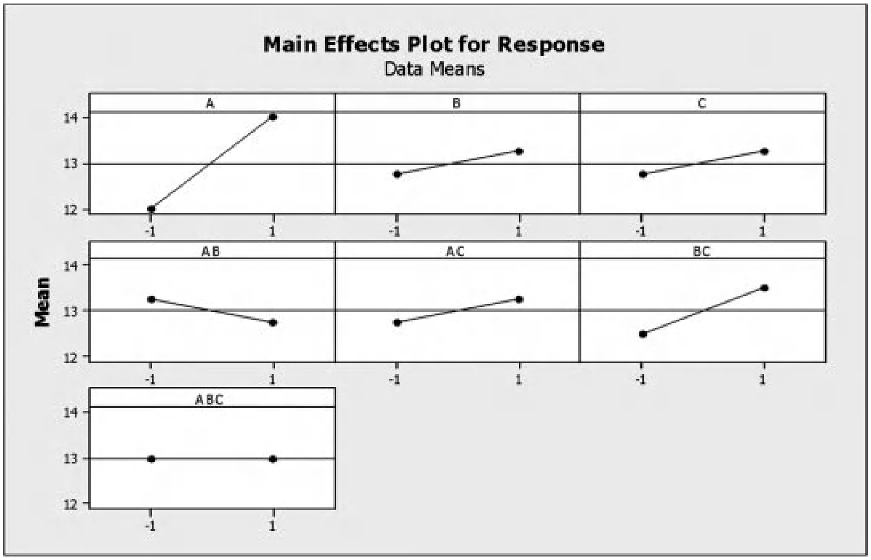


Figure 11.2 Main effects plots for Example 11.3 using Minitab®.

and interactions which are already believed to be unimportant are eliminated, we can make a compromise between total information, or experiment costs, and experimental value.

A point to remember is that higher order interactions are unlikely to have any engineering meaning or to show statistical significance and therefore the full factorial experiment with several factors can give us information not all of which is meaningful.

We can design a fractional factorial experiment in different ways, depending on which effects and interactions we wish to analyse. Selection of the appropriate design is made starting from the full factorial design matrix. Table 11.13 gives the full design matrix for a 2^4 factorial experiment.

We select those interactions which we do not consider worth analysing. In the 2^4 experiment, for example, the ABCD interaction would not normally be considered statistically significant. We therefore omit all the rows in the table in which the ABCD column shows – Thus we eliminate half of our test combinations, leaving a *half factorial* experiment. What else do we lose as a result? Table 11.14 shows what is left of the experiment.

Examination of this table shows the following pairs of identical columns

A, BCD; B, ACD; C, ABD; D, ABC; AB, CD; AC, BD: AD, BC

This means that the A main effect and the BCD interaction effect will be indistinguishable in the results. In fact in any experiment to this design we will not be able to distinguish response values for these effects; we say that they are *aliased* or *confounded*. If we can assume that the first- and second-order interactions aliased in this case are insignificant, then this will be an appropriate fractional design, reducing the number of tests from $2^4 = 16$ to $\frac{1}{2} \times 2^4 = 8$. We will still be able to analyse all the main effects, and up to three first-order interactions if we considered that they were likely to be significant. For example, if engineering knowledge

Table 11.13 The full design matrix for a 2^4 factorial experiment.

Test	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
no.	A	B	C	D	AB	AC	AD	BC	BD	CD	ABC	ABD	ACD	BCD	ABCD
1	−	−	−	−	+	+	+	+	+	+	−	−	−	−	+
2	−	−	−	+	+	+	−	+	−	−	−	+	+	+	−
3	−	−	+	−	+	−	+	−	+	−	+	−	+	+	−
4	−	−	+	+	+	−	−	−	−	+	+	+	−	−	+
5	−	+	−	−	+	+	+	−	−	+	+	+	−	+	−
6	−	+	−	+	−	+	−	−	+	−	+	−	+	−	+
7	−	+	+	−	−	−	+	+	−	−	−	+	+	−	+
8	−	+	+	+	−	−	−	+	+	+	−	−	−	+	−
9	+	−	−	−	−	−	−	+	+	+	+	+	+	−	−
10	+	−	−	+	−	−	+	+	−	−	+	−	−	+	+
11	+	−	+	−	−	+	−	−	+	−	−	+	−	+	+
12	+	−	+	+	−	+	+	−	−	+	−	−	+	−	−
13	+	+	−	−	+	−	−	−	−	+	−	−	+	+	+
14	+	+	−	+	+	−	+	−	+	−	−	+	−	−	−
15	+	+	+	−	+	+	−	+	−	−	+	−	−	−	−
16	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

tells us that the AB interaction is likely to be significant but the CD interaction not, we can attribute the variance estimate to the AB interaction.

A similar breakdown of a 2^3 experiment would show that it is not possible to produce a fractional factorial without aliasing main effects with first-order interactions. This is unlikely to be acceptable, and fractional factorial designs are normally only used when there are four or more factors to be analysed. Quarter factorial designs can be used when appropriate, following similar logic to that described above. The value of using fractional factorial designs increases rapidly when large numbers of effects must be analysed. For example, if there are seven main effects, a full factorial experiment would analyse a large number of high level interactions which would not be meaningful, and would require $2^7 = 128$ tests for no repeats. We can design a sixteenth

Table 11.14 Table 11.13 omitting rows where ABCD gives minus.

A	B	C	D	AB	AC	AD	BC	BD	CD	ABC	ABD	ACD	BCD	ABCD
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
+	−	−	+	−	−	+	+	−	−	+	−	−	+	+
+	−	+	−	−	+	−	−	+	−	−	+	−	+	+
+	+	−	−	+	−	−	−	−	+	−	−	+	+	+
−	−	−	−	+	+	+	+	+	+	−	−	−	−	+
−	+	+	−	−	−	+	+	−	−	−	+	+	−	+
−	+	−	+	−	+	−	−	+	−	+	−	+	−	+
−	−	+	+	+	−	−	−	−	+	+	+	−	−	+

Table 11.15 *Sixteenth fractional factorial layout for seven main effects.*

Test no.	A	B	C	D	E	F	G
1	+	+	+	+	+	+	+
2	+	+	−	+	−	−	−
3	+	−	+	−	+	−	−
4	+	−	−	−	−	+	+
5	−	+	+	−	−	+	−
6	−	+	−	−	+	−	+
7	−	−	+	+	−	−	+
8	−	−	−	+	+	+	−
Alias				AB	AC	BC	ABC

fractional factorial layout with only eight tests, which will analyse all the main effects and most of the first-order interactions, as shown in Table 11.15. The aliases are deliberately planned, for example by evaluating the signs for the D column by multiplying the signs for A and B, and so on. We can select which interactions to alias by engineering judgement. It should be noted that other effects are also aliased. The full list of aliased effects can be derived by multiplying the aliased effects. For example, if D and AB are aliased, then the ABD interaction effect is also aliased, and so on. (If a squared term arises, let the squared term equal unity, e.g. $AB \times AC = A^2BC = BC$.)

There are various methods of constructing fractional factorial experiments to reduce the number of factors and combinations. Detailed coverage of those techniques is beyond the scope of this chapter, therefore for more information refer to Hinkelmann and Kempthorne (2008), Mathews (2005) or other references at the end of this chapter.

11.3 Randomizing the Data

At this stage it is necessary to point out an essential aspect of any statistically designed experiment. Significant sources of variance must be made to show their presence not only against the background of experimental error but against other sources of variation which might exist but which might not be tested for in the experiment. For instance, in Example 11.2, a source of variation might be the order in which seals are tested, or the batch from which seals are drawn. To eliminate the effects of extraneous factors and to ensure that only the effects being analysed will affect the results, it is important that the experiment is randomized. Thus the items selected for test and the sequence of tests must be selected at random, using random number tables or another suitable randomizing process. In Example 11.2, test samples should have been drawn at random from several batches of seals of each type and the test sequences should also have been randomized. It is also very important to eliminate human bias from the experiment, by hiding the identity of the items under test, if practicable.

If the items under test undergo a sequence of processes, for example heat treatment, followed by machining, then plating, the items should undergo each process in random order, that is separately randomized for each process.

Because of the importance of randomizing the data, it is nearly always necessary to design and plan an experiment to provide the data for analysis of variance. Data collected from a process as it normally occurs is unlikely to be valid for this purpose. Careful planning is important, so that once the experiment starts unforeseen circumstances do not cause disruption of the plan or introduce unwanted sources of variation or bias. A dummy run can be useful to confirm that the experiment can be run as planned.

11.4 Engineering Interpretation of Results

A statistical experiment can always, by its nature, produce results which conflict with the physical or chemical basis of the situation. The probability of a variance estimate being statistically significant in relation to the experimental error is determined in the analysis, but we must always be on the lookout for the occurrence of chance results which do not fit our knowledge of the processes being studied. For example, in the hydraulic seal experiment we could study further the temperature–pressure interaction. Since the variance estimate for this interaction was higher than for the other interactions we might be tempted to suspect some interaction. In another experiment the seals might be selected in such a way that this variance estimate showed significance.

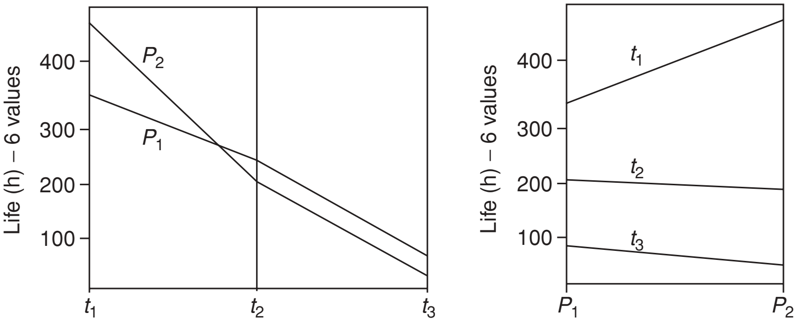


Figure 11.3 *Temperature–pressure interactions.*

Figure 11.3 shows the interaction graphically, in two forms. If there were no interaction, that is the two effects acted quite independently upon seal life, the lines would be parallel between t_1t_2, t_2t_3 or p_1p_2. In this case an interaction appears to exist between 80(t_1) and 100 °C(t_2), but not between 100 and 120 °C(t_3). This we can dismiss as being unlikely, taking into account the nature of the seals and the range of pressures and temperatures applied. That is not to say that such an interaction would always be dismissed, only that we can legitimately use our engineering knowledge to help interpret the results of a statistical experiment. The right balance must always be struck between the statistical and engineering interpretations. If a result appears highly significant, such as the temperature main effect, then it is conversely highly unlikely that it is a perverse result. If the engineering interpretation clashes with the statistical result and the decision to be made based on the result is important, then it is wise to repeat the experiment, varying the plan to emphasize the effects in question. In the hydraulic seal experiment, for example, we might perform another experiment, using type 3 seals only, but at three values of pressure as well as three values of temperature, and making four replications of each test instead of only two.

11.5 The Taguchi Method

Genichi Taguchi (1986), developed a framework for statistical design of experiments adapted to the particular requirements of engineering design. Taguchi suggested that the design process consists of three phases: *system* *design, parameter design* and *tolerance design*. In the system design phase the basic concept is decided, using theoretical knowledge and experience to calculate the basic parameter values to provide the performance required. Parameter design involves refining the values so that the performance is optimized in relation to factors and variation which are not under the effective control of the designer, so that the design is 'robust' in relation to these. Tolerance design is the final stage, in which the effects of random variation of manufacturing processes and environments are evaluated, to determine whether the design of the product and the production processes can be further optimized, particularly in relation to cost of the product and the production processes. Note that the design process is considered to explicitly include the design of the production methods and their control. Parameter and tolerance design are based on statistical design of experiments.

Taguchi separates variables into two types. *Control factors* are those variables which can be practically and economically controlled, such as a controllable dimensional or electrical parameter. *Noise factors* are the variables which are difficult or expensive to control in practice, though they can be controlled in an experiment, for example ambient temperature, or parameter variation within a tolerance range. The objective is then to determine the combination of control factor settings (design and process variables) which will make the product have the maximum 'robustness' to the expected variation in the noise factors. The measure of robustness is the *signal-to-noise* ratio, which is analogous to the term as used in control engineering.

Figure 11.4 illustrates the approach. This shows the response of an output parameter to a variable. This could be the operating characteristic of a transistor or of a hydraulic valve, for example. If the desired output

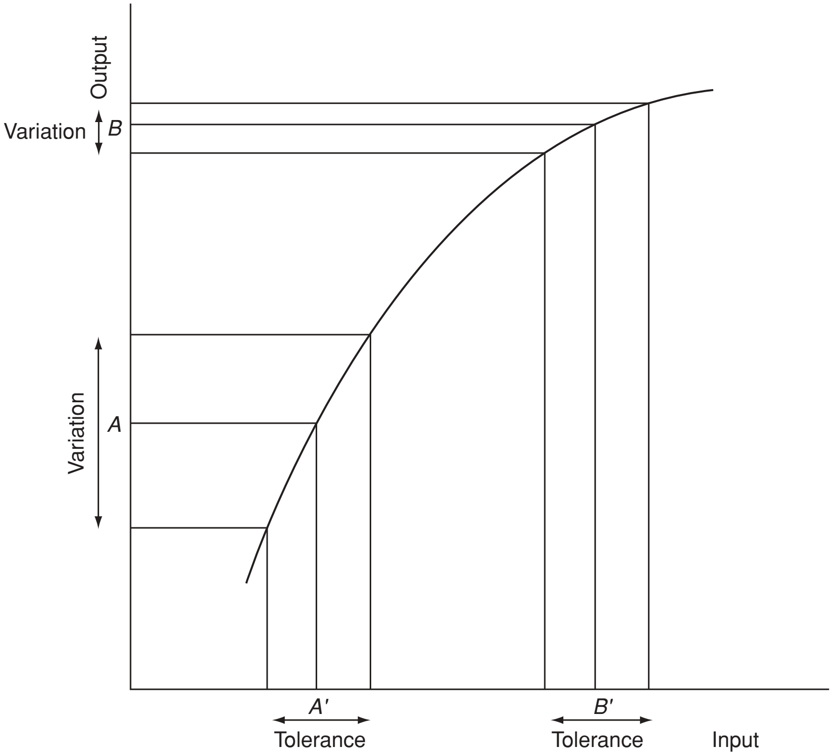


Figure 11.4 *Taguchi method (1).*

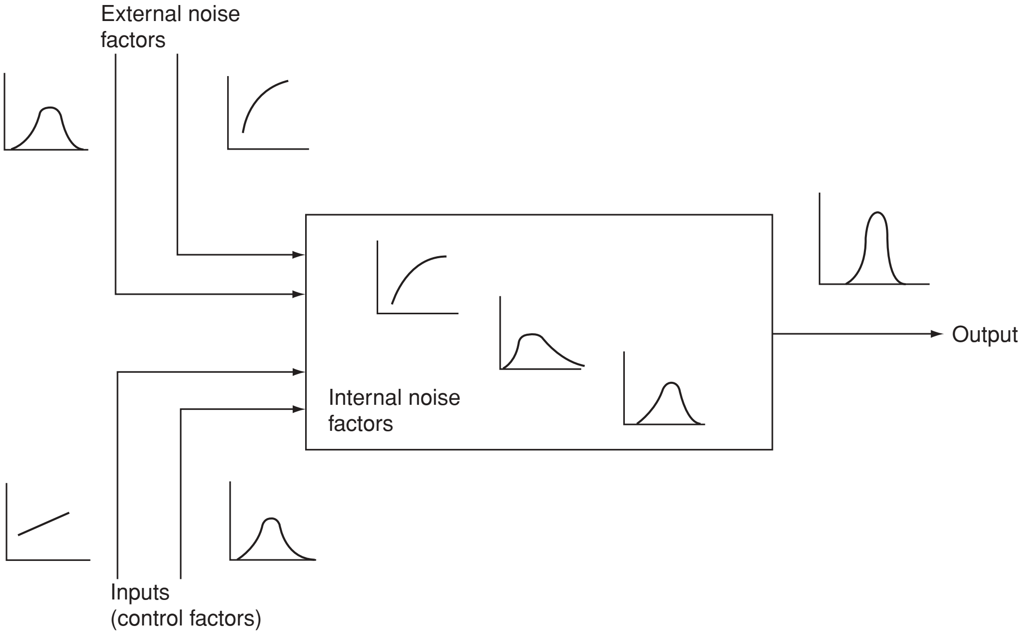


Figure 11.5 *Taguchi method (2).*

parameter value is A, setting the input parameter at A', with the tolerance shown, will result in an output centred on A, with variation as shown. However, the design would be much better, that is more robust to variation of the input parameter, if this were centred at B', since the output would be much less variable with the same variation of the input parameter. The fact that the output value is now too high can be adjusted by adding another component to the system, with a linear or other less sensitive form of operating characteristic. This is a simple case for illustration, involving only one variable and its effect. For a multi-dimensional picture, with relationships which are not known empirically, the statistical experimental approach must be used.

Figure 11.5 illustrates the concept when multiple variations, control and noise factors affect the output of interest. This shows the design as a control system, whose performance must be optimized in relation to the effect of all variations and interactions.

The experimental framework is as described earlier, using fractional factorial designs. Taguchi argued that, in most engineering situations, interactions do not have significant effects, so that much reduced, and therefore more economical, fractional factorial designs can be applied. When necessary, subsidiary or confirmatory experiments can be run to ensure that this assumption is correct. Taguchi developed a range of such design matrices, or *orthogonal arrays*, from which the appropriate one for a particular experiment can be selected. For example, the 'L8' array is a sixteenth fractional factorial design for seven variables, each at two levels, as shown in Table 11.14. (The 'L' refers to the Latin square derivation.) Further orthogonal arrays are given in Taguchi (1986), Ross (1988), Condra (1993) and Roy (2001).

The arrays can be combined, to give an inner and an outer array, as shown in Table 11.16. The inner array contains the control factors, and the outer array the noise factors. The signal-to-noise ratio is calculated for the combination of control factors being considered, using the outer array, the formula depending on

Table 11.16 *Results of Taguchi experiment on fuel system components (Example 11.4).*

					OUTER ARRAY (2 × 2)						
	INNER ARRAY (L4: 3 × 2)			*X*	+	+	−	−			
				Y	+	−	+	−			
	A	B	C		RESPONSE (−30)				mean	$\hat{\sigma}$	*S/N* $(-10 \log \sigma^2)$
1	+	+	+		8	6	4	4	5.5	1.91	−5.63
2	+	−	−		0	0	−2	−4	−1.5	1.91	−5.63
3	−	+	−		0	−2	0	−2	−1.0	1.15	−1.12
4	−	−	+		4	2	4	2	3.0	1.15	−1.12

whether the desired output parameter must be maximized, minimized or centralized. The expressions are as follows:

$$\text{Maximum output, S/N ratio} = -10 \log \left[\frac{\Sigma(1/x^2)}{n} \right]$$

$$\text{Minimum output, S/N ratio} = -\log \left[\frac{\Sigma x^2}{n} \right]$$

$$\text{Centralized output, S/N ratio} = -10 \log \left[\hat{\sigma}^2 \right]$$

where x is the mean response for the range of control factor settings, and $\hat{\sigma}$ is the estimate of the standard deviation. The ANOVA is performed as described earlier, using the S/N ratio calculated for each row of the inner array. The ANOVA can, of course, also be performed on the raw response data.

Example 11.4

Table 11.16 shows the results of a Taguchi experiment on a fuel control system, with only the variation in components A, B and C being considered to be significant. These are then selected as control factors (Inner array). The effects of two noise factors, *X* and *Y*, (Outer Array) are to be investigated. The design must be robust in terms of the central value of the output parameter, fuel flow, that is minimal variation about the nominal value.

Figure 11.6 shows graphically the effects of varying the control factors on the mean response and signal-to-noise ratio. Variation of C has the largest effect on the mean response, with A and B also having effects. However, variation of B and C has negligible effects on the signal-to-noise ratio, but the low value of A provides a much more robust design than the higher value.

This is a rather simple experimental design, to illustrate the principles. Typical experiments might utilize rather larger arrays for both the control and noise factors. Most commercially available software packages include the Taguchi method as one of the DOE options.

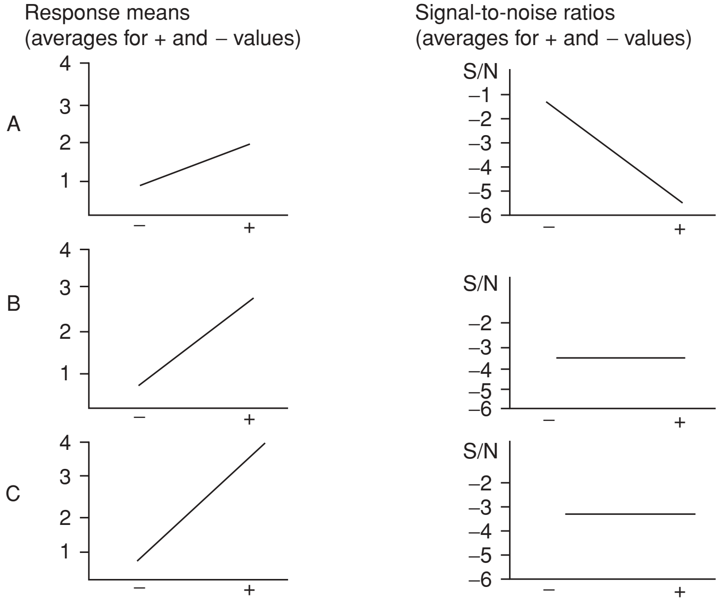


Figure 11.6 *Results of Taguchi experiment (Example 11.4).*

11.6 Conclusions

Statistical experimental methods of design optimization and problem-solving in engineering design and production can be very effective and economic. They can provide higher levels of optimization and better understanding of the effects of variables than is possible with purely deterministic approaches, when the effects are difficult to calculate or are caused by interactions. However, as with any statistical method, they do not by themselves explain why a result occurs. A scientific or engineering explanation must always be developed, so that the effects can be understood and controlled.

It is essential that careful plans are made to ensure that the experiments will provide the answers required. This is particularly important for statistical experiments, owing to the fact that several trials are involved in each experiment, and this can lead to high costs. Therefore a balance must be struck between the cost of the experiment and the value to be obtained, and care must be taken to select the experiment and parameter ranges that will give the most information. The ‘brainstorming’ approach is associated with the Taguchi method, but should be used in the planning of any engineering experiment. In this approach, all involved in the design of the product and its production processes meet and suggest which are the likely important control and noise factors, and plan the experimental framework. The team must consider all sources of variation, deterministic, functional and random, and their likely ranges, so that the most appropriate and cost-effective experiment is planned. A person who is skilled and experienced in the design and analysis of statistical experiments must be a team member, and may be the leader. It is important to create an atmosphere of trust and teamwork, and the whole team must agree with the plan once it is evolved. Note the similarity with the Quality Function Deployment method described in Chapter 7. The philosophical and psychological basis is the same, and QFD should highlight the features for which statistical experiments should be performed.

Statistical experiments are equally effective for problem solving. In particular, the brainstorming approach very often leads to the identification and solution of problems even before experiments are conducted, especially when the variation is functional, as described earlier.

The results of statistical experiments should be used as the basis for setting the relevant process controls for production. This aspect is covered in Chapter 15. In particular, the Taguchi method is compatible with modern concepts of statistical process control in production, as described in Chapter 15, as it points the way to minimizing the variation of responses, rather than just optimizing the mean value. The explicit treatment of control and noise factors is an effective way of achieving this, and is a realistic approach for most engineering applications.

The Taguchi approach has been criticized by some of the statistics community and by others (see, e.g. Logothetis, 1990 and Ryan, 2001), for not being statistically rigorous and for under-emphasizing the effects of interactions. Whilst there is some justification for these criticisms, it is important to appreciate that Taguchi has developed an operational method which deliberately economizes on the number of trials to be performed, in order to reduce experiment costs. The planning must take account of the extent to which theoretical and other knowledge, for example experience, can be used to generate a more cost-effective experiment. For example, theory and experience can often indicate when interactions are unlikely or insignificant. Also, full randomization of treatments might be omitted in an experiment involving different processing treatments, to save time. Taguchi recommends that confirmatory experiments should be conducted, to ensure that the assumptions made in the plan are valid.

It is arguable that Taguchi's greatest contribution has been to foster a much wider awareness of the power of statistical experiments for product and process design optimization and problem solving. The other major benefit has been the emphasis of the need for an integrated approach to the design of the product and of the production processes.

Statistical experiments can be conducted using computer-aided design software, when the software includes the necessary facilities, such as Monte Carlo simulation and statistical analysis routines. Of course there would be limitations in relation to the extent to which the software truly simulates the system and its responses to variation, but on the other hand experiments will be much less expensive, and quicker, than using hardware. Therefore initial optimization can often usefully be performed by simulation, with hardware experiments being run to confirm and refine the results.

The methods described in this chapter are appropriate for analysing cause-and-effect relationships which are linear, or which can be considered to be approximately linear over the likely range of variation. They cannot, of course, be used to analyse non-linear or discontinuous functions, such as resonances or changes of state.

The correct approach to the use of statistical experiments in engineering design and development, and for problem solving, is to use the power of all the methods available, as well as the skills and experience of the people involved. Teamwork and training are essential, and this in turn implies good management.

Questions

You can use statistical software to solve some of the problems below. If software is not available, trial versions of Minitab®or ReliaSoft DOE++®can be downloaded from www.minitab.com or www.reliasoft.com respectively.

1. A manufacturer has undertaken experiments to improve the hot-starting reliability of an engine. There were two control factors, mixture setting (M) and ignition timing (T), which were each set at three levels. The trials were numbered as below (which is a full factorial, equivalent to a Taguchi L9 orthogonal array with only two of its four columns allocated):

Mixture

		1	2	3
Timing	1	Expt 1	Expt 2	Expt 3
	2	Expt 4	Expt 5	Expt 6
		Expt 7	Expt 8	Expt 9

The response variable was the number of starter-motor shaft revolutions required to start the engine. Twenty results were obtained at each setting of the control factors, with averages as follows:

*Expt*1	2	3	4	5	6	7	8	9
*Av. revs*12.15	17.15	20.85	16.25	19.25	16.45	22.10	20.05	14.15

a What initial conclusions can be drawn about the control factors and their interaction (without performing a significance test)?

b There are two identifiable noise factors (throttle position and clutch pedal position), both of which can be set at two levels. How would the noise factors have been incorporated into the experimental design?

c Do you think a Taguchi signal-to-noise ratio would be a better measure than the average revs? Which one would you use?

d What additional information do you need to perform an analysis of variance?

2. a It has been suggested that reliability testing should not be applied to proving a single design parameter but should instead be based on simple factorial experimentation on key factors which might improve reliability. What do you think of this idea?

b The following factors have been suggested as possible influences on the reliability of an electromechanical assembly:

A electrical terminations (wrapped or soldered).
B type of switching circuit (relay or solid-state).
C component supplier (supplier 1 or supplier 2).
D cooling (convection or fan).

It is suspected that D might interact with B and C.

An accelerated testing procedure has been developed whereby assemblies are subjected to repeated environmental and operational cycles until they fail. As this testing is expensive, a maximum of 10 prototypes can be made and tested. Design a suitable experiment, and identify all aliased interactions.

3. One measure of the reliability of a portable communications receiver is its ability to give adequate reception under varying signal strengths, which are strongly influenced by climatic conditions and other environmental factors outside the user's control. It was decided that, within the design of a receiver, there were seven factors which could possibly influence the quality of reception.

An experiment was designed using a 2^{7-3} layout. In the design shown in Table 11.13, the six factors (denoted A–G), each at two levels, were allocated respectively to columns 1, 2, 3, 4, 12, 14 and 15. The remaining columns were left free for evaluation of interactions. It was felt safe to assume that three- and four-factor interactions would not occur.

In the experiment the 16 prototype receivers were installed in an area of known poor reception, and each was evaluated for performance on four separate occasions (between which the environmental conditions varied in a typical manner). The results are shown below (the response being an index of reception quality):

Test no.	Experimental results (reception index)			
1	6.66	5.90	6.72	4.81
2	7.76	5.77	8.36	8.62
3	5.59	6.34	7.35	8.50
4	6.36	5.37	6.17	6.46
5	7.00	6.76	5.47	5.92
6	7.52	4.71	6.69	8.14
7	7.25	5.08	5.66	5.04
8	6.18	6.47	7.55	5.92
9	7.21	5.37	7.34	4.48
10	6.95	6.96	8.36	6.87
11	7.08	5.74	6.72	6.70
12	5.34	7.56	7.22	6.89
13	8.09	8.27	5.69	5.96
14	7.72	5.62	5.77	6.79
15	6.43	6.59	6.08	5.37
16	5.52	5.82	5.82	7.29

Calculate the maximum output ('biggest is best') signal-to-noise ratios. From these, calculate the effects and sums of squares for signal-to-noise for all the factors and interactions, and carry out an analysis of variance using these values to test the significance of the various factors and interactions. What would you recommend as the final design?

4. When would you use Taguchi in place of DOE to evaluate contributors to a response? Explain.
5. Why should we always Dry-Run the DOE or Taguchi set-up before performing the experiment? Explain.
6. Why is it important to randomize for a DOE? Explain.
7. How does DOF (degrees of freedom) within the experiment help to determine significance? Explain.
8. You are conducting the experiments with two different types of solder alloys (Alloy A and Alloy B) and different temperature excursion during thermal cycling. The results of these experiments are presented in the Table in form of MTTF.

DOE mean times to failure.

	Temperature Excursion	Temperature Excursion
Solder Alloy	$\Delta T = 25\,^{\circ}\mathrm{C}$	$\Delta T = 135\,^{\circ}\mathrm{C}$
Alloy A	4000 hours	1000 hours
Alloy B	3850 hours	875 hours

Analyse the data and determine which factors are significant.

Generate the main effect plots.

Derive the equation linking MTTF with the test variables. Consider the main effects as well as the interactions.

Run ANOVA on the same data and compare the results.

Bibliography

Bhote, K.R. (1988) *World Class Quality*, American Management Association.

Box, G.E., Hunter, W.G. and Hunter, J.S. (1978) *Statistics for Experimenters*, Wiley.

Breyfogle, F. (1992) *Statistical Methods for Testing, Development and Manufacturing*, Wiley-Interscience.

Condra, L. (1993) *Reliability Improvement with Design of Experiments*, Marcel Dekker.

Deshpande, J., Gore, A. and Shanubhogue, A. (1995) *Statistical Analysis of Nonnormal Data*, Wiley.

Grove, D. and Davis, T. (1992) *Engineering Quality and Design of Experiments*, Longman Higher Education.

Hinkelmann, K. and Kempthorne, O. (2008) *Design and Analysis of Experiments, Introduction to Experimental Design*, Wiley Series in Probability and Statistics.

Lipson, C. and Sheth, N.J. (1973) *Statistical Design and Analysis of Engineering Experiments*, McGraw-Hill.

Logothetis, N. and Wynn, H. (1990) *Quality through Design*, Oxford University Press.

Mason, R.L., Gunst R.F. and Hess J.L. (1989) *Statistical Design and Analysis of Experiments*, Wiley.

Mathews, P. (2005) *Design of Experiments with MINITAB*, American Society for Quality (ASQ) Press.

Montgomery, D. (2000) *Design and Analysis of Experiments*, 5th edn, Wiley.

NIST (2011) Section 1.3.6.7.3 *Upper Critical Values of the F-Distribution*. Engineering Statistics Handbook, published by NIST. Available at: http://www.itl.nist.gov/div898/handbook/eda/section3/eda3673.htm.

Park, S. (1996) *Robust Design and Analysis for Quality Engineering*, Chapman & Hall.

Ross, P.J. (1988) *Taguchi Techniques for Quality Engineering*, McGraw-Hill.

Roy, R. (2001) *Design of Experiments Using the Taguchi Approach*, Wiley-Interscience.

Ryan, T. (2001) *Statistical Methods for Quality Improvement*, 2nd edn, Wiley-Interscience.

Taguchi, G. (1986) *Introduction to Quality Engineering*, Unipub/Asian Productivity Association.

Taguchi, G. (1978). *Systems of Experimental Design*, Unipub/Asian Productivity Association.

12
Reliability Testing

12.1 Introduction

Testing is an essential part of any engineering development programme. If the development risks are high the test programme becomes a major component of the overall development effort, in terms of time and other resources. For example, a new type of hydraulic pump or a new model of a video recording system will normally undergo exhaustive tests to determine that the design is reliable under the expected operating environments and for the expected operating life. Reliability testing is necessary because designs are seldom perfect and because designers cannot usually be aware of, or be able to analyse, all the likely causes of failure of their designs in service. The disciplines described in earlier chapters, when systematically applied, can contribute to a large extent to inherently reliable design. They can also result in fewer failures during testing, and thus reduce the time and cost of the test programme.

Reliability testing should be considered as part of an integrated test programme, which should include:

1 Functional testing, to confirm that the design meets the basic performance requirements.
2 Environmental testing, to ensure that the design is capable of operating under the expected range of environments.
3 Statistical tests, as described in Chapter 11, to optimize the design of the product and the production processes.
4 Reliability testing, to ensure (as far as is practicable) that the product will operate without failure during its expected life.
5 Safety testing, when appropriate.

It is obviously impracticable to separate entirely the various categories of test. All testing will provide information on performance and reliability, and there will be common requirements for expertise, test equipment and other resources. The different categories of test do have certain special requirements. In particular, statutory considerations often determine safety tests, some of which may have little in common with other tests.

To provide the basis for a properly integrated development test programme, the design specification should cover all criteria to be tested (function, environment, reliability, safety). The development test programme should be drawn up to cover assurance of all these design criteria. It is important to avoid competition

Practical Reliability Engineering, Fifth Edition. Patrick D. T. O'Connor and Andre Kleyner.

between people running the different categories of test, with the resulting arguments about allocation of models, facilities, and priorities. An integrated test programme reduces the chances of conflict.

The development test programme should include:

1 Model allocations (components, sub-assemblies, system).
2 Requirements for facilities such as test equipment.
3 A common test and failure reporting system.
4 Test plan and schedule.

One person should be put in charge of the entire programme, with the responsibility and authority for ensuring that all specification criteria will be demonstrated.

There is one conflict inherent in reliability testing as part of an integrated test programme, however. To obtain information about reliability in a cost-effective way, that is quickly, it is necessary to generate failures. Only then can safety margins be ascertained. On the other hand, failures interfere with functional and environmental testing. The development test programme must address this dilemma. It can be very tempting for the people running the development test programme to minimize the chance of failure occurring, in order to make the programme run smoothly and at least cost. However, weaknesses in the design (or in the way it is made) must be detected and corrected before the production phase. This can realistically be achieved only by generating failures. An ideal test programme will show up every failure mode which might otherwise occur in service.

The development test dilemma should be addressed by dividing tests into two main categories:

1 Tests in which failures are undesirable (test to success).
2 Tests which deliberately generate failures (test to failure).

Statistical testing, functional testing, system level testing and most environmental testing are in category 1. Most reliability testing (and some safety testing) are in category 2, although reliability testing may belong to both categories, depending on the objectives of this testing (see reliability demonstration in Chapter 14). In particular, there must be a common reporting system for test results and failures, and for action to be taken to analyse and correct failure modes. Test and failure reporting and corrective action are covered in more detail later.

The category 2 testing should be started as soon as hardware (and software, when appropriate) is available for test, no later than the VERIFY (preferably earlier) stage of the design for reliability (DfR) process (Chapter 7). The effect of failures on schedule and cost increases progressively, the later they occur in the development programme (see Figure 7.1). Therefore tests should be planned to show up failure modes as early as is practicable. The category 1 testing is more appropriate for the VALIDATE stage of the DfR process, although test to failure is not uncommon at this stage as well.

Engineering development testing methods are described in more detail in O'Connor (2001).

12.2 Planning Reliability Testing

12.2.1 Using Design Analysis Data

The design analyses performed during the design phase (CAE, reliability prediction, FMECA, stress analysis, parameter variation analysis, sneak circuit analysis, FTA) described in Chapters 6, 7 and 9, as well as any earlier test results, should be used in preparing the reliability test plan. These should have highlighted the risks and uncertainties in the design, and the reliability test programme should specifically address these. For example, if the FMECA shows a particular failure mode to be highly critical, the reliability test programme should confirm that the failure is very unlikely to occur within the use environment and lifetime. Inevitably the test programme will also show up failure modes and effects not perceived during the design analyses, otherwise there would be little point in testing. Therefore, the test programme must cover the whole range of use conditions, including storage, handling, testing, repair and any other aspect which might affect reliability.

12.2.2 Considering Variability

We have seen in Chapters 5 and 11 how variability affects the probability of failure. A major source of variability is the range of production processes involved in converting designs into hardware. Therefore the reliability test programme must cover the effects of variability on the expected and unexpected failure modes. If parameter variation analyses or statistical tests have been performed, these can be very useful in planning reliability tests to confirm the effects of variation. However, to ensure that the effects of variability are covered as far as is practicable, it is important to carry out reliability testing on several items. The number of systems to be tested must be determined by considering:

1 The extent to which the key variables can be controlled.
2 The criticality of failure.
3 The cost of test hardware and of testing.

Only rarely will fewer than four items be adequate. For fairly simple systems (transistors, fasteners, hydraulic check valves, etc.) it might be relatively easy to control the few key variables and the criticality of failures might be relatively low. However, it is not expensive to test large quantities. For systems of moderate complexity (e.g. automobiles, TV sets, machine tools) it is much harder to control key variables, since there are so many. Every interface within the system introduces further sources of variability which can affect reliability. Therefore it is very important to test a relatively large number, five to 20 being a typical range. For complex systems (aero engines, aircraft, etc.) hardware and test cost tend to be the major constraints, but at least four items should be subjected to a reliability test. Reliability testing of fewer than four might be appropriate for large, expensive, complex systems which will be manufactured in very small quantities (e.g. spacecraft, ships, power stations), of which the items tested will be used operationally.

The effects of known sources of variability can sometimes be assessed by testing items in which variable parameters (e.g. dimensions, process variables) have been deliberately set at worst case values. Statistical design of experiment (DOE) and other statistical engineering optimization techniques, as described in Chapter 11, should be used to analyse the effects of multiple sources of variation.

12.2.3 Durability

The reliability test programme must take account of the pattern of the main failure modes with respect to time (or cycles, distances, etc., with which the time dimension is associated).

If the failure modes have increasing hazard rates, testing must be directed towards assuring adequate reliability during the expected life. Therefore reliability tests must be of sufficient duration to demonstrate this, or they must be accelerated. Accelerated testing is covered later. Generally speaking, mechanical components and assemblies are subject to increasing hazard rates, when wear, fatigue, corrosion or other deterioration processes can cause failure. Systems subject to repair and overhaul can also become less reliable with age, due to the effects of maintenance, so the appropriate maintenance actions must be included in the test plan.

12.3 Test Environments

The reliability test programme must cover the range of environmental conditions which the product is likely to have to endure. The main reliability-affecting environmental factors, affecting most products, are:

Temperature.
Vibration.
Mechanical shock.
Humidity.
Power input and output.
Voltage (electronics).
Dirt/dust.
Contaminants.
People.

In addition, electronic equipment might be subjected to:

Electromagnetic effects (EMI).
Voltage transients, including electrostatic discharge (ESD).

Certain other environments can affect reliability in special cases. Examples are:

Radiation (ultraviolet, cosmic, X-rays).
Lubricant age or contamination.
High altitude.
Space vacuum.
Industrial pollution.
Electromagnetic pulse (lightning, nuclear).
Salt spray.
Fungus.
High intensity noise.
Noxious gases.

US MIL-STD-810, UK Defence Standard 07-55 and ISO/IEC60068 (see Bibliography) provide test methods appropriate to most of these environmental conditions. However, these standards do not address reliability directly, since the objective is to show that the product will not fail or incur damage under the test conditions. Also, most of the tests do not require that the equipment be operating during the tests, and the tests are single-environment, not combined.

The environmental test programme will address the formal environmental test requirements, particularly when these are necessary in order to comply with legal or contractual requirements. The environmental aspects of the reliability test programme must take account of the environmental requirements stated in the design specification and of the planned environmental test. However, to be effective as a means of ensuring a reliable product, the environmental aspects of reliability testing must be assessed in much greater detail.

The environmental aspects of reliability testing must be determined by considering which environmental conditions, singly and in combination with others, are likely to be the most critical from the reliability point of view. In most cases, past experience and codes of practice will provide adequate guidelines. For example, US MIL-HDBK-781 provides information on how to assess environmental conditions and to design the

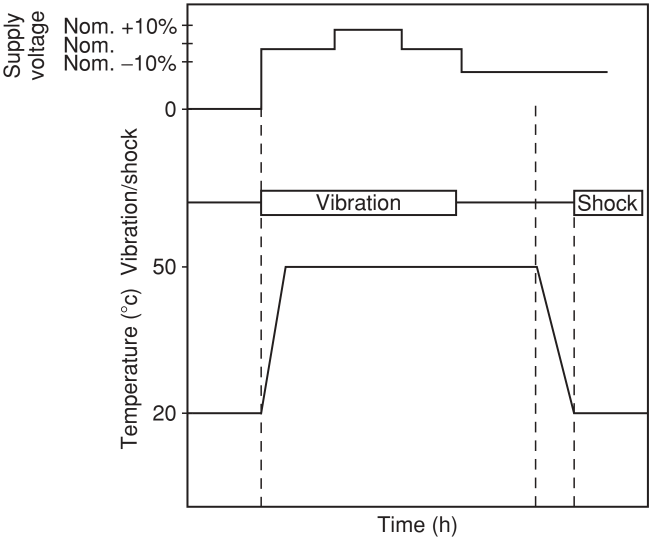


Figure 12.1 Typical CERT environmental cycles: electronic equipment in a vehicle application.

tests accordingly. Typically, a reliability test environment for an electronic system to be used in a vehicle or aircraft might be as shown in Figure 12.1. Such testing is known as *combined environmental reliability testing* (CERT).

Test chambers are available for CERT testing, particularly for electronic systems. These include facilities for temperature cycling and for vibration input to the unit under test by locating the chamber over a floor-mounted vibrator, with a movable or flexible floor for the chamber. Electrical signals, (power, control and monitoring) can be fed through connectors in the chamber wall. Special chambers can be provided with other facilities, such as humidity and reduced pressure. Control of the chamber conditions can be programmed, and the unit under test can be controlled and monitored using external equipment, such as programmable power supplies, data loggers, and so on.

Figure 12.2 shows a typical CERT facility.

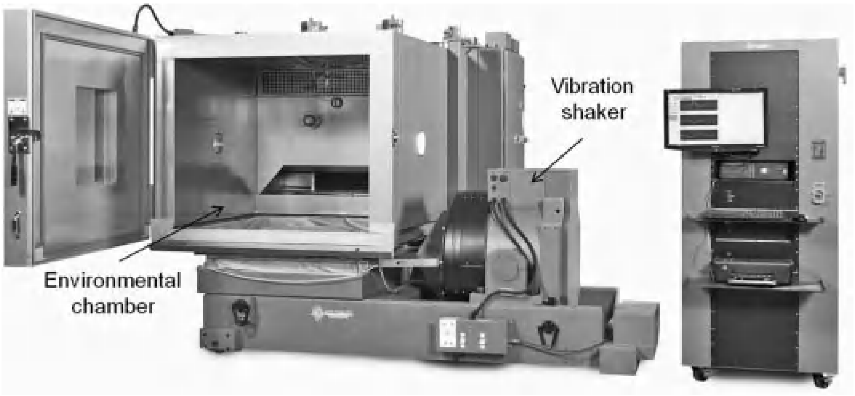


Figure 12.2 CERT test facility (Reproduced by permission of Thermotron Industries).

When past experience on standard methods is inappropriate, for example for a high risk product to be used in a harsh environment, the test environments must be carefully evaluated, particularly:

1 Rate of change of conditions, not just maximum and minimum values. For example, a high rate of change of temperature can cause fracture or fatigue due to thermal mismatch and conductivity effects.
2 Operating and dormant conditions in relation to the outside environment. For example, moisture-assisted corrosion might cause more problems when equipment is idle than when it is operating.
3 The effects of combined environments, which might be much more severe than any one condition. Statistical experiments (Chapter 11) can be used to evaluate these effects.
4 Direction and modes of vibration and shock. This is dealt with in more detail later.
5 Particular environmental conditions applicable to the product, such as handling, storage, maintenance and particular physical conditions.

12.3.1 Vibration Testing

Adequate vibration testing is particularly important for products which must survive vibration conditions. However, specifying and obtaining the right conditions can be difficult, and it is easy to make expensive mistakes.

The main principles of effective vibration testing are:

1 Vibration should be input to the device under test (DUT) through more than one axis, preferably simultaneously.
2 Vibration inputs should cover the complete range of expected frequencies and intensities, so that all resonances will be excited.
3 In most applications vibration input should be random, rather than swept frequency, so that different resonances will be excited simultaneously (see below).
4 Test fixtures to mount the DUT to the vibration tables should be designed so that they do not alter the vibration output (no fixture resonances or damping). Whenever practicable, the DUT should be mounted directly on to the vibrator platform.

The simplest vibration test is a fixed frequency ‘shake’, usually with a sine wave input. However, this is of little value in reliability testing. Modern vibrators can be programmed to generate any desired profile.

Swept frequency sine testing is useful for resonance searches, to enable the design to be modified if unacceptable resonances are detected.

Peak acceleration for a given frequency of sine wave vibration can be calculated using the formula:

$$A = 0.002f^2 D \tag{12.1}$$

where: $A =$ peak acceleration (g).
$f =$ frequency (Hz).
$D =$ peak-to-peak displacement (mm)
for example, if $f = 50$ Hz and $D = 2$ mm then $A = 10$ g.

Another type of sinusoidal vibration is sine-dwell, where the DUT is vibrated for a period of time at its resonant frequency in order to generate the maximum amount of stress.

Alternatively, the spectrum could be a random input within a specified range and density function. Random vibration testing in which the input contains many frequencies is more effective than swept frequency for

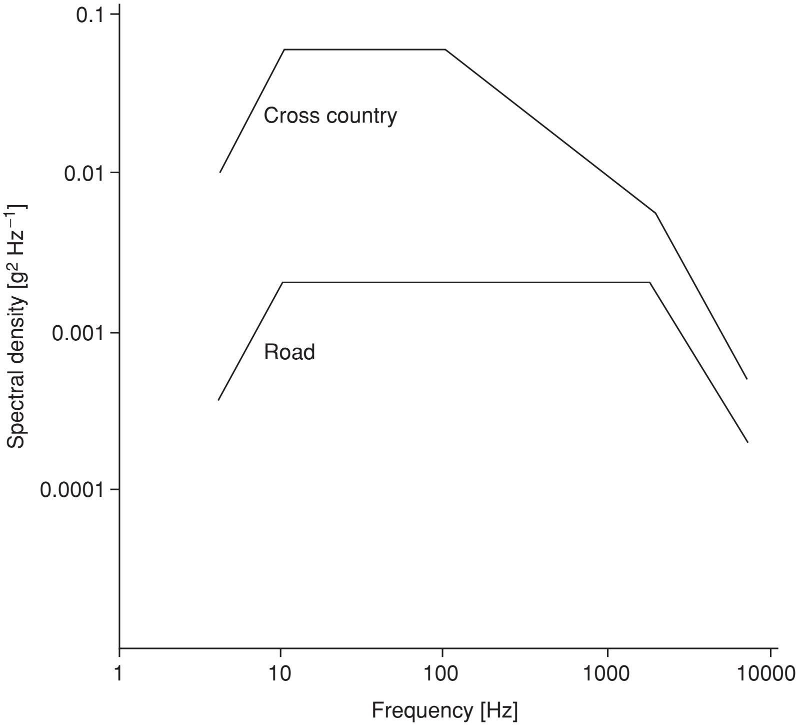


Figure 12.3 *Road transport vibration levels.*

reliability testing, to show up vibration-induced failure modes, since it simultaneously excites all resonances. It is also more representative of real life.

The unit of measurement for random vibration inputs with continuous spectra is *power spectral density* (PSD). The units are g^2/Hz. Typically inputs of up to 0.1 g^2/Hz are used for equipment which must be shown to withstand fairly severe vibration, or for screening tests on assemblies such as electronic equipment. A typical random vibration spectrum is shown in Figure 12.3.

It is important to apply power to electrical or electronic equipment and to monitor its performance while it is being vibrated, so that intermittent failures can be detected and investigated.

Since dynamic responses are usually affected more by resonances within the product, due to the design and to production variation, than by the input spectrum from the vibrator, it is seldom cost-effective to simulate accurately the operating environment, even if it is known in detail. Since the objective in vibration reliability testing is to excite resonances simultaneously, and since almost any random spectrum will do this, test costs can be minimized by permitting large spectral tolerances, for example ± 6 dB.

Vibration and shock testing are described in more detail in O'Connor (2001), Harris (2010) and Steinberg (2000).

12.3.2 Temperature Testing

The most common types of temperature tests are constant temperature, temperature cycling and thermal shock. Constant temperature tests are more common in the electronics industry and are designed to evaluate the operational or storage capabilities of the product under extreme low or extreme high temperatures. Temperature cycling and thermal shock are intended to subject the product to *low cycle fatigue* (as opposed to *high cycle fatigue* experienced during vibration). Due to mismatch between coefficients of thermal expansion for different materials, thermal cycling causes stress cycling which often results in fatigue failures. Thermal cycling and thermal shock are defined by the extreme values (high and low temperatures) and the rate of transition, which is typically much higher in thermal shock tests. Temperature testing for electrical and electronic equipment is particularly important, since reliability can be affected by operating temperature (Chapter 9) and by thermal cycling. In most cases equipment should be powered and operated during temperature testing, otherwise the tests will be unrepresentative of the thermal patterns and gradients in use. It should also be monitored continuously to ensure that intermittent failures are detected.

12.3.3 Electromagnetic Compatibility (EMC) Testing

EMC testing is very important for electronic systems, since data corruption due to electromagnetic interference (EMI) or from voltage transients in power supplies can have serious consequences (see Chapter 9). The equipment must be subjected to EMI and transients to confirm that it will perform without failure under these conditions. The levels of EMI and transient waveforms must be ascertained by evaluating or measuring the operating environment, or they might be specified.

Internally induced EMI and transients must also be protected against, and tests to ensure that transmitted EMI is within limits might also be necessary.

EMI/EMC testing methods are described in US MIL-STD-462D and IEC 61000 (see Bibliography).

12.3.4 Other Environments

See O'Connor (2001).

12.3.5 Customer Simulation Testing

Functional, environmental, reliability and safety tests are all designed to demonstrate that the equipment will meet its design parameters. In general these tests are carried out by people with extensive engineering backgrounds. Such people are, with the best will in the world, often far removed from the average user of the equipment. It is therefore important, particularly in the field of consumer products (televisions, copiers, washing machines, etc.), that some reliability testing is conducted using people who are more nearly representative of typical customers or by trial customers. This is called 'beta' testing. This approach is very useful in highlighting failure modes that do not show up when the equipment is used by experienced personnel or in non-representative environments. For example, car companies often use 'fleet' vehicles (police, rental and delivery) to test new parts and systems. These vehicles accumulate mileage much faster than the ordinary cars, therefore potential design problems may be discovered much sooner.

12.4 Testing for Reliability and Durability: Accelerated Test

In Chapters 8 and 9 we reviewed how mechanical, electrical and other stresses can lead to failures, and in Chapter 5 how variations of strength, stresses and other conditions can influence the likelihood of failure or duration (time, distance, cycles, etc.) to failure. In this section we will describe how tests should be designed and conducted to provide assurance that designs and products are reliable and durable in service.

12.4.1 Test Development

For most engineering designs we do not know what is the '*uncertainty gap*' between the theoretical and real capabilities of the design and of the products made to it, for the whole population, over their operating lives and environments. The effects of these uncertainties can seldom be evaluated with confidence by any of the design analysis methods described in Chapter 7. How then can we plan a test programme that will reduce the uncertainty gap to an extent that we can be assured of reliability and durability, whilst taking due account of practical constraints like cost and time?

The conventional approach to this problem has been to treat reliability as a functional performance characteristic that can be measured, by testing items over a period of time whilst applying simulated or actual in-service conditions, and then calculating the reliability achieved on the test. For example, time of operation divided by the number of failures is the estimated mean time between failures (MTBF). This approach will be discussed in Chapter 14. However, these methods are fundamentally inadequate for providing assurance of reliability. The main reason is that they are based on measuring the reliability achieved during the application of simulated or actual stresses that are within the specified service environments, in the expectation (or hope) that the number of failures will be below the criterion for the test. This is the wrong answer to the problem expressed above.

The correct answer is straightforward: *we must test to cause failures, not test to demonstrate successful achievement*. This concept is well accepted in many industries, particularly where mechanical strength testing is involved. It is important to remember that most of the failures in modern electronics are also of mechanical nature. To derive the strength and fatigue properties of materials, samples are tested to failure. As explained in Chapter 8, we cannot precisely determine the strength of, say, an alloy or a plastic material by theoretical analysis, only by testing samples to failure. If we design a component using such a material, we can analyse the stresses using methods like FEA and we can calculate the strength using the material properties derived from published data or, where necessary, the tests to failure. If the design is simple and there is an adequate margin between stress and strength, we might decide that no further testing is necessary. If, however, constraints such as weight force us to design with smaller margins, and if the component's function is critical (like supporting an aircraft engine), we might well consider it prudent to test some quantity to failure. We would then expect that failures would occur only well beyond the expected maximum stress/minimum life, to provide an adequate margin of safety to take account of the uncertainties and variations in this kind of design and application.

However, let us assume that a system is being designed, and the specified maximum temperature for satisfactory operation is 40 °C. Up to what temperature should the prototype be tested? Some inexperienced (and some experienced) engineers answer that 40 °C should be the maximum test temperature, because any temperature above that would not be 'representative' of specified conditions. Therefore any failures that occur above that temperature would not be considered relevant.

However, suppose that a prototype was tested at 42 °C, and failed. Should we ignore this? Might this failure occur at 35 °C on another unit built to the same drawings (effect of variability), or might it occur on this unit six months into the warranty period (effect of a time-dependent failure mechanism)? Might it occur at a combination of 35 °C and a small, within-specification, increase in supply voltage? Can we really be sure that the failure at 42 °C is not relevant, just because the thermal stress applied was not 'representative'?

If the failure occurred at a temperature 2 °C above the specified limit it is unlikely that it would be ignored (though this does happen). Suppose, however, that failure occurred at 50 °C, or 60 °C? At what stress do we decide that the level is so high that we can ignore failures? Should we even be testing at stresses so much higher than the maximum specified values?

The answer is that these are the wrong questions. The clue is in the earlier questions about the possible cause of the failure. When failures occur on test we should ask whether they could occur in use. The question

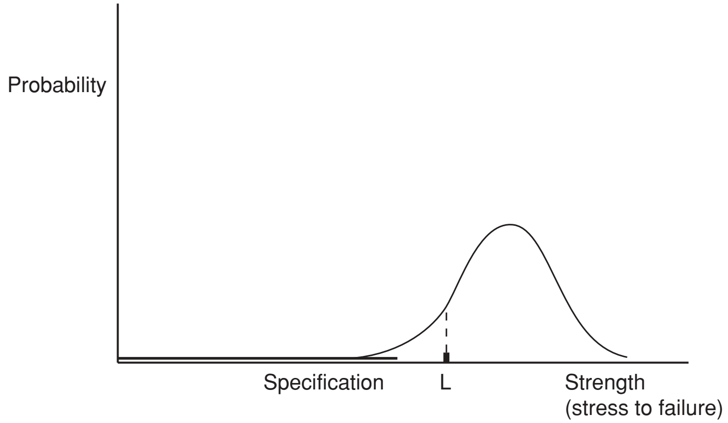


Figure 12.4 *Stress, strength and test failures (1).*

of relevance can be answered only by investigating the actual physical or chemical cause of failure. Then the questions that must be asked are:

1 Could this failure occur in use (on other items, after longer times, at other stresses, etc.)?
2 Could we prevent it from happening in use?

The stress/es that were applied are relevant only in so far as *they were the tools to provide the evidence that an opportunity exists to improve the design. We have obtained information on how to reduce the uncertainty gap*. Whether the opportunity is taken is a management issue, in which other aspects such as cost, weight, time, and so on must be considered.

The uncertainty described above is shown in Figure 12.4. If we consider only one stress and the failures it might cause, the stress to failure distribution of production items might be as shown. As a simple example, this might be the operating temperature at which an electronic component malfunctions, or the pressure at which a seal begins to leak. As discussed in Chapter 2, the exact nature of this distribution is almost always uncertain, particularly in the tails, which are the most important areas as far as reliability and durability are concerned.

Suppose that the first test failure occurs at stress level *L*. At this stage we might have only a few items to test, maybe only one. We can state that the strength of this item represents a point on the distribution, but we cannot say whether it was an average strength item, a strong one or a weak one. The only way to find out the nature of the strength distribution is to test more items to failure, and to plot and analyse the results. Inevitably most of the items tested will be near to the average, because that is where most of the population will lie. Therefore, it is unlikely that any item tested will represent the weakest in the future population.

However, if we analyse the actual cause of the failure, by whatever means is appropriate, and take action to strengthen the item, then in effect we will move the strength distribution to the right. We still do not know its shape, but that is not what is important. We just want to move it out of the way. We are engineers, not theoretical scientists or statisticians, *so we can use high stresses in place of large samples*. Whilst scientific knowledge of the cause and effect relationships that affect reliability and durability is obviously necessary in order to create designs *and to determine how to improve them*, this is appropriate to determining where distributed values are centred, and sometimes the variation near the centre. In the electronics example, we

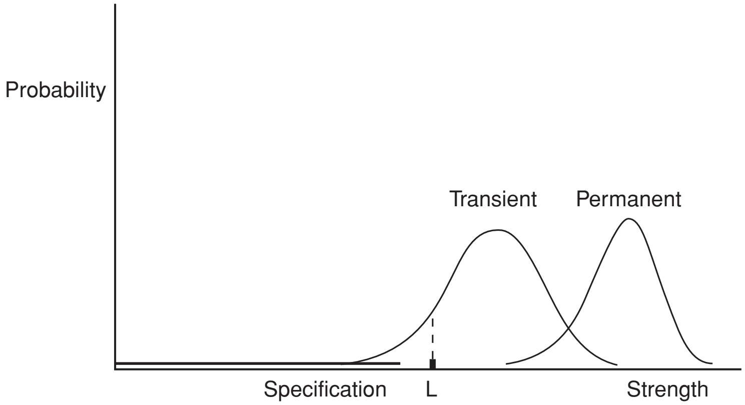


Figure 12.5 Stress, strength and test failures (2).

might use a higher rated component, or add a heat sink. For the seal we might change the material or the dimensions, or add a second seal, or reduce the pressure. The system will therefore be made more reliable.

For many items subjected to stress tests, particularly electronic systems, two types of failure can occur: *transient (or operating) failures* and *permanent failures*. If a transient failure occurs at some stress level the correct operation will be restored if the stress is reduced. Permanent failures are those from which the operation does not recover when the stress is reduced. For stresses like temperature, power voltage level, and so on, which might have low or negative limiting values, failures might occur at high and at low levels, and these will of course have different physical causes from those at the high levels. For a population of items the stresses at which the failures occur will be distributed. The general case is illustrated in Figure 12.5.

If the failure is due to a wearout mechanism, say wear of a bearing or fatigue of a component attachment, the horizontal axis of the distribution will represent time (or cycles), for any particular stress value. In addition to the uncertainty regarding the stress and strength, we now have the further uncertainty of time. Failure on test after time t at stress level L will represent one point on an unknown three-dimensional distribution (Figure 12.6). As Figure 12.6 illustrates, an important feature of wearout mechanisms is that the resulting distributions of times to failure become wider as damage accumulates, thus further increasing the uncertainty.

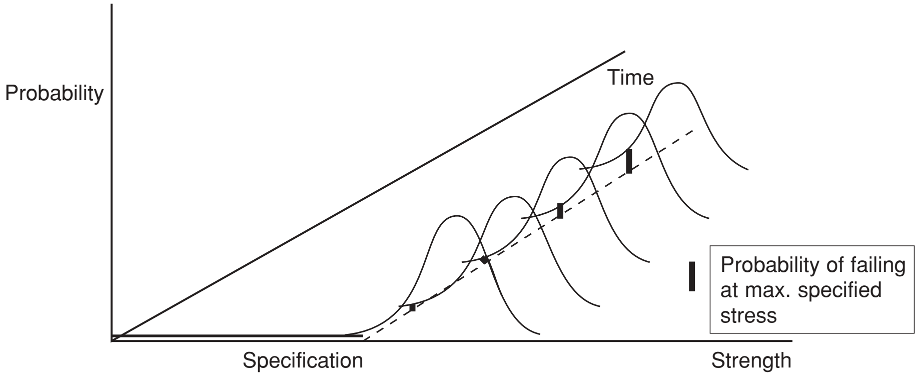


Figure 12.6 Stress, strength and test failures (3): wearout failures.

To obtain a full understanding might require more testing, with larger samples. However, the same principle applies: we are not really interested in the shapes of distributions. We just want to make the design better, if it is cost effective to do so.

In most engineering situations failures are caused by combinations of stresses and strength values, not just by one stress and one strength variable. Some might be time-dependent, others not. Using the examples above:

- The stresses applied to the electronic component could be a combination of high temperature operation, high temperature rate of change after switch-on, rate of on–off cycles, humidity when not operating, power supply voltage level, and vibration. The resisting strengths might be mechanical integrity of the internal connections, thermal conductivity of the encapsulating material, absence of defects, and so on.
- For the seal, the stresses and other variables might be oil temperature, pressure, pressure fluctuations, oil conditions (viscosity, cleanliness, etc.), shaft axial and radial movement, vibration, tolerances between moving parts, tolerances on seal grooves and seal dimensions, and so on.

Between any two or more of these variables there might also be interactions, as described in Chapter 11.

Therefore, even for relatively simple and common failure situations like these, there is not just one distribution that is important, but a number of possible distributions and interactions. What might cause a transistor, capacitor or seal to fail in one application might have a negligible effect in another, or a component that has worked well in previous applications might cause problems in a new but similar one. This is how life is in engineering!

This reasoning leads to the main principle of development testing for reliability. *We should increase the stresses so that we cause failures to occur, then use the information to improve reliability and durability*. This is particularly true in the DESIGN-ANALYZE-VERIFY cycle of the DfR process (Chapter 7). Clearly there will be practical limits to the stresses applied. These limits are set by:

- The fundamental limits of the technology. For example, there is no point in testing an electronic system at temperatures above the melting point of the solder used.
- The limits of the test capability, such as the maximum temperature of the test chamber.

The logic that justifies the use of very high ‘unrepresentative’ stresses is based upon four aspects of engineering reality:

1 The causes of failures that will occur in the future are often very uncertain.
2 The probabilities of and durations to failures are also highly uncertain.
3 Time spent on testing is expensive, so the more quickly we can reduce the uncertainty gap the better.
4 Finding causes of failure during development and preventing recurrence is far less expensive than finding new failure causes in use.

It cannot be emphasized too strongly: testing at ‘representative’ stresses, in the hope that failures will not occur, is very expensive in time and money and is mostly a waste of resources. It is unfortunate that nearly all standardized approaches to stress testing (these standards are discussed later) demand the use of typical or maximum specified stresses. This approach is widely applied in industry, and it is common to observe prototypes on long-duration tests with ‘simulated’ stresses applied. For example, engines are run on test beds for hundreds of hours, cars are run for thousands of miles around test tracks, and electronic systems are run for thousands of hours in environmental test chambers. Tests in which the prototype does not fail are considered to be ‘successes’. However, despite the long durations and high costs involved, relatively few opportunities for improvement are identified, and failures occur in service that were not observed during testing. An important point to realize in this context is that a failure that is generated by a stress level during test *might be generated by a different stress (or stresses) in service*. For example, a fatigue failure caused by a few minutes vibration on test might be caused by months or years of temperature cycling in service. The vibration stress applied on test might be totally unrepresentative of service conditions. Once again, though, the principle applies that the 'unrepresentative' test stress might stimulate a relevant failure. Furthermore, it will have done so much more rapidly than would have been the case if temperature cycling had been applied. However it is more common in test engineering to subject a product to the same types of environment as in the field, only at higher levels (*accelerated test*).

12.4.2 Accelerated Test

Product testing at accelerated levels is very common in many industries. Testing a product above the level of its specification makes it fail sooner and provides additional information about its strength. It also helps to lower the cost of development by reducing test time and thus helping to deliver products to market in the shortest possible time.

The same environments discussed in Section 12.3 (temperature, vibration, humidity, etc.) can be used in accelerated testing. For example, making a product operate at elevated temperatures or subjecting it to higher vibration levels will shorten its test time.

The theoretical concept of the effect of accelerated test on a product's life is shown in Figure 12.7. Higher stress levels shorten the expected product life and increase the expected failure rate at all phases of the bathtub curve.

Understanding the potential failure mechanisms and product design limits is critical to developing a successful accelerated test. Figure 12.8 shows the typical stress ranges and the expected types of failures. These stress limits can be two-sided (e.g. temperature ranges) as shown in Figure 12.8 or one-sided (e.g. vibration or voltage). Increasing the stress beyond the product's design limits may precipitate failures which would not be representative of the field environment. For example, plastic parts may exceed their glass transition points or even melt at high ambient temperatures, something which would not happen under normal usage conditions. These types of failures are often referred as *foolish failures* and should be avoided during product testing. Accelerated stress levels should be chosen that so they accelerate the 'realistic' failure modes, which are expected in the field. Understanding of the technology, previous experience with similar products, and design team inputs should help in the process of determining the appropriate stress levels. Also

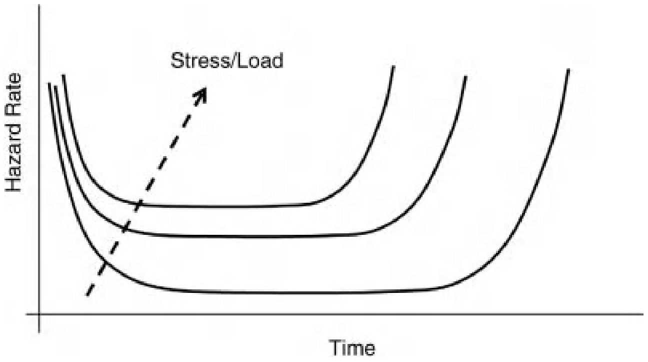


Figure 12.7 *Effect of accelerated test on the bathtub curve.*

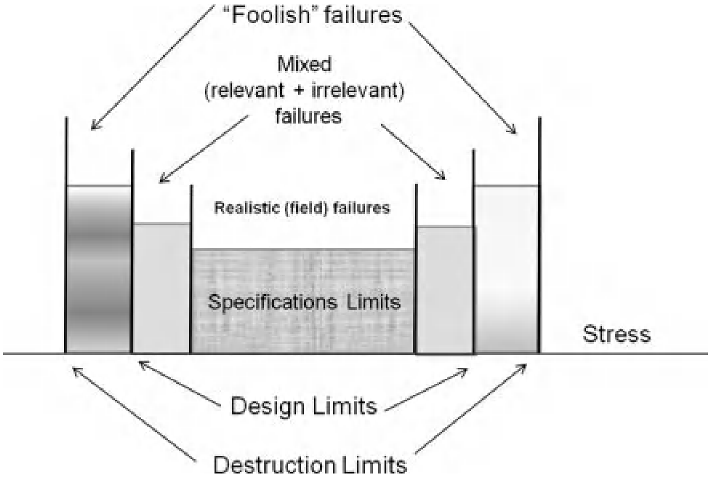


Figure 12.8 *Stress ranges and types of failures.*

proper use of DOE (Chapter 11) is critical at this step. A similar concept as applied to manufacturing stress screening is described in Chapter 15 (Figure 15.8). Accelerated test models and how to apply them to the test development will be discussed in Chapter 13. Accelerated stress testing can provide quantitative cause and effect information when the mechanisms of failure are already understood (e.g. material fatigue), and when the tests are planned specifically to provide such information. We can perform statistically designed experiments (Chapter 11), applying accelerated stresses to explore their effects. However, tests to provide such information require larger samples, more detailed planning and more time, and therefore would cost more than would be the case in the accelerated test approach described above. We must decide whether we need the extra information that more 'scientific' or statistical tests can provide. In many cases the information from accelerated tests, as described above, coupled with engineering knowledge, is sufficient to enable us to take appropriate action to improve designs and processes. However, sometimes we need to obtain more detailed information, especially when the cause and effect relationships are uncertain.

Electronic systems have been subjected to accelerated temperature and vibration stresses not only in development, but also for production units. This is called *environmental stress screening* (ESS). Other names have been used for the same approach, including STRIFE (stress + life). ESS methods have been standardized to some extent as a result of the guidelines published by the US Institute of Environmental Sciences and Technology (IEST) and ISO/IEC 61163. Also, apart from the fairly limited stress combinations applied in ESS or CERT, the stresses are usually applied singly. The objective is to simulate the expected worst-case service environments, or to accelerate them by only moderate amounts. The general principle usually applied to all of these methods has been to test the item to ensure that it does not fail during the test. *This is not consistent with the approach discussed above, and in more detail below.*

12.4.3 Highly Accelerated Life Testing

A test in which stresses applied to the product are well beyond normal shipping, storage and in-use levels is called *Highly Accelerated Life Testing* (HALT). The principle was developed by an American engineer, Dr Gregg Hobbs, and it is fully described in McLean (2009). HALT is usually conducted in specially designed environmental HALT chambers. They can combine wide temperature ranges with fast transition and high

G_{RMS} random vibration. A HALT chamber can provide a temperature range of [−100; +200] °C with up to 100 G_{RMS} repetitive shock 6 degree of freedom vibration. Repetitive shock vibration tables typically allow limited control over the shape of the random vibration spectrum. In HALT we make no attempts to simulate the service environment, except possibly as the starting point for the step-stress application. This type of accelerated test with no clear acceleration model is often referred as *Qualitative Accelerated Test* (as opposed to *Quantitative Accelerated Test*, discussed in the previous section). No limits are set to the types and levels of stresses to be applied. We apply whatever stresses might cause failures to occur as soon as practicable, whilst the equipment is continually operated and monitored. We then analyse the failures as described above, and improve the design. These design improvements can expand the design limits (Figure 12.8) moving them further away from the specification limits, thus reducing the chances of failure.

By applying stresses well in excess of those that will be seen in service, failures are caused to occur much more quickly. Typically the times or cycles to failure in HALT will be several orders of magnitude less than would be observed in service. Failures which might occur after months or years in service are stimulated in minutes by HALT. Also, the very small sample usually available for development test will show up failure modes that might occur on only a very small proportion of manufactured items. *Therefore we obtain time compression of the test programme by orders of magnitude, and much increased effectiveness. This generates proportional reductions in test programme cost and in time to market, as well as greatly improved reliability and durability*.

It is important to appreciate that reliability/durability values cannot be demonstrated or measured using HALT. An accelerated stress test can provide such information only if the cause of failure is a single predominant mechanism such as fatigue, we know exactly what single type of stress was applied, and we have a credible mathematical relationship to link the two. Such relationships exist for some failure mechanisms, as described in Chapters 8 and 9. However, since HALT applies a range of simultaneous stresses, and since the stress profiles (particularly the vibration inputs) are complex and unrecorded, such relationships cannot be derived. In HALT we are trying to stimulate failures as quickly as possible, using highly ‘unrepresentative’ stresses, so it is impossible and misleading to relate the results to any quantitative reliability/durability requirement such as MTBF, MTTF, and so on.

The HALT approach can be applied to any kind of product or technology. For example:

- *Engines, pumps, power transmission units such as gearboxes, and so on.*
  - Start tests with old lubricants or other fluids (coolants, hydraulics, etc.), rather than new.
  - Run at low fluid levels.
  - Use fluids that are heated, cooled or contaminated.
  - Use old filters.
  - Misalign shafts, bearings, and so on.
  - Apply out-of-balance to rotating components.
- *Electro-mechanical assemblies such as printers, document, material or component handlers, and so on.*
  - Apply high/low temperatures, vibration, humidity/damp, and so on.
  - Use components with out-of-tolerance dimensions.
  - Misalign shafts, bearings, and so on.
  - Use papers/documents/materials/components that exceed specifications (thickness, weight, friction, etc.).
- *Small components or assemblies such as electronic packages, mechanical latches, switches, transducers, and so on.*
- Apply high/low temperatures, vibration, humidity/damp, and so on.
- Apply high frequency vibration by fixing to suitable transducers, such as loudspeaker coils, and driving with an audio amplifier.

12.4.4 Test Approach for Accelerated Test

The approach that should be applied to any accelerated test programme for reliability/durability should be:

1 Try to determine, as far as practicable, what failures might occur in service. This should have been performed during design analysis and review, particularly during the quality function deployment (QFD) and failure modes, effects and criticality analysis (FMECA) (Chapter 7).
2 List the application and environmental stresses that might cause failures. Use Chapters 8 and 9 for guidance.
3 Plan how the stresses that might stimulate foreseeable and unforeseen failures can most effectively be applied in test. Set up the item (or items) to be tested in the test chamber or other facility so that it can be operated and monitored.
4 Apply a single stress, at or near to the design maximum, and increase the level stepwise until the first failure is detected. This approach is called *step-stress* accelerated testing.
5 Determine the cause and take action to strengthen the design so that it will survive higher stresses. This action might be a permanent improvement, or a temporary measure to enable testing to be continued.
6 Continue increasing the stress(es) to discover further failure causes (or the same cause at a higher stress), and take action as above.
7 Continue until all of the transient and permanent failure modes for the applied stress are discovered and, as far as technologically and economically practicable, designed out. Repeat for other single stresses.
8 Decide when to stop (fundamental technology limit, limit of stress that can be applied, cost or weight limit).
9 Repeat the process using combined stresses, as appropriate and within the equipment capabilities (temperature, vibration, power supply voltage, etc.).

Note that there is a variety of different test profiles besides step-stress, which can be utilized in HALT. The selection of stresses to be applied, singly or in combination, is based upon experience and on the hardware being tested, and not on specifications or standards.

12.4.5 HALT and Production Testing

HALT does not only provide evidence on how to make designs more robust. It also provides the information necessary to optimize stress screens for manufacturing. The basic difference between the objectives of accelerated test in development and in manufacturing is that, whilst we try to cause all development test items to fail in order to learn how to improve the design, we must try to avoid damaging good manufactured items, whilst causing weak or defective ones to fail so that they can be corrected or segregated. The knowledge that we gain by applying the full HALT sequence, including the design ruggedization, can be used to design a stress test regime that is optimized for the product, and which is far more effective than conventional production testing. This is called *highly accelerated stress screening* (HASS). HASS provides the same benefits in manufacturing as HALT does in development, in greatly increasing the effectiveness of manufacturing screens whilst reducing test cost and time. We will describe manufacturing testing in detail in Chapter 15.

HALT and HASS represent an integrated approach to testing to ensure that both the design and the manufacturing processes will generate highly reliable products, at minimum cost and time. Conventional, separate, approaches to development and manufacturing tests do not assure this integration, and therefore can result in much lower reliability and higher costs.

Table 12.1 DoE/HALT Selection

Important Variables, Effects, etc.	DoE/HALT?
Parameters: electrical, dimensions, etc.	DoE
Effects on measured performance parameters, yields	DoE
Stress: temperature, vibration, etc.	HALT
Effects on reliability/durability	HALT
Several uncertain variables	DoE
Not enough items available for DoE	HALT
Not enough time available for DoE	HALT

12.4.6 DoE or HALT?

Statistical experiments and HALT are complementary approaches in development testing. Table 12.1 gives some guidance on which approach to select for particular situations.

Note that these are by no means clear-cut criteria, and there will often be shades of grey between them. We must decide on the most appropriate method or combination of methods in relation to all of the factors: risks, knowledge, costs, time.

12.5 Test Planning

Testing is an integral part of product development and can begin as early as the DESIGN phase of the DfR process (Chapter 7). However most testing is done during the VERIFICATION and VALIDATION stages. In the earlier development phases testing is usually directed at addressing particular design concerns or failure mechanisms. For example a circuit board can be tested to 1000 cycles of thermal shock to estimate the fatigue life of lead-free solder joints or a heavy device can be subjected to random vibration to test the strength of a mounting bracket. Test to failure followed by life data analysis (Chapter 3) would be the best way to explore design limits (Figure 12.8) at the system, sub-system, or component level.

At the VALIDATION stage testing is typically done on a system level in order to confirm that it is ready for production. Due to schedule constraints and time to market pressure at this phase the product is often tested to prove that the design is ‘good enough’ for the expected environment. This is often done as a test to success and would often combine a gamut of tests simulating all possible field environments. Planning such a comprehensive system validation test requires a variety of considerations including understanding of the reliability specifications, field environments, possible failure mechanisms, acceleration models, and other considerations. One example of such a comprehensive test planning document is GMW 3172, a test standard developed by General Motors for testing electrical and electronic components installed on its vehicles (GMW 3172, 2004).

In the ideal scenario all the environmental tests should be applied to the same test units, preferably simultaneously to reflect the effect of combined environments (see CERT, Section 12.3). However due to equipment limitations those tests might have to be conducted sequentially. Test planning where environmental tests are conducted in sequence are very common in the industry, but they present a problem of long test durations when the development schedule is short. On many occasions these tests are conducted in parallel on separate test samples (Figure 12.9). The key to creating an efficient parallel test plan is an understanding of the interactions between failure mechanisms and environments. Failure mechanisms which may potentially have a combined environment effect should be addressed in the same test leg and those, which appear independent in the different legs. The test flow in Figure 12.9 constructed based on GMW 3172 has three paths (test legs).

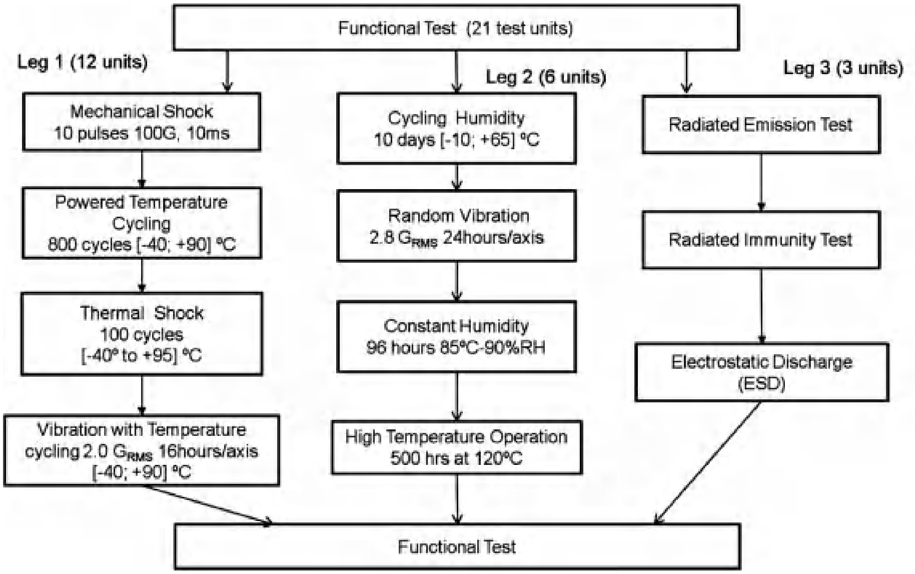

Functional Test (21 test units)
Leg 1 (12 units)
Leg 2 (6 units)
Leg 3 (3 units)
Mechanical Shock 10 pulses 100G, 10ms
Powered Temperature Cycling 800 cycles [-40; +90] °C
Thermal Shock 100 cycles [-40° to +95] °C
Vibration with Temperature cycling 2.0 G_RMS 16hours/axis [-40; +90] °C
Cycling Humidity 10 days [-10; +65] °C
Random Vibration 2.8 G_RMS 24hours/axis
Constant Humidity 96 hours 85°C-90%RH
High Temperature Operation 500 hrs at 120°C
Radiated Emission Test
Radiated Immunity Test
Electrostatic Discharge (ESD)
Functional Test

Figure 12.9 *Example of a parallel test flow for an electronic device.*

Leg 1 addresses durability and has the largest number of units (12) for the reliability demonstration (Chapter 14). Leg 2 mostly addresses potential corrosion, dendrite formation and intermittent failures and Leg 3 is testing for electromagnetic compatibility (EMC).

It is difficult if not impossible to anticipate all the possible interactions between different test environments and their effects on product failures, therefore past experience and/or a comprehensive design of experiment can provide additional data.

12.6 Failure Reporting, Analysis and Corrective Action Systems (FRACAS)

12.6.1 Failure Reporting

FRACAS is an apt acronym for the task of failure reporting, analysis and corrective action. It is essential that all failures which occur during development testing are carefully reported and investigated. It can be very tempting to categorize a failure as irrelevant, or not likely to cause problems in service, especially when engineers are working to tight schedules and do not want to be delayed by filling in failure reports. However, time and costs will nearly always be saved in the long run if the first occurrence of every failure mode is treated as a problem to be investigated and corrected. Failure modes which affect reliability in service can often be tracked back to incidents during development testing, when no corrective action was taken.

A failure review board should be set up with the task of assessing failures, instigating and monitoring corrective action, and monitoring reliability growth. An important part of the board's task is to ensure that the corrective action is effective in preventing any recurrence of failure. The board should consist of:

- The project reliability engineer.
- The designer.
- Others who might be able to help with the solutions, such as the quality engineer, production or test engineer.

The failure review board should operate as a team which works together to solve problems, not as a forum to argue about blame or to consign failure reports to the ‘random, no action required’ category. Its recommendations should be actioned quickly or reported to the project management if the board cannot decide on immediate action, for example if the solution to the problem requires more resources. This approach has much in common with the quality circles method described in Chapter 15.

US MIL-HDBK-781 provides a good description of failure reporting methods. Since a consistent reporting system should be used throughout the programme MIL-STD-781 can be recommended as the basis for this.

These data should be recorded for each failure:

1 Description of failure symptoms, and effect of failure.
2 Immediate repair action taken.
3 Equipment operating time at failure (e.g. elapsed time indicator reading, mileage).
4 Operating conditions.
5 Date/time of failure.
6 Failure classification (e.g. design, manufacturing, maintenance-induced).
7 Report of investigation into failed component and reclassification, if necessary.
8 Recommended action to correct failure mode.
9 Corrective action follow-up (test results, etc.).

Failure report forms should be designed to allow these data to be included. They should also permit easy input to computer files, by inclusion of suitable coding boxes for use by the people using the forms. An example is given in Appendix 5.

Failure data analysis methods are described in Chapter 13.

12.6.2 Corrective Action Effectiveness

When a change is made to a design or to a process to correct a cause of failure, it is important to repeat the test which generated the failure to ensure that the corrective action is effective. Corrective action sometimes does not work. For example, it can have the effect of transferring the problem to the next weakest item in the sequence of stress-bearing items, or the true underlying cause of failure might be more complex than initial analysis indicates. Therefore re-test is important to ensure that no new problems have been introduced and that the change has the desired effect.

Analysis of test results must take account of the expected effectiveness of corrective action. Unless the causes of a failure are very well understood, and there is total confidence that the corrective action will prevent recurrence, 100 % effectiveness should not be assumed.

Questions

1. Describe the concept of integrated test planning. What are the main categories of test that should be included in an integrated test programme for a new design, and what are the prime objectives of each category?
2. List the important information that should be considered in planning a reliability test programme.
3. Identify one major reference standard providing guidance on environmental testing. Identify the major factors to be considered in setting up tests for temperature, vibration, or electromagnetic compatibility.

4. Briefly describe the concept of combined environmental reliability testing (CERT). What are the main environmental stresses you may consider in planning a CERT for (i) a domestic dishwasher electronic controller; (ii) a communications satellite electronic module; (iii) an industrial hydraulic pump?
5. State your reservations concerning the use of standard environmental test specifications in their application to the equipments in question 4.
6. What is ‘accelerated testing’ and what are the main advantages of this type of testing in comparison with non-accelerated tests?
7. Explain why the stresses applied in accelerated stress tests should not necessarily simulate expected in-service levels.
8. How are tests accelerated for (i) mechanical components under fatigue loading and (ii) electronic systems operating at high temperatures? Comment on the methods used for analysis for the test results for each.
9. What is ‘highly accelerated life testing’? Describe the main benefits claimed for this method.
10. Why is it important to ensure that all failures experienced during engineering development are reported? Describe the essential features of an effective failure reporting, analysis and corrective action system (FRACAS).
11. Give examples of ‘foolish failures’ besides those listed in the chapter. Discuss the ways to avoid these types of failures in testing.
12. An electro-dynamic shaker has been set up to vibrate at 12g acceleration level. Compare the peak to peak displacements at 120 Hz and 200 Hz during the sine sweep testing.
13. Discuss what happens to the bathtub curve (Figure 12.7) when stress levels approach the design limits. What happens to the bathtub curve pattern when stresses reach the destruct limits?
14. You are developing a gear box for a wind turbine power generator for an off-shore installation (sea/ocean water). What kind of environments would you consider including in your test flow? Which tests can be run sequentially and which ones can be done in parallel.
15. Would you consider a warranty claims database as FRACAS? What information would warranty return systems have in addition to a typical FRACAS data structure?
16. Would the use of HALT be more beneficial in product development of new technological or in evolutionary designs? Justify your answer.
17. You are developing a test plan for a temperature cycling test of 300 cycles of [T_{MIN}, T_{MAX}]. How would you take into account the size and the weight of your product when determining the duration of temperature dwells at T_{MIN} and T_{MAX} and the transition time between them?

Bibliography

Environmental Stress Screening Guidelines. Institute of Environmental Sciences and Technology (USA).

GMW 3172 (2004) *General Specification for Electrical/Electronic Component Analytical/Development/Validation (A/D/V) Procedures for Conformance to Vehicle Environmental, Reliability, and Performance Requirements*, General Motors Worldwide Engineering standard. Available at www.global.ihs.com (Accessed 20 March 2011).

Harris, C., Piersol, A. and Paez, T. (eds) (2010) *Shock and Vibration Handbook*, 6th edn, McGraw-Hill.

IEC 61 000. *Electromagnetic Compatibility*.

ISO/IEC 60068. *Environmental Testing*.

ISO/IEC 61163. *Reliability Stress Screening*.

McLean, H. (2009) *HALT, HASS and HASA Explained: Accelerated Reliability Techniques*, American Society for Quality.

O’Connor, P.D.T. (2001) *Test Engineering*, Wiley.

Steinberg, D. (2000) *Vibration Analysis for Electronic Equipment*, 3rd edn, Wiley.

UK Defence Standard 07-55. *Environmental Testing*. HMSO.

US MIL-HDBK-781. *Reliability Testing for Engineering Development, Qualification and Production*. Available from the National Technical Information Service, Springfield, Virginia.

US MIL-STD-462D. *Measurement of Electromagnetic Interference Characteristics*. Available from NTIS, Springfield, Virginia.

US MIL-STD-810. *Environmental Test Methods*. Available from the National Technical Information Service, Springfield, Virginia.

13

Analysing Reliability Data

13.1 Introduction

This chapter describes a number of techniques, further to the probability plotting methods described in Chapter 3, which can be used to analyse reliability data derived from development tests and service use, with the objectives of monitoring trends, identifying causes of unreliability, and measuring or demonstrating reliability.

Since most of the methods are based on statistical analysis, the caution given in Section 2.17 must be heeded, and all results obtained must be judged in relation to appropriate engineering and scientific knowledge.

13.2 Pareto Analysis

As a first step in reliability data analysis we can use the Pareto principle of the 'significant few and the insignificant many'. It is often found that a large proportion of failures in a product are due to a small number of causes. Therefore, if we analyse the failure data, we can determine how to solve the largest proportion of the overall reliability problem with the most economical use of resources. We can often eliminate a number of failure causes from further analysis by creating a Pareto plot of the failure data. For example, Figure 13.1 shows failure data on a domestic washing machine, taken from warranty records. These data indicate that attention paid to the program switch, the outlet pump, the high level switch and leaks would be likely to show the greatest payoff in warranty cost reduction. However, before committing resources it is important to make sure that the data have been fully analysed, to obtain the maximum amount of information contained therein. The data in Figure 13.1 show the parts replaced or adjusted.

In this case further analysis of records reveals:

1. **For the program switch:** 77 failures due to timer motor armature open-circuit, 18 due to timer motor end bearing stiff, 10 miscellaneous. Timer motor failures show a decreasing hazard rate during the warranty period.
2. **For the outlet pump:** 79 failures due to leaking shaft seal allowing water to reach motor coils, 21 others. Shaft seal leaks show an increasing hazard rate.
3. **For the high level switch:** 58 failures due to failure of spot weld, allowing contact assembly to short to earth (decreasing hazard rate), 10 others.

Practical Reliability Engineering, Fifth Edition. Patrick D. T. O'Connor and Andre Kleyner.

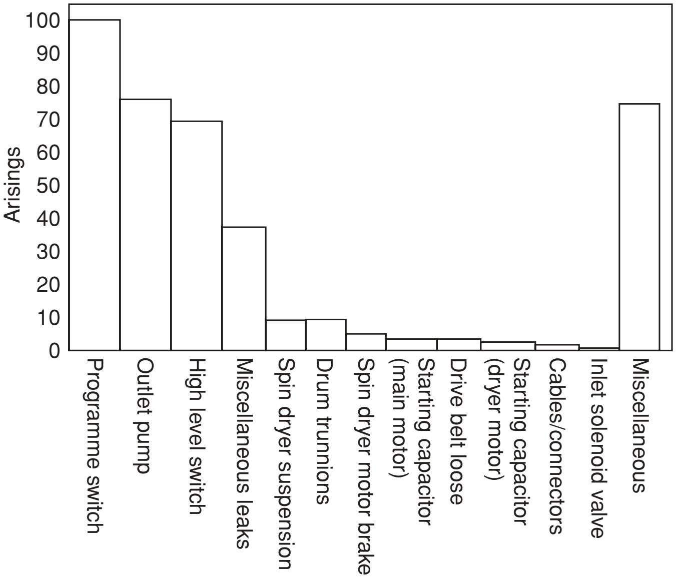


Figure 13.1 *Pareto plot of failure data.*

These data reveal definite clues for corrective action. The timer motor and high level switch appear to exhibit manufacturing quality problems (decreasing hazard rate). The outlet pump shaft leak is a wear problem (increasing hazard rate). However, the leak is made more important because it damages the pump motor. Two types of corrective action might be considered: reorientation of the pump so that the shaft leak does not affect the motor coils and attention to the seal itself. Since this failure mode has an increasing rate of occurrence relative to equipment age, improving the seal would appreciably reduce the number of repair calls on older machines. Assuming that corrective action is taken on these four failure modes and that the improvements will be 80 % effective, future production should show a reduction in warranty period failures of about 40 %.

The other failure modes should also be considered, since, whilst the absolute payoff in warranty cost terms might not be so large, corrective action might be relatively simple and therefore worthwhile. For example, some of the starting capacitor failures on older machines were due to the fact that they were mounted on a plate, onto which the pump motor shaft leak and other leaks dripped. This caused corrosion of the capacitor bodies. Therefore, rearrangement of the capacitor mounting and investigation into the causes of leaks would both be worth considering.

This example shows the need for good data as the basis for decision-making on where to apply effort to improve reliability, by solving the important problems first. The data must be analysed to reveal as much as possible about the relative severity of problems and the likely causes. Even quite simple data, such as a brief description of the cause of failure, machine and item part number, and purchase date, are often enough to reveal the main causes of unreliability. In other applications, the failure data can be analysed in relation to contributions to down-time or by repair cost, depending upon the criteria of importance.

13.3 Accelerated Test Data Analysis

Failure and life data from accelerated stress tests can be analysed using the methods described in Chapters 2, 3, 5 and 11. If the mechanism is well understood, for example material fatigue and some electronic degradation processes, then the model for the process can be applied to interpret the results and to derive reliability or life values at different stress levels. Also, the test results can be used to determine or confirm a life model. The most commonly used stress–life relationship models include:

$$\textit{Exponential Model}: \quad \textit{Life} = A\mathrm{e}^{-X(Stress)} \tag{13.1}$$

In this model there is an exponential relationship between the life (or other performance measure) and the stress. *A* is an empirical constant and *X* can be a constant or function describing this relationship (more later in this chapter). Depending on the failure mechanism, stress can be a maximum application temperature, temperature excursion ΔT during thermal cycling, vibration G_{RMS}, humidity, voltage and other external or internal parameters. Failure mechanisms of electronic components typically follow this relationship over most of their life for dependence on temperature. Some materials follow this type of relationship as a function of weathering.

$$\textit{Power model}: \quad \textit{Life} = A(\mathrm{Stress})^{-X} \tag{13.2}$$

In this model there is a simple power relationship between the performance measure and the stress variable. Failure mechanisms of electronic components typically follow this relationship over most of their life for dependence on voltage. Electric motors follow a similar form for dependence on stresses other than temperature. Mechanical creep of some materials may follow a form such as this.

There are other models including the mix of exponential and power model or other physical laws. Some of these models are covered in detail in this chapter.

13.4 Acceleration Factor

Most life-stress models (see (13.1) and (13.2)) contain an empirical constant *A*, which is usually unknown a priori and can only be obtained by testing for a specific failure mechanism under specific test conditions. Also the measure of life will affect the value of *A*, for example product's life can be measured as MTBF, MTTF, B_{10}-life, Weibull characteristic life η or even as time to the first failure. Therefore, it is more common to analyse test data based on *acceleration factor, AF*.

$$\textit{Acceleration Factor}: \quad AF = \frac{L_{Field}}{L_{Test}} \tag{13.3}$$

Where L_{Field} is the product life at the field stress level and L_{Test} is the life at the test (accelerated) stress level. Other literature may use different nomenclature, such as L_U (use) for field life and L_S (stress) or L_A (accelerated) for test life. In (13.3) the acceleration factor is independent of the empirical constant *A* and of the measure of product life. The relationship (13.3) assumes the same failure mechanism caused by the same type of environment only at a different stress level.

The effect of accelerated test conditions on product life has been illustrated in Chapter 12, bathtub curve, Figure 12.7, therefore acceleration factor is used to calculate the required test time based on the expected product field life. Acceleration factor effect on the key reliability functions is shown in Table 13.1. The ability to calculate the acceleration factor is critical to the development of an efficient test plan adequately reflecting the field level stress conditions.

Table 13.1 *Acceleration factor relationship with reliability functions.*

Time to failure	$t_{Field} = t_{Test} AF$
Failure probability density function p.d.f.	$f_{Field}(t) = \frac{1}{AF} f_{Test}\left(\frac{t}{AF}\right)$
Failure probability cdf	$F_{Field}(t) = F_{Test}\left(\frac{t}{AF}\right)$
Reliability	$R_{Field}(t) = R_{Test}\left(\frac{t}{AF}\right)$
Failure or hazard rate	$h_{Field}(t) = \frac{1}{AF} h_{Test}\left(\frac{t}{AF}\right)$

Example 13.1

Mean time to failure (MTTF) for a microprocessor at the field level temperature of 60 °C is 20 years assuming 2 hours of operation per day. Continuous test at 120 °C produced failure modes consistent with field failures and the MTTF = 1000 hours. Calculate the acceleration factor.

In order to calculate the field life we need to convert 20-year life into the number of operating hours: $L_{Field} = 20$ years × 365 days/year × 2 hours/day = 14 600 hours

$$\text{Therefore per (13.3): } AF = \frac{L_{Field}}{L_{Test}} = \frac{14600}{1000} = 14.6$$

13.5 Acceleration Models

This section discusses the acceleration models attributed to various stresses such as temperature, humidity, vibration, voltage and current. All those models contain empirical constants most of which are generic values for the specific models and specific failure mechanisms. These generic constants are obtained from past experience or past testing. Even though these constants have been widely used in the industry for many years, it is always best to obtain these empirical values by conducting test to failure experiments at different stress levels and fitting the obtained life data into a model. The development of acceleration models based on accelerated test data is discussed in Section 13.7.

13.5.1 Temperature and Humidity Acceleration Models

Following are the most commonly used acceleration models involving stresses caused by temperature and humidity.

13.5.1.1 Arrhenius Model

Based on the Arrhenius Eq. (8.5), life is non-linear in the single stress variable, temperature (T). It describes many physical and chemical temperature-dependent processes, including diffusion and corrosion. The

Arrhenius model has various applications, but it is most commonly used to estimate the acceleration factor for electronic component operation at a constant temperature:

$$\textit{Arrhenius model}: AF = \exp\left[\frac{E_A}{k}\left(\frac{1}{T_{Field}} - \frac{1}{T_{Test}}\right)\right] \tag{13.4}$$

E_A = activation energy for the process.
$k = 8.62\times10^{-5}$ eV/K (Boltzmann constant).
T_{Field} and T_{Test} = absolute Kelvin temperatures at field and test level respectively.

The original definition of Arrhenius' activation energy E_A came from chemistry and physics. It corresponds to the minimum energy required for the electron to move to a different energy level and begin a chemical reaction. It is important to note that even though E_A has a specific meaning as an atomic or material property, in the Arrhenius equation it simply becomes an empirical constant appropriate for use with a particular failure mechanism. This means that at a component level activation energy is not a simple material property but a fairly complex function of geometry, material properties, technology, interconnections and other factors. Therefore it is sometimes referred as E_{AA} – 'apparent activation energy' (JEDEC, 2009).

There are no predetermined generic values of E_A for the specific failure mechanisms due to variation in parts characteristics, therefore different reference sources list different values for the activation energies. However some recommended values suggested in the literature on the subject (see, e.g. Ohring, 1998) are listed in Table 13.2.

Please note, that the only reliable way to obtain the value of E_A is to conduct a series of accelerated tests and record failure times as a function of temperature (see Example 13.4). It is also important to note that the Arrhenius acceleration factor is very sensitive to the value of E_A due to its exponential nature, therefore the accuracy of estimating E_A is important.

Table 13.2 *Commonly used activation energy values for different failure mechanisms.*

Failure Mechanism	Activation Energy, E_A (eV)
Gate oxide defect	0.3–0.5
Bulk silicon defects	0.3–0.5
Silicon junction defect	0.6–0.8
Metallization defect	0.5
Au-Al intermetallic growth	1.05
Electromigration	0.6–0.9
Metal corrosion	0.45–0.7
Assembly defects	0.5–0.7
Bond related	1.0
Wafer fabrication (chemical contamination)	0.8–1.1
Wafer fabrication (silicon/crystal defects)	0.5–0.6
Dielectric breakdown, field > 0.04 micron thick	0.3
Dielectric breakdown, field <= 0.04 micron thick	0.7
Adhesive tack: bonding-debonding	0.65–1.0

13.5.1.2 Eyring Model

The Eyring model is usually applied to combine the effect of more than one independent stress variable assuming no interactions between the stresses. Many other models are simplified versions of the Eyring model. A generic form of the Eyring equation is

$$Life = A\exp\left[\frac{E_A}{kT}Y_1(Stress1)Y_2(Stress2)\right] \tag{13.5}$$

$Y_1(Stress1)$, $Y_2(Stress2)$ are the factors for other applied stresses, such as temperature, humidity, voltage, current, vibration, and so on.

A form of the Eyring model for the influence of voltage, V in addition to temperature is:

$$Life = AV^{-B}\exp\left[\frac{E_A}{kT}+\left(C+\frac{D}{T}\right)V\right] \tag{13.6}$$

To use this formula, four constants, A, B, C, D must be known or estimated (see NIST, 2006).

13.5.1.3 Peck Temperature-Humidity Model

As a special case of the Eyring model, Peck's equation (Peck, 1986) is probably the most commonly used acceleration model addressing the combined effect of temperature and humidity. According to Peck, the acceleration factor correlating product life in the field with test duration can be expressed as:

$$AF = \left(\frac{RH_{Test}}{RH_{Field}}\right)^m \exp\left[\frac{E_A}{k}\left(\frac{1}{T_{Field}}-\frac{1}{T_{Test}}\right)\right] \tag{13.7}$$

where: $m =$ humidity power constant, typically ranging between 2.0 and 4.0
$RH =$ Relative humidity measured as percent

Despite the limited applications and varying accuracy, Peck's model has been often utilized to calculate field-to-test ratios for a wide variety of products and failure modes. It is important to note that this model can only be applied to wear-out failure mechanisms, including electromigration, corrosion, dielectric breakdown and dendritic growth (Kleyner, 2010a). It has also been applied to tin whisker growth in lead-free electronics, although with varying degrees of success.

13.5.1.4 Lawson Temperature-Humidity Model

A lesser known temperature-humidity model, which is based on the water absorption research presented in Lawson (1984) is often applied its modified version:

$$AF = \exp\left[\frac{E_A}{k}\left(\frac{1}{T_{Field}}-\frac{1}{T_{Test}}\right)\right]\exp\left[b\left(RH^2_{Test}-RH^2_{Field}\right)\right] \tag{13.8}$$

b is an empirical humidity constant based on water absorption. In many electronics applications involving silicon chips $b = 5.57\times10^{-4}$ although it is best when b is determined based on test results.

13.5.1.5 Tin-Lead Solder

Solder joint fatigue has always been a source of concern in the electronics industry (Chapter 9). The well-known Coffin-Manson relationship provides a relationship between life in thermal cycles, *N*, and plastic strain range $\varDelta\gamma_p$.

$$N_f\left(\Delta\gamma_p\right)^m = \text{Constant} \tag{13.9}$$

m is an empirical fatigue constant, observed to be about 2.0–3.0 for eutectic tin-lead solder.

During thermal cycling the strain range caused by the mismatch of the coefficients of thermal expansion between solder and other materials is proportional to the cycling temperature excursion $\varDelta \mathrm{T} = \mathrm{T}_{\mathrm{Max}} - \mathrm{T}_{\mathrm{Min}}$. Based on (13.9) the acceleration factor can be approximated by:

$$AF = \left(\frac{\Delta T_{Test}}{\Delta T_{Field}}\right)^m \tag{13.10}$$

In the case of low cycle fatigue, the acceleration factor is typically applied to the number of thermal cycles rather than the temperature exposure time. There are extensions of the Coffin-Manson model which account for the effect of temperature transition during thermal cycling (see Norris and Landzberg, 1969). However, there is no conclusive evidence that faster temperature transition has a significant effect on fatigue life of tin-lead solder joints.

13.5.1.6 Lead-Free Solder

As mentioned in Chapter 9, lead-free solder mechanical behaviour is different from that of tin-lead including its low cycle fatigue properties. Simplified lead free model is based on the tin-lead coffin Manson equation (13.10) with the fatigue constant $m = 2.6 - 2.7$. However, the research showed that lead-free solder acceleration factors are also influenced by the variables other than $\varDelta T$, such as the maximum and minimum cycling temperatures, dwell times (both at maximum and minimum temperatures) and to some degree the temperature transition rate.

As mentioned in Chapter 9, lead-free solder has not been studied for nearly as long as tin-lead, therefore it will take time before technical knowledge about lead-free solder reaches maturity. Several empirical lead-free acceleration models have been developed in the last years, especially for SAC305 solder. Some of these models are covered in Pan *et al.* (2005), Clech *et al.* (2005), Salmela (2007) and several more are still in the process of development.

13.5.2 Voltage and Current Acceleration Models

A simple inverse power law model often used for capacitors has only voltage *V* dependency and takes the form:

$$AF = \left(\frac{V_{Test}}{V_{Field}}\right)^B \tag{13.11}$$

An alternative exponential voltage model takes the form of:

$$AF = \exp\left[B\left(V_{Test} - V_{Field}\right)\right] \tag{13.12}$$

Where B is voltage acceleration parameter (typically determined from an experiment). The exponential voltage model (13.12) combined with Arrhenius can be applied to time dependent dielectric breakdown (TDDB) (JEDEC, 2009) in the form of:

$$AF = \exp\left[B\left(V_{Test} - V_{Field}\right)\right]\exp\left[\frac{E_A}{k}\left(\frac{1}{T_{Field}} - \frac{1}{T_{Test}}\right)\right] \tag{13.13}$$

Another form of the Eyring equation can be used to model electromigration (Section 9.3.1.5). The ionic movement is accelerated by high temperatures and increased current density, J.

$$AF = \left(\frac{J_{Test}}{J_{Field}}\right)^n \exp\left[\frac{E_A}{k}\left(\frac{1}{T_{Field}} - \frac{1}{T_{Test}}\right)\right]$$

The commonly used values for Al or Al-Cu alloys $E_A = 0.7 – 0.9$ eV and $n = 2$.

More acceleration models for electronic devices can be found in SEMATECH, 2000.

13.5.3 Vibration Acceleration Models

Most vibration models are based on the *S-N* curve introduced in Chapter 8. The relationship between peak stress σ and the number of cycles to failure, N can be expressed as $N\sigma^b =$ Constant (high cycle fatigue). Assuming the linear relationship between the stress and acceleration G during vibration, the model takes the form:

$$\begin{aligned} &Sinusoidal\ Vibration: AF = \left(\frac{G_{Peak-Test}}{G_{Peak-Filed}}\right)^b \\ &Random\ Vibration: AF = \left(\frac{G_{RMS\ Test}}{G_{RMS\ Filed}}\right)^b \end{aligned} \tag{13.14}$$

Fatigue exponent b is the slope of the *S-N* line in the log-log scale (Chapter 8, Figure 8.5) and has different values for different materials (see Steinberg, 2000 for fatigue curves). $b = 4.0 – 6.0$ is very common for electronics related failure (electrical contacts, component leads, mounting brackets, etc.) It is typical for sinusoidal vibration to measure life in vibration time or a number of cycles. For random vibration it is the test time or the number of stress reversals (see Steinberg, 2000).

Example 13.2

An electrical insulator is rated for normal use at 12 kV. Prior tests of a sample have suggested this insulator will operate over 30 kV and a stress-life exponent of the power model was found to be $N = 5.5$. How long must one run an accelerated life test at 30 kV to demonstrate an equivalent B_{10} life (based upon voltage only) if the B_{10} life at 12 kV is desired to be at least 25 years?

Let life, $L = A(V)^{-N}$ represent the typical stress life relationship (inverse power law).

Applying (13.11) to calculate the acceleration factor due to voltage only:

$$AF = \left(\frac{30kV}{12kV}\right)^{5.5} = 154.41$$

Therefore:

$$\text{Test time at } 30\ k\text{V} = \frac{L_{25yrs}}{AF} = \frac{25\,years\ \times 8760\,\frac{h}{year}}{154.41} = 1418.3\,h$$

Extrapolation of accelerated test results to expected in-service conditions can be misleading if the test stresses are much higher, since different failure mechanisms might be stimulated. That increases the probability of irrelevant or 'foolish' failures. This is particularly the case if very high stresses, particularly combined stresses, are applied, as in HALT (Chapter 12). It is important that the primary objective of the test is understood: whether it is to determine or confirm a life characteristic, or to help to create designs that are inherently failure free.

Life characteristics such as those listed above can only be analysed when the data represent a single (or dominant) failure mode. This significantly complicates the analysis of failures of assemblies or systems, when several different failure modes might be present. In those cases, different failure modes need to be addressed by a series of environmental tests targeting individual failure mechanisms as presented in Section 12.5. Life-stress models then should be applied to every stress environment in that test programme in order to calculate the appropriate test durations (see the next section).

Life data analysis methods (Chapter 3) at different stress levels can also be used for analysing such data when sufficient data are available (see Section 13.7).

13.6 Field-Test Relationship

Determining what a particular test represents in terms of the field life of a product and vice versa can often be a complicated task. It combines the use of the appropriate acceleration models with knowledge of the stress and usage conditions in the field. As mentioned before, an acceleration model should be applied only to one failure mechanism at a time, however in reality there are almost always more than one potential failure mechanism present. Therefore it is practicable to address the most dominant failure mechanism associated with a particular environment for a particular test in the testing programme (see Chapter 12, Figure 12.9). Additionally, the acceleration factor calculations should account for the failure mechanism which will make the product fail fastest. For example, amongst possible failure mechanisms caused by a constant temperature, the one with the lowest E_A (see Table 13.2) should be selected for the test time calculations because it would yield the most conservative (lowest) acceleration factor.

Design of the test plan simulating the expected product field life should include the following steps:

- Evaluate the stress level and the usage profile for the field environment.
- Select the appropriate acceleration model(s) for the test based on the expected failure mechanism.
- Define the appropriate stress level and test duration based on the test equipment capability and the maximum allowable stress level to avoid 'foolish' failures.
- Calculate the acceleration factor and duration of the test based on field stress level and the environmental exposure time during the life time operation.
- Repeat this procedure for each applicable test/environment.

Example 13.3

Develop the thermal cycling test for an automotive controller designed to operate for 10 years and mounted under the hood of a passenger car. First, we need to determine the typical field environment and the usage for this electronic controller. It has been established that two 'cold' temperature cycles per day are typical for a 90–98 percentile user of a passenger car. One 'cold' cycle represents a vehicle start after being parked at below-freezing temperature and followed by at least 30 minutes of continuous driving. Therefore the 10 year field exposure can be calculated as:

$$10 \text{ yrs} \times 365 \text{ days/year} \times 2 \text{ cycles/day} = 7300 \text{ cycles}$$

Solder fatigue is often considered to be the dominant failure mechanism for automotive electronics during thermal cycling. Automotive thermal cycling is typically caused by the engine heat combined with the internal heat dissipation of the electronic unit. Therefore the Coffin-Manson model (13.8) is selected to calculate the acceleration factor with $m = 2.5$ for tin-lead solder. To calculate the acceleration factor we need to know the temperature excursion ΔT for both test and field conditions. Based on the environmental chamber capability and previous experience [−40; +125] °C cycle has been selected. Field study showed that the internal temperature of the under-the-hood-mounted electronics operated in severe climates grows by up to $\Delta T_{Field} = 70$ °C during driving. Substituting these numbers into (13.8) produces:

$$T_{Test} = 7300\left(\frac{70}{125-(-40)}\right)^{2.5} = 856\,cycles$$

Therefore, 856 thermal cycles of [-40; +125] °C will represent one product life of 10 years for an automotive controller mounted under the hood of a vehicle.

13.7 Statistical Analysis of Accelerated Test Data

As mentioned before, the acceleration models discussed in this chapter have limited accuracy because their empirical equations are based on generic data. Therefore, it is always better to develop an acceleration model based on experimental data rather than generic equations. There is commercially available software designed to analyse accelerated life data. When sufficient data is available software packages such as ReliaSoft ALTA® or WinSMITH® can fit a statistical distribution to a life data set at each stress level and model the resulting life-stress relationship.

An effective quantitative accelerated life test produces life data obtained at two or more stress levels that cause the product to fail. Analysing this data, which contains failed and not failed (suspended) parts we can estimate the parameters for the lifetime distribution that best fits the data at each stress level (e.g. Weibull, exponential, lognormal, etc.). The life-stress relationship then can be used to estimate the pdf at the field use (not accelerated) stress level based on the characteristics of the distributions at each accelerated stress level. A simplified version of this approach is shown in Table 13.1 for the pdf.

The data analyst must first choose a life-stress relationship which is appropriate for the test failure modes and find the best fit to the data being analysed. The appropriate life characteristic can be chosen based on the statistical distribution or other criteria. For example, for the Weibull distribution, the scale parameter η is considered to be stress- dependent. Therefore, the life-stress model for data that fits the Weibull distribution is assigned to η. For the exponential distribution it is MTTF, for the normal it is the mean life, and so on.

As mentioned before, life-stress relationships are specific to the types of failures; therefore it is important that the failure mechanisms remain the same at different stress levels. The best way to verify, this is to perform failure analysis, which often involves cross-sectioning of the failed part. The analytical alternative (though less preferred than failure analysis) is to compare the Weibull slopes β at different stress levels. If the β

Table 13.3 *Accelerated test results (Example 13.4).*

60 °C (333K)		80 °C (353K)		100 °C (373K)	
68 h	Fail	55 h	Fail	13 h	Fail
127 h	Fail	63 h	Fail	15 h	Fail
186 h	Fail	80 h	Fail	30 h	Fail
205 h	Fail	126 h	Fail	31 h	Fail
250 h	Suspended	137 h	Fail	47 h	Fail
250 h	Suspended	192 h	Fail	73 h	Fail
250 h	Suspended	240 h	Fail	95 h	Fail
250 h	Suspended	250 h	Suspended	98 h	Fail

slope remains constant over the different stress levels of acceleration, it is a good indicator that the product is experiencing the same or similar failure modes.

Example 13.4

In order to expedite product development, 24 electronic parts, which are designed to operate at field temperatures up to 40 °C have been subjected to accelerated testing. The first group of eight samples have been tested at 60 °C (333K), second group at 80 °C (353K) and third at 100 °C (373K). The test was terminated after 250 hours. The time-to-failure and time-to-suspension data obtained during this test are presented in Table 13.3.

Assuming the Arrhenius temperature model, calculate the activation energy E_A and determine the life-stress relationship for this device. Based on the accelerated test data, what is the reliability of this part after 100 hours of operation at the field temperature of 40 °C (313K)?

First step is to run life data analysis at each stress level and verify that the parts have the same failure modes at all three temperatures. Figure 13.2 shows the Weibull plots for each group of parts. It shows three approximately equal Weibull slopes, which confirms the consistency of the failure mechanisms at all three stress levels.

The next step is to model the life-stress relationship. In this case we will make use of the ReliaSoft ALTA® software which estimates the parameters of the Weibull distributions at each stress level and intrapolates them to the field use level Figure 13.3.

Figure 13.3 shows the distributions at each stress level and model the median life as the function of temperature. Based on the ALTA® analysis (Figure 13.3), the activation energy $E_A = 0.476$, $\beta = 1.6693$ and the life stress relationship is:

$$Life = \eta(T) = C \exp\left(\frac{B}{T}\right) = 2.27 \times 10^{-5} \exp\left(\frac{5528.78}{T}\right)$$

Characteristic life at the use temperature is $\eta(313\text{K}) = 1062.6$ h, therefore based on Weibull equation:

$$R(100\,h) = \exp\left[-\left(\frac{100}{1062.6}\right)^{1.6693}\right] = 0.9808$$

Please note that life-stress relationship in this example could be derived using other analytical tools. Characteristic life values calculated using life data analysis (Figure 13.2) at different stress levels $\eta(333\text{K}) = 310.0$,

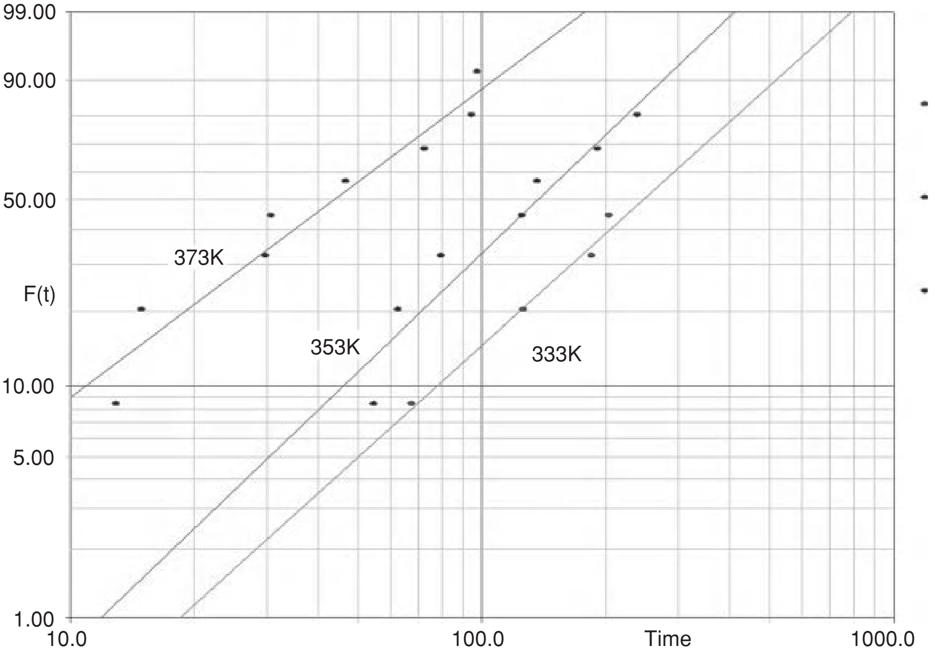


Figure 13.2 *Weibull plot at three stress levels (Weibull++®) (reproduced by permission of ReliaSoft.*

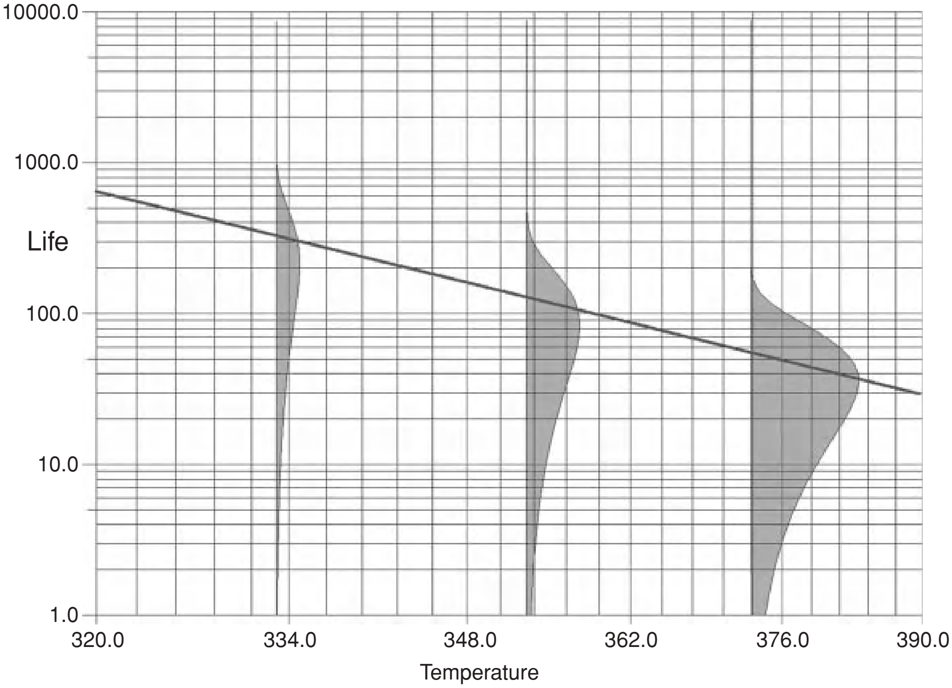


Figure 13.3 *Life vs. stress plot generated with ALTA® software (reproduced by permission of ReliaSoft).*

$\eta(353\text{K}) = 169.7$, $\eta(373\text{K}) = 57.6$ can be fitted with the Arrhenius model using Excel spreadsheet. However, specially designed software, such as ALTA® can complete this task more efficiently and with higher accuracy.

13.8 Reliability Analysis of Repairable Systems

13.8.1 Failure Rate of a Repairable System

Chapter 3 described methods for analysing data related to the time to first failure. The distribution function of times to first failure are obviously important when we need to understand failure processes of parts which are replaced on failure, or when we are concerned with survival probability, for example, for missiles, spacecraft or underwater telephone repeaters.

However, for repairable systems (Chapters 2 and 6), which represent the majority of everyday reliability experience, the distribution of times to first failures are much less important than is the *failure rate* or *rate of occurrence of failures* (ROCOF) of the system.

Any repairable system may be considered as an assembly of parts, the parts being replaced when they fail. The system can be thought of as comprising 'sockets' into which non-repairable parts are fitted. We are concerned with the pattern of successive failures of the 'sockets'. Some parts are repaired (e.g. adjusted, lubricated, tightened, etc.) to correct system failures, but we will consider first the case where the system consists only of parts that are replaced on failure (e.g. most electronic systems). Therefore, as each part fails a new part takes its place in the 'socket'. If we ignore replacement (repair) times, which are usually small in comparison with standby or operating times, and if we assume that the time to failure of any part is independent of any repair actions, then we can use the methods of event series analysis in Chapter 2 to analyse the system reliability.

Consider the data of Example 2.19, Section 2.15.1. The interarrival and (chronologically ordered) arrival values between successive component failures were as shown in columns 1 and 2:

1	2	3
X_i	Chronological x_i	Ranked x_i
175	175	12
21	196	14
108	304	21
111	415	23
89	504	38
12	516	47
102	618	51
23	641	89
38	679	102
47	726	108
14	740	111
51	791	175

Example 2.19 showed that the failure rate was increasing, the interarrival values tending to become shorter. In other words, the interarrival values are not IID. If, however, we had not performed the centroid test and assumed that the data were IID, we might order the data in rank order (column 3) and plot on probability paper. These are shown plotted on Weibull paper in Figure 13.4 (Line A). The plot shows an apparently exponential

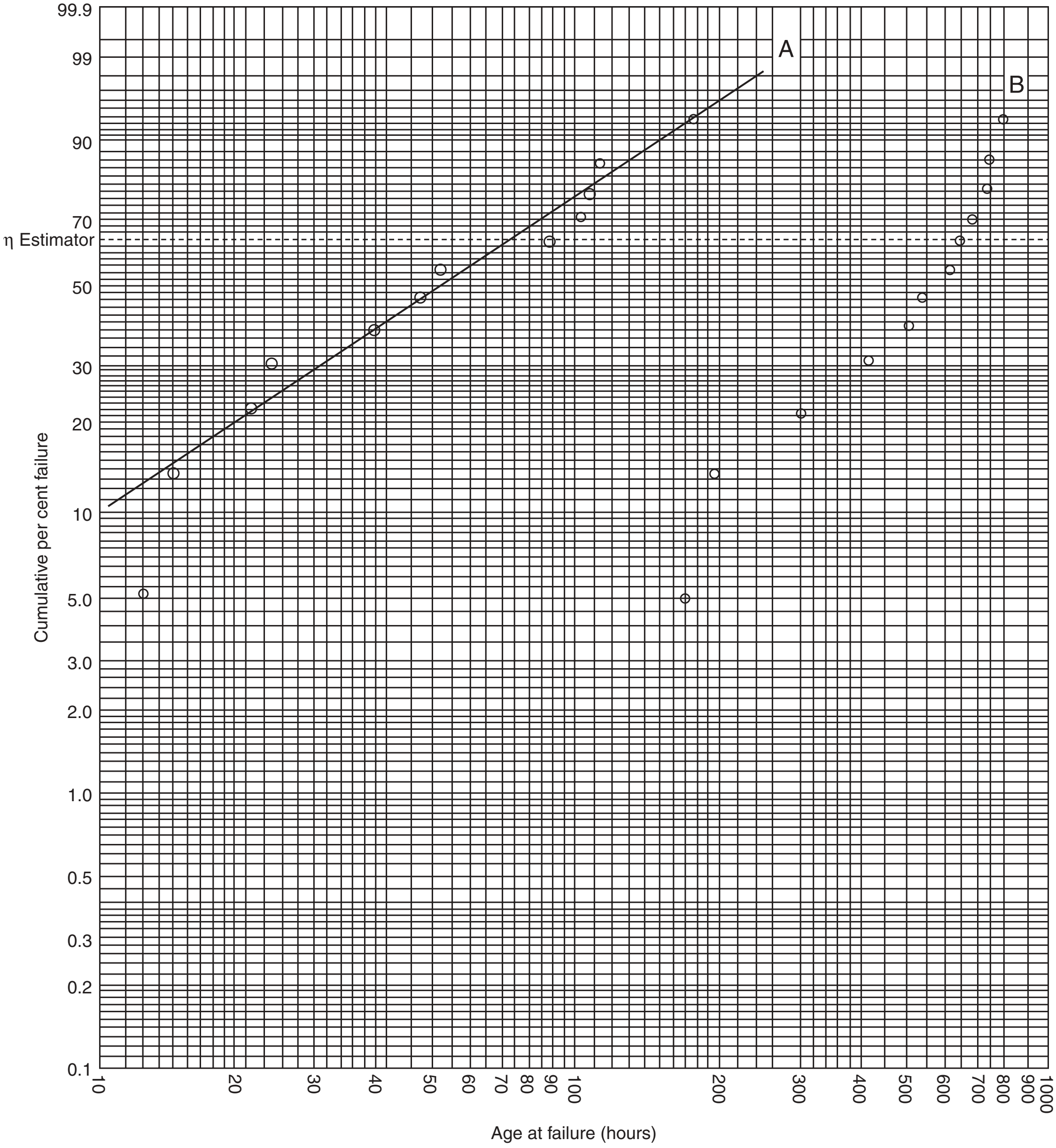


Figure 13.4 *Plotted data of Example 2.19.*

component life distribution. This is obviously a misleading result, since there is clearly an increasing failure rate trend for the 'socket' when the data are studied chronologically.

This example shows how important it is for failure data to be analysed correctly, depending on whether we need to understand the reliability of a non-repairable part or of a repairable system consisting of 'sockets' into which parts are fitted. The presence of a trend when the data are ordered *chronologically* shows that times to failure are not IID, and ordering by magnitude, which implies IID, will therefore give misleading results. *Whenever failure data are reordered* all trend information is ignored. The appropriate method to apply is trend (time series) analysis.

We can derive the system reliability over a period by plotting the cumulative times to failure in chronological order (column 2) rather than in rank order. This is shown in Figure 13.4 (Line B). It shows the progressively increasing failure rate (though the 'socket' times to failure are not Weibull-distributed).

13.8.2 Multisocket Systems

Now we will consider a more typical system, comprised of several parts which exhibit independent failure patterns. Each part fills a 'socket'. The failure pattern of such a system, comprising six 'sockets', is shown in Figure 13.5.

Socket 1 generates a high, constant rate of system failures. Socket 2 generates an increasing rate of system failures as the system ages, and so on. The combined failure rate can be seen on the bottom line. The estimate of U (see the centroid test Chapter 2 Eq. (2.46)) for each part and for the system is shown. When U is negative (i.e. negative process trend), it denotes a 'happy' socket, with an increasing inter-arrival time between failures (decreasing failure rate, DFR). A positive value of U indicates a 'sad' socket (increasing failure rate, IFR).

If there are no perturbations (which will be discussed below) the failure rate will tend to a constant value after most parts have been replaced at least once, regardless of the failure trends of the sockets (see Example 2.19) This is one of the main reasons why the constant failure rate, CFR assumption has become so widely used for systems, and why part hazard rate has been confused with failure rate. However, the time by which most parts have been replaced in a system is usually very long, well beyond the expected life of most systems.

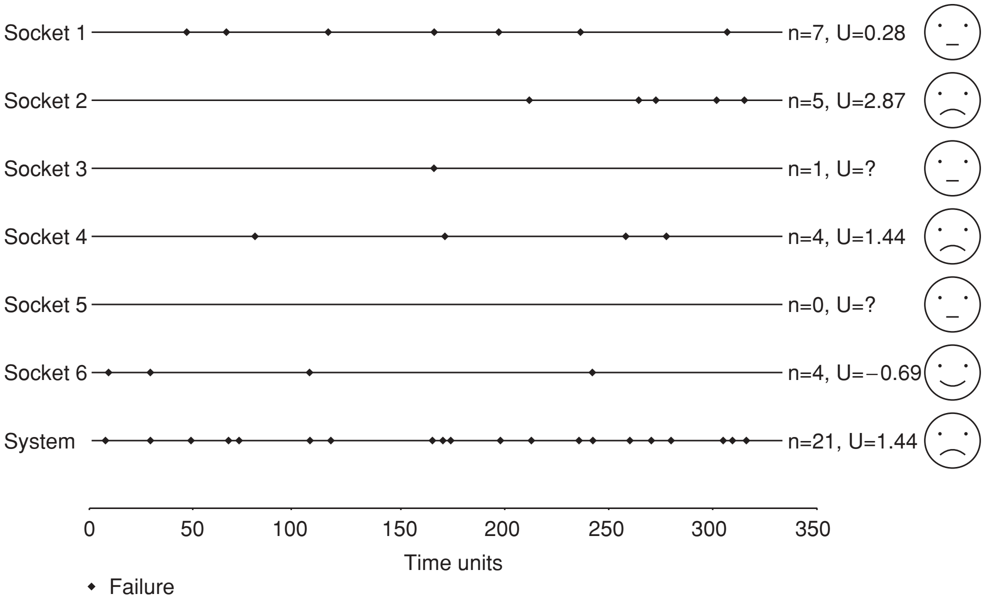


Figure 13.5 *The failure pattern of a multisocket system.*

If part times to failure (in a series system, see Chapter 6) are independently and identically exponentially distributed (IID exponential) the system will have a CFR which will be the sum of the reciprocals of the part mean times to failure, that is

$$\lambda_{\mathrm{s}} = \sum_{1}^{n} \frac{1}{x_i}$$

The assumption of IID exponential for part times to failure within their sockets in a repairable system can be very misleading. The reasons for this are (adapted from Ascher and Feingold, 1984 with permission):

1 The most important failure modes of systems are usually caused by parts which have failure probabilities which increase with time (wearout failures).
2 Failure and repair of one part may cause damage to other parts. Therefore times between successive failures *are not necessarily independent*.
3 Repairs often do not ‘renew’ the system. Repairs are often imperfect or they introduce other defects leading to failures of other parts.
4 Repairs might be made by adjustment, lubrication, and so on, of parts which are wearing out, thus providing a new lease of life, but not ‘renewal’, that is, the system is not made as good as new.
5 Replacement parts, if they have a decreasing hazard rate, can make subsequent failure initially more likely to occur.
6 Repair personnel learn by experience, so diagnostic ability (i.e. the probability that the repair action is correct) improves with time. Generally, changes of personnel can lead to reduced diagnostic ability and therefore more reported failures.
7 Not all part failures will cause system failures.
8 Factors such as on–off cycling, different modes of use, different system operating environments or different maintenance practices are often more important than operating times in generating failure-inducing stress.
9 Reported failures are nearly always subject to human bias and emotion. What an operator or maintainer will tolerate in one situation might be reported as a failure in another, and perception of failure is conditioned by past experience, whether repair is covered by warranty, and so on. Wholly objective failure data recording is very rare.
10 Failure probability is affected by scheduled maintenance or overhaul. Systems which are overhauled often display higher failure rates shortly after overhaul, due to disturbance of parts which would otherwise not have failed. If there is a post-overhaul test period before the system is returned to service, many of these failures might be repaired then. The failure data might or might not include these failures.
11 Replacement parts are not necessarily drawn from the same population as the original parts – they may be better or worse.
12 System failures might be caused by parts which individually operate within specification (i.e. do not fail) but whose combined tolerances cause the system to fail.
13 Many reported failures are not caused by part failures at all, but by events such as intermittent connections, improper use, maintainers using opportunities to replace ‘suspect’ parts, and so on.
14 Within a system not all parts operate to the overall system cycle.

Any practical person could add to this list from his or her own experience. The factors listed above often predominate in systems to be modelled and in collected reliability data. Large data-collection systems, in which failure reports might be coded and analysed remotely from the work locations, are usually most at fault in perpetrating the analytical errors described. Such data systems might generate ‘MTBFs’ for systems and for parts by merely counting total reported failures and dividing into total operating time. For example, MTBFs in flying hours are quoted for aircraft electronic equipment, when the equipment only operates for part of the flight, or MTBFs in hours are quoted for valves, ignoring whether they are normally closed, normally open or how often they are opened and closed. These data are often used for reliability predictions for new systems (see Chapter 6), thus adding insult to injury.

A CFR is often a practicable and measurable first-order assumption, particularly when data are not sufficient to allow more detailed analysis.

The effect of successive repairs on the reliability of an ageing system are shown vividly in the next example (from Ascher and Feingold, 1984).

Example 13.5 (Reprinted from Ascher and Feingold (1984) by courtesy of Marcel Dekker, Inc.)

Data on the miles between major failures (interarrival values) of bus engines are shown plotted in Figure 13.6. These show the miles between first, second, . . . , fifth major failure. Note that the interarrival mileages to the first failures (X_i) are nearly normally distributed. Successive interarrival times (second, third, fourth, fifth failures) show a tendency to being exponentially distributed. Nevertheless, the results show clearly that the reliability decreases with successive repairs, since the mean of the interarrival distances is progressively reduced:

Failure No.	$\bar{X}_i$ miles
1	94 000
2	70 000
3	54 000
4	41 000
5	33 000

The importance of this result lies in the evidence that:

1 Repair does not return the engines to an 'as new' condition.
2 Successive X_i s are not IID exponential.
3 The failure rate tends to a constant value only after nearly all engines have been repaired several times. Even after five repairs the steady state has not been reached.
4 Despite the *appearance* of 'exponentiality' after several failures, replacement or more effective overhaul appears to be necessary.

13.9 CUSUM Charts

The 'cumulative sum', or CUSUM, chart is an effective graphical technique for monitoring trends in quality control and reliability. The principle is that, instead of monitoring the measured value of interest (parameter value, success ratio), we plot the divergence, plus or minus, from the target value. The method is the same as the scoring principle in golf, in which the above or below par score replaces the stroke count. The method enables us to report progress simply and in a way that is very easily comprehended.

The CUSUM chart also provides a sensitive indication of trends and changes. Instead of indicating measured values against the sample number, the plot shows the CUSUM, and the slope provides a sensitive indicator of the trend, and of points at which the trend changes.

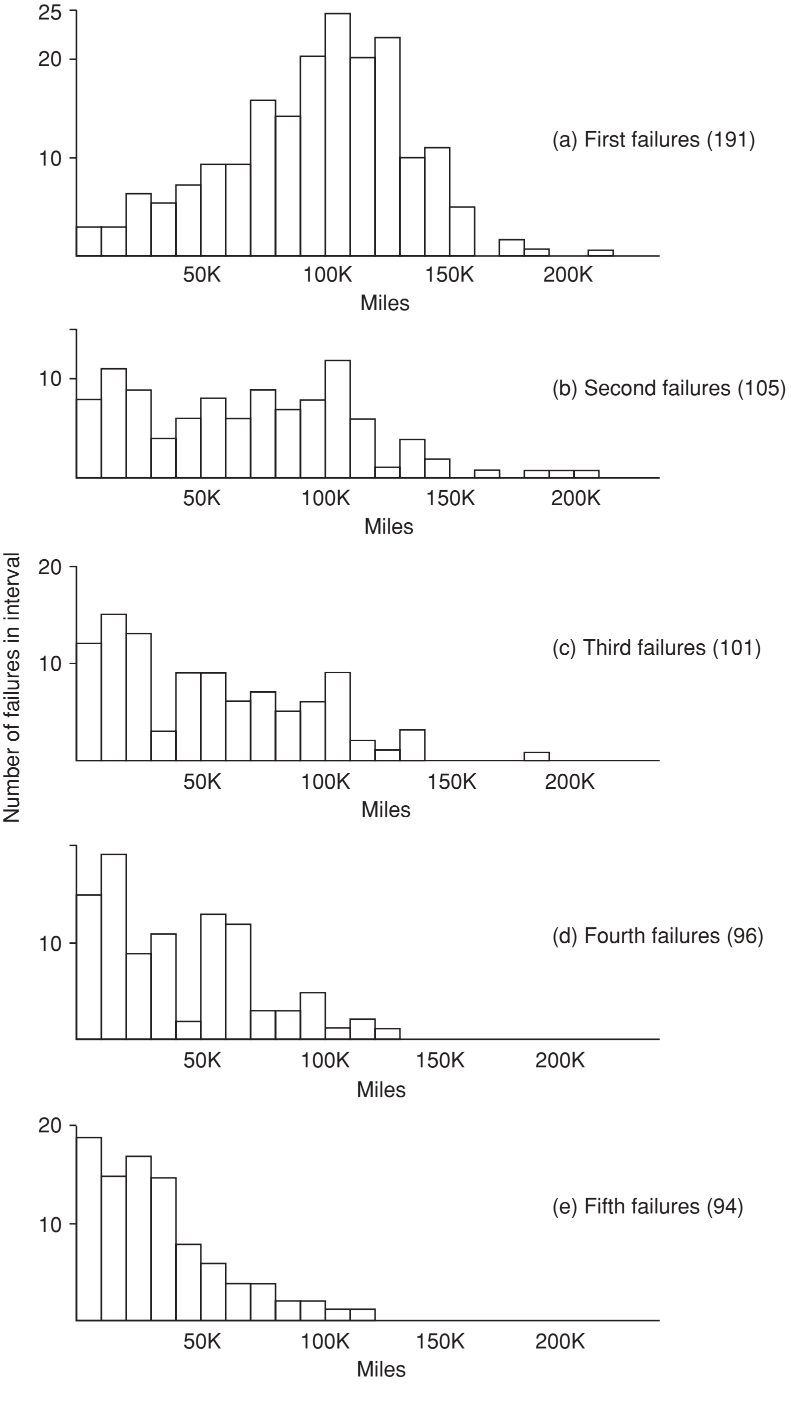


Figure 13.6 *Bus engine failure data (a) First failures (191), (b) Second failures (105), (c) Third failures (101), (d) Fourth failures (96), (e) Fifth failures (94).*

Table 13.4 Reliability test data Target = 95 % (T).

Sample, i	x_i	$x_i - T$	CUSUM $\Sigma(x_i - T)$
1	86	−9	−9
2	88	−7	−16
3	85	−10	−26
4	87	−8	−34
5	88	−7	−41
6	91	−4	−45
7	91	−4	−49
8	93	−2	−51
9	93	−2	−53
10	94	−1	−54
11	92	−3	−57
12	95	0	−57
13	94	−1	−58
14	96	1	−57
15	94	−1	−58
16	93	−2	−60
17	95	0	−60
18	97	2	−58
19	96	1	−57
20	96	1	−56
21	94	−1	−57
22	96	1	−56
23	97	2	−54
24	95	0	−54
25	96	1	−53
26	97	2	−51
27	98	3	−48
28	98	3	−45
29	96	1	−44
30	98	3	−41

Table 13.4 shows data from a reliability test on a one-shot item. Batches of one hundred are tested, and the target success ratio is 0.95.

Figure 13.7 (a) shows the results plotted on a conventional run chart.

Figure 13.7 (b) shows the same data plotted on a CUSUM chart, with the CUSUM values calculated as shown in Table 13.4.

The CUSUM can be restarted with a new target value if a changed, presumably improved, process average is attained. Decisions on when to restart, the sample size to take, and scaling of the axes will depend upon particular circumstances.

Guidance on the use of CUSUM charts is given in British Standard BS 5703 (see Bibliography), and in good books on statistical process control, such as those listed in the Bibliography for Chapter 15.

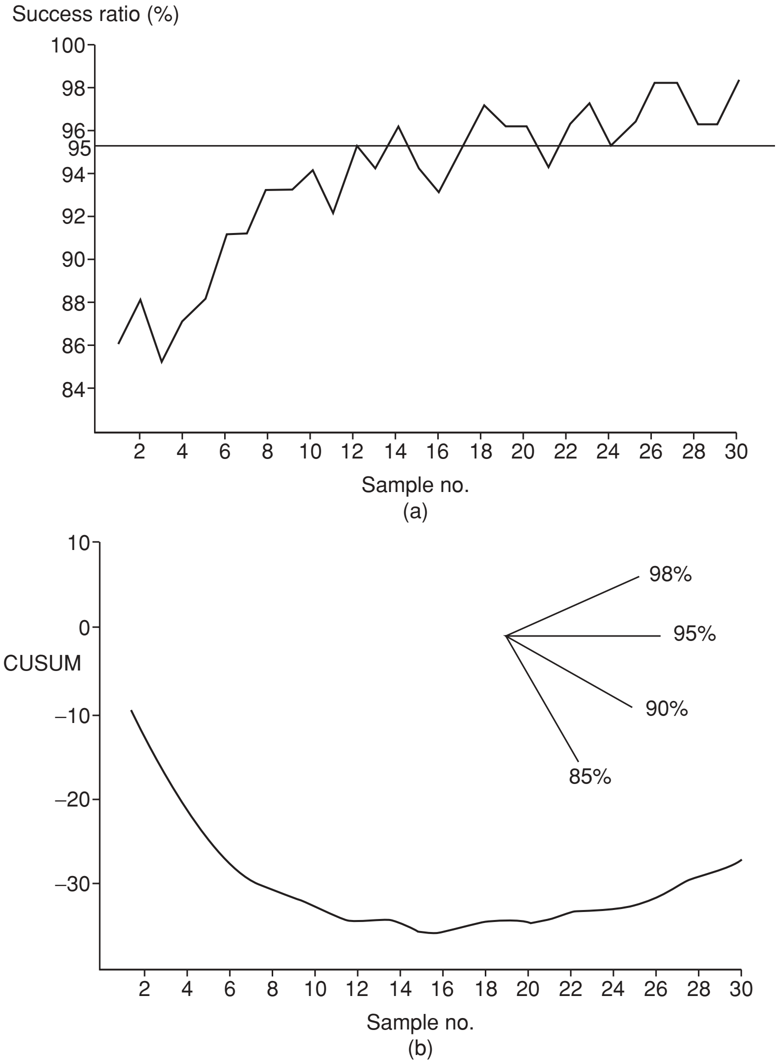


Figure 13.7 *(a) Run chart of data in Table 13.4. (b) CUSUM chart of data of Table 13.4.*

13.10 Exploratory Data Analysis and Proportional Hazards Modelling

Exploratory data analysis is a simple graphical technique for searching for connections between time series data and explanatory factors. It is also used as an approach to analysing data for the purpose of formulating hypotheses worth testing. In the reliability context, the failure data are plotted as a time series chart, along with the other information. For example, overhaul intervals, seasonal changes, or different operating patterns can be shown on the chart. Figure 13.8 shows failure data plotted against time between scheduled overhauls. There is a clear pattern of clustering of failures shortly after each overhaul, indicating that the overhaul is actually adversely affecting reliability. In this case, further investigation would be necessary to determine the reasons for this, for example, the quality of the overhaul work might be inadequate. Another feature that

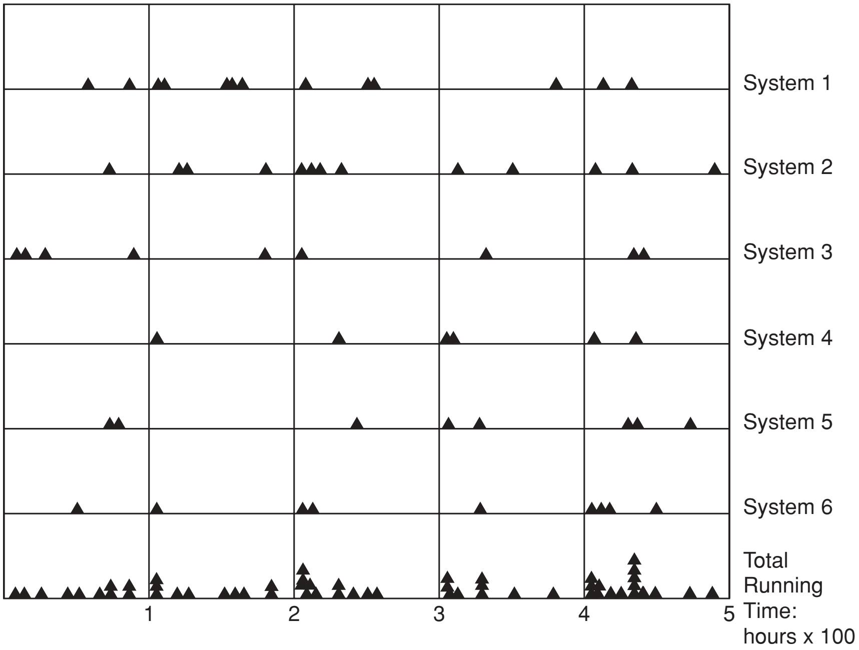


Figure 13.8 *Time series chart: failure* vs *time (overhaul interval 1000 h).*

shows up is a tendency for failures to occur in clusters of two or more. This seems to indicate that failures are often not diagnosed or repaired correctly the first time.

This method of presenting data can be very useful for showing up causes of unreliability in systems such as vehicle fleets, process plant, and so on. The data can be shown separately for each item, or by category of user, and so on, depending on the situation, and analysed for connections or correlations between failures and the explanatory factors.

Proportional hazards modelling (PHM) is a mathematical extension of EDA. It is used to model the effect of secondary variables on product failures. The basic proportional hazards model is of the form

$$\lambda(t;\ Z_1,\ Z_2, \ldots, Z_k) = \lambda_0(t)\ \exp(\beta_1 Z_1 + \beta_2 Z_2 + \cdots \beta_k Z_k) \tag{13.15}$$

where $\lambda(t\ Z_1, Z_2, \ldots, Z_k)$ represents the hazard rate at time t, $\lambda_0(t)$ is the baseline hazard rate function, $Z_1, Z_2, \ldots, Z_k$ are the explanatory factors (or *covariates*), and $\beta_1, \beta_2, \ldots, \beta_k$ are the model parameters.

In the proportional hazards model, the covariates are assumed to have multiplicative effects on the total hazard rate. In standard regression analysis or analysis of variance the effects are assumed to be additive. The multiplicative assumption is realistic, for example, when a system with several failure modes is subject to different stress levels, the stress having similar effects on most of the failure modes. The proportional hazards approach can be applied to failure data from repairable and non-repairable systems.

The theoretical basis of the method is described in Kalbfleisch and Prentice (2002). The derivation of the model parameters requires the use of advanced statistical software, as the analysis is based on iterative methods. This limits application of the technique to teams with specialist knowledge and access to the appropriate software.

13.11 Field and Warranty Data Analysis

Field returns and warranty data can be an excellent source for product reliability analysis and modelling. The field environment is the ultimate test for product performance; therefore the reliability should be evaluated based on field failures whenever possible. Warranty claims database can be a source of engineering analysis of the failure causes as well as for the forecasting of the future claims. However, working with warranty data requires an understanding of its specifics and limitations in order to produce meaningful results.

13.11.1 Field and Warranty Data Considerations

Warranty data comes in various shapes and forms depending on the industry, type of product, individual manufacturer, and many other factors. A typical warranty database contains a large amount of information about each claim, however a reliability practitioner should understand that besides 'true' failures there is a certain amount of 'noise' in this data.

Figure 13.9 shows an example of warranty database content. The 'noise' factors include NFFs (no fault found) common in the electronics industry (Chapter 9), fraudulent claims, inaccurate reporting including missing data and misuse. The amounts of 'noisy' claims would vary significantly based on the industry, type of the product and even individual manufacturer. Therefore it is important to be able to 'clean' the data distinguishing between the 'relevant' and 'irrelevant' claims and also further categorize them by failure modes. However, it is also important to remember that the product user is affected by *all product failures* regardless of their origin, therefore it is often beneficial to process the data 'as is' without removing those considered irrelevant.

Another complicating factor in high volume warranty analysis is that the items are continuously produced and sold; therefore warranty periods for different items begin at different times. Thus it is easier to model warranty in the product age format as opposed to the calendar time format.

It is also important for a reliability professional to remember that warranty periods are usually shorter than the expected life of a product, therefore warranty data does not normally provide enough information to evaluate the reliability at the later phases of product life, where wearout is expected.

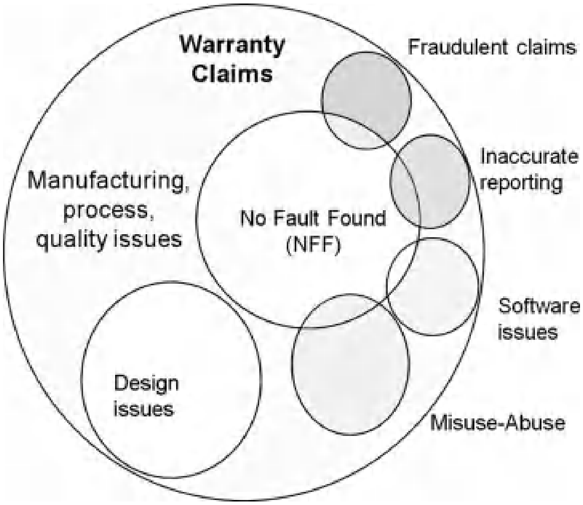


Figure 13.9 *Warranty claims root causes.*

13.11.2 Warranty Data Formats

Warranty databases often contain two types of data. One would have detailed information about each claim, while another type has the warranty statistics, such as sales volumes, number of failures and the relevant times to failure. These formats are covered below.

13.11.2.1 Individual Claims Data Format

Individual claims data contain various types of information about each failure. That usually includes the production date, sales date, failure date, description of the problem, location of the claim (country, state, region, etc.), cost of repair, who performed the repair, and so on. It may also contain the usage data, such as mileage, number of runs, cycles, loadings, and so on. Detailed claims data can be used for the engineering analysis and feedback on the root causes of design and manufacturing problems. Pareto charts Figure 13.1 can also be compiled after determining a root cause of each failure. Based on this information engineers can improve the existing product and also learn lessons to be incorporated in the design for reliability (DfR) process. For more engineering and business applications of warranty data analysis please see Kleyner (2010b). Individual claims data are often based on a sample, rather than on processing all of the returned items, especially if the number of repairs is large.

Individual claim analysis would also allow the analyst to separate the failure modes of interest and to enhance the statistical data analysis (see next section) by accounting for the failure modes of interest. For example if the individual claims analysis shows that 20 % of all claims were caused by misdiagnosis or customer misuse, then we can accordingly adjust the statistics of the ‘true’ claims.

13.11.2.2 Statistical Data Format

Individual claims data is not sufficient for the life data analysis and forecasting. It needs to be combined with the information about the whole population of units in the field in order to be useful for a statistical analysis. Statistical warranty data (sometimes referred to as actuarial) typically contains the sales volumes, number of failed parts and the associated timing information, such as dates of manufacturing, sales, failure or repair.

In statistical data reporting it is often impractical to trace every individual item, especially in high volume production industries; therefore it is customary to group the data on a monthly or other predetermined time interval basis. The exact formats and the amount of information vary from industry to industry and even from company to company. One of the common data reporting formats is called MIS (Month in Service). It is widely used to track automotive warranties, but has been successfully applied in other industries.

Table 13.5 shows the example of warranty data in MIS format. For each month of sale (or month of production if warranty begins with shipping) the number of repaired or returned parts is recorded for each month in service. Other MIS information may include the financials, such as total warranty expenses for that month or the average cost per repair. Each sales month is tracked separately; however the totals can be calculated as weighted averages based on monthly volumes (bottom row of Table 13.5). Obviously the later the product has been manufactured (or sold) the fewer months of warranty claims will be recorded, affecting that month’s contribution to the totals.

Another popular data format used by warranty professionals is called the ‘Nevada’ (the data table resembling the shape of American state of Nevada), which is also sometimes referred as the ‘Layer cake’. Table 13.6 shows the monthly returns from Table 13.5 presented in the Nevada format. The returns are shown in the calendar time format as opposed to the age format associated with MIS.

The Nevada format allows the user to convert shipping and warranty return data into the standard reliability data form of failures and suspensions so that it can easily be analysed with traditional life data analysis

Table 13.5 *Example of MIS (Month in Service) data (January – July 2011).*

		MIS (Months in Service)											
Sales		1		2		3		4		5		6	
Month	Volume	Repairs	%	Repairs	%	Repairs	%	Repairs	%	Repairs	%	Repairs	%
Jan, 2011	5,000	15	0.30	12	0.24	19	0.38	12	0.24	16	0.32	17	0.34
Feb, 2011	7,000	11	0.16	16	0.23	11	0.16	21	0.30	10	0.14		
Mar, 2011	8,000	9	0.11	17	0.21	9	0.11	12	0.15				
Apr, 2011	6,000	9	0.15	12	0.20	9	0.15						
May, 2011	8,000	17	0.21	21	0.26								
Totals	**34,000**	**61**	**0.18**	**78**	**0.23**	**48**	**0.18**	**45**	**0.23**	**26**	**0.22**	**17**	**0.34**

methods. At the end of the analysis period, all of the units that were shipped and have not failed in the time since shipment are considered to be suspensions. Commercially available warranty analysis software packages can handle various data entry formats. For example ReliaSoft Weibull++® has four different warranty data entry formats including the Nevada.

13.11.3 Warranty Data Processing

Most warranty claims result in repair or replacement of the failed part, therefore warranty data should be analysed using statistical analysis appropriate for the repairable systems, such as renewal process, non-homogeneous Poisson process (NHPP), and so on. Those are covered in Section 2.15, Section 6.7 and Section 13.8. This is especially true in the cases where devices are expected to experience repeat failures or undergo maintenance, such as operating machinery, plant equipment, airplanes, and so on. However in a large volume production of relatively simple parts the number of secondary failures is expected to be small. For example, the number of automotive electronics modules, which experienced secondary failures during warranty is typically below 5 % or even 1 % (see Kleyner and Sandborn, 2008) therefore applying non-repairable data analysis techniques would simplify the analysis and would not result in large calculation errors.

The cumulative failure function $F(t)$ is easier to model than the renewal function or NHPP, therefore reliability $R(t)$ can be obtained from life data analysis of warranty data. However since most units are expected to operate without failure through the warranty period, this data will be heavily censored with a large number of suspensions.

As shown in Figure 13.9, warranty data can be messy and often needs to be 'cleaned' before life data analysis can be performed. Another factor often complicating warranty data analysis is so-called 'data

Table 13.6 *Example of warranty data presented in the 'Nevada' (or the 'Layer cake') format.*

		Number of Failures by Month					
Production Month	Production or sales volume	Feb, 2011	Mar, 2011	Apr, 2011	May, 2011	Jun, 2011	Jul, 2011
Jan, 2011	5,000	15	12	19	12	16	17
Feb, 2011	7,000		11	16	11	21	10
Mar, 2011	8,000			9	17	9	12
Apr, 2011	6,000				9	12	9
May, 2011	8,000					17	21

Table 13.7 Cumulative percent failures for 6 months. Based on the data in Table 13.5.

Time, months	1	2	3	4	5	6
Cumulative percent failed	0.18 %	0.41 %	0.59 %	0.82 %	1.04 %	1.38 %

maturation'. Warranty data is often limited to several months of observation, where the failure trend may not be established yet. Also there is often a lag between repair date and warranty system entry date, thus resulting in underreporting of the latest claims. All that may bias *F(t)* and the overall reliability data analysis. For more on warranty data collection and analysis see Blischke *et al*. (2011).

Example 13.6

Estimate the product reliability at the 36-months warranty period based on the data presented in Table 13.5. Data in the bottom row of Table 13.5 can be used to calculate the cumulative failures (cdf) Table 13.7.

Two parameter Weibull distribution can be fit to the dataset in Table 13.7 using Weibull++®, 'Distribution Fit' option in @Risk® or other distribution fitting software producing $\beta = 1.11$ and $\eta = 297.9$ months.

$$\text{Therefore } R(36 \text{ months}) = \exp(-(36/297.9)^{1.11}) = 90.8\,\%$$

This forecast can only be considered tentative, since this data is based on just six months of observations. To make the matter worse, the sampled population was progressively decreasing with the number of months in service, which further decreases the accuracy of this warranty forecast. For example, Table 13.5 shows that the six-months' warranty is only available for the January 2011 parts, while one month warranty data exists for all parts from January to May 2011. Furthermore, this data has not been filtered for the 'irrelevant' failures; therefore the design related reliability is expected to be somewhat higher than predicted.

More information on warranty data can be found in Blischke and Murthy (1996).

Questions

You can use life data analysis software to solve some of the problems below. If software is not available trial versions of Weibull++® and ALTA® can be downloaded from www.reliasoft.com.

1. Based on the existing data a reliability engineer has developed and implemented an accelerated test plan. Under the assumption of the exponential distribution a test provides MTTF = 300 hours. Given that the acceleration factor is 5.6, determine the reliability under normal conditions for a time equal to 200 hours.
2. Identify at least three models of stress-life relationships and provide examples when these might occur.
3. Consider the following data.
   a Does the stress life relationship follow an expected log log relationship?

Stress (V)	Time to fail (h)
25	5.6
50	10
70	14.3
90	27.8

b Give reasons why a log stress vs. log life relationship might not always be linear at high stress and at low stress.

4. Explain why a simple power law of life vs. stress is unlikely to apply to an assembly that consists of several different stressed components.
5. You desire to estimate the accelerating influences of a corrosive solution through an accelerated life test. The concentration of the solution will be the accelerating factor and four levels will be employed in the test. Metallic samples will be soaked in the various concentrations for 10 % of the day and then allowed to air dry at room temperature. Describe how to approach this test and use the results in the future.
6. Let S measure the time-dependent strength of a material under test and S_0 is the initial strength as measured at the beginning of test. Prior tests and evaluations have suggested that the following stress time relationship appears to work.

$$S^2 = (S_0)^2 - 2Ct(t = \text{time})$$

If the strength declines 20 % after exactly three weeks of accelerated test, what is the expected time to failure at this condition, when the failure is defined as a decline of strength of 50 %?

7. Let the functional relationship of question 6 become $S^4 = (S_0)^4 - 4Ct$ instead (this is a typical degradation curve). Redo the analysis of question 6.
8. The following table of data was generated from a valid accelerated life test. The failure definition employed was the time to 10 % failures of a sample operated at each condition. If the basic model below is correct, what are the values of the three unknowns E_o, N and B for the typical formula shown below?

$$\text{Life} = B(V)^{-N}\text{eE}_\text{a}/\text{KT}$$

Operating temp (°C)	Operating volts	Life (h)
150	70	10
150	50	14.3
150	25	44.7
125	70	27.8
125	50	46.4
125	25	199.5
100	70	117.1
100	50	188.9

9. Explain why the times between successive failures of a repairable system might not be independently and identically distributed (IID).
10. Question 3 in Chapter 3 describes the behaviour of a component in a ‘socket’ of a repairable system. Referring again to that question, suppose you have been given the additional information that machine B was put into service when machine A had accumulated 500 h, machine C when machine A had accumulated 1000 h, machine D when machine A had accumulated 1500 h, and machine E when machine A had accumulated 2000 h.

a Use this additional information about the *sequencing* of failures to calculate the trend statistic (Eq. 2.46), and hence judge whether, as far as this socket is concerned, the system is ‘happy’, ‘sad’ or indeterminate (IID) in terms of Figure 13.5.

b Repeat the exercise, but splitting the data to deal separately with (i) the first eight *sequenced* failures. and (ii) the second eight. What do these results tell you about the dangers of assuming IID failures when the assumption may not be valid?

11. Accelerated vibration test to failure has been performed on a sample of automotive radios. There were three G-levels of vibration referred here as 1-low, 2-medium, 3-high. What can be concluded from the Weibull plots for each stress level shown in Figure 13.10?

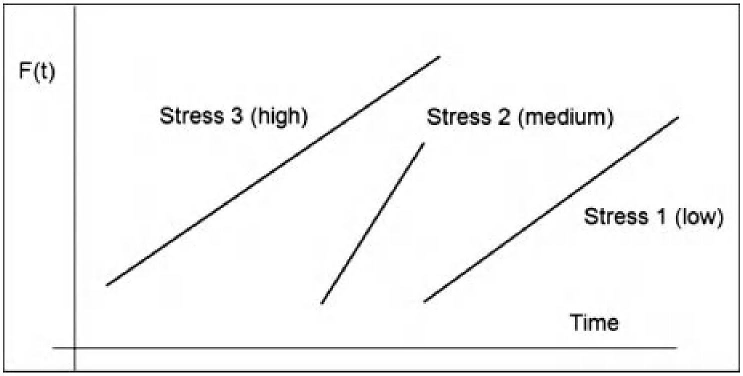


Figure 13.10 *Weibull charts of the test results.*

12. Develop a test for a wind power generator to address corrosion caused by the combination of temperature and humidity. The generator will be installed at the climate conditions with humidity levels up to 80 % RH (relative humidity). Generator is designed to operate 24 hours per day for 10 years. The field usage temperatures are distributed as follows: 20 % of the time at 30 °C, 30 % of the time at 20 °C and 10 % of the time at 0 °C. Calculate the duration of the test at 85 °C-85 % RH using Peck's model assuming $E_A = 0.7\text{eV}$ and humidity power constant $m = 3.0$.

13. In an investigation into cracking of brake discs on a locomotive, a proportional hazards analysis was undertaken on a sample containing 205 failures and 905 censorings. (A failure was removal of an axle because a crack had propagated to such an extent that replacement was needed to avoid any possibility of fracture, a censoring was removal of an axle for any other reason.) Referring to Eqn (13.9), the covariates were:

Z_1 = region of operation (0 for Eastern region, 1 for Western region).
Z_2 = braking system (0 for type A, 1 for type B).
Z_3 = disc material (0 for material X, 1 for material Y).

The data were analysed using computer methods (the only practicable way) to give the following coefficients: $\beta_1 = 0.39$, $\beta_2 = 0.72$, $\beta_3 = 0.95$.

a At any given age, what is the ratio between the hazard functions of axles running on the two regions?
b What is the ratio for the two braking systems?
c What is the ratio for the two disc materials?

[This example is based on real data first reported by Newton and Walley in 1983. The analysis is further developed in Bendell, Walley, Wightman and Wood (1986), 'Proportional hazards modelling in reliability analysis – an application to brake disks on high-speed trains', *Quality and Reliability International*, **2**, 42–52.]

14. The following data give the times between successive failures of an aircraft air-conditioning unit: 48, 29, 502, 12, 70, 21, 29, 386, 59, 27, 153, 26 and 326 hours. When a unit fails it is repaired *in situ*. Repairs can be assumed to be instantaneous.
    a Examine the distribution of failure times, treating the unit as a component on the aircraft.
    b Examine the trend of failures, viewing the air- conditioning system itself as a repairable system.
    c Describe clearly the conclusions to be drawn from both these analyses.
15. Two prototypes of a newly designed VHF communications set have been manufactured. Each has been subjected to a life-test as follows:

    Prototype A – had failures at 37, 53, 102, 230 and 480 h running; withdrawn from test at 600 h.

    Prototype B – started test when A was withdrawn (only one test rig was available) and had failures after 55, 290, 310, 780 and 1220 h; is still running, having accumulated 1700 h.

    a On the assumption of random failures, estimate the failure rate and the probability of surviving a 100 h mission without failure.
    b Revise your answers to (a) if you apply a reliability growth model to the data, assuming that all modifications found necessary during the test on prototype A were applied to prototype B.
16. Calculate the overall mean time between failures for the data in Question 10. Produce a CUSUM chart of the data, plotting $\Sigma(t_i - T)$ against i where t_i is the elapsed time between the ith and the $(i - 1)$th sequenced failures and T is the expected time since the previous failure based on the overall MTBF.

    From this plot, identifying when any change to the MTBF might have occurred, and estimate the MTBF before and after the change.

    Consider whether this plot shows the situation more clearly than a straightforward trend plot of cumulative failures against cumulative time.
17. Run the paperclip example as described at http://www.weibull.com/AccelTestWeb/paper_clip_example.htm Use the angle of bend as a stress variable and measure the fatigue life as a number of cycles till the clip's inner loop breaks. Run your own calculations and predict the fatigue life at 45 degree bend and check your results by actually running the experiment at 45 degree bend.
18. Vibration test has been conducted at two different stress levels: 4.0 and 6.0 G peak to peak acceleration. The results in cycles to failure are presented in the table below

4.0 G Vibration	6.0 G Vibration
9.60×10^5 cycles	4.92×10^4 cycles
1.52×10^6 cycles	5.60×10^4 cycles
8.35×10^5 cycles	5.32×10^4 cycles

    a Determine the fatigue exponent *b*.
    b Calculate the expected number of cycles to failure at the vibration level of 2.0G.
19. What are the purpose and the benefits of Pareto Analysis? Make the examples of the situations where using the Pareto Analysis would be beneficial.
20. A Pareto analysis of field return warranty data shows several categories which are approximately equal. In developing the corrective actions to reduce the number those problems how would you choose which of those categories to address first? Consider the criteria such as cost of fixing the problem, ease of fixing the problem, overall cost saving, cost-benefit analysis, how quick you will see the results and others.

21. Discuss the warranty claims root causes presented in Figure 13.9.
    a Which of those categories are easiest to detect in a warranty database and which ones would be most difficult?
    b How would be your course of actions in taking corrective actions for each category? Which ones would be easiest to take corrective action about and which would be most difficult?
    c In your opinion, which categories would offer biggest cost savings?
22. What are the limitations of using acceleration models in developing accelerated test plans? Which limitations would be eliminated and which would remain if instead of using generic models we apply the models based on the actual accelerated test data?

Bibliography

Ascher, H. and Feingold, H. (1984) *Repairable Systems Reliability*, Dekker.

Blischke, W. and Murthy, D. (1996) *Product Warranty Handbook*, Marcel Dekker.

Blischke, W., Karim, R. and Murthy, D. (2011) *Warranty Data Collection and Analysis*, Springer.

British Standard BS 5703. *Guide to Data Analysis and Quality Control using Cusum Techniques*. British Standards Institute, London.

British Standard, BS 5760. *Reliability of Systems, Equipments and Components*, Part 2. British Standards Institution, London.

Clech, J-P., Henshall, G. and Miremadi, J. (2009) Closed-Form, Strain-Energy Based Acceleration Factors for Thermal Cycling of Lead-Free Assemblies. Proceedings of SMTA International Conference (SMTAI 2009), Oct. 4-8, 2009, San Diego, CA.

ISO IEC 60605. *Equipment Reliability Testing*. International Standards Organisation, Geneva.

JEDEC (2009) JEP122F *Failure Mechanisms and Models for Semiconductor Devices*. Published by JEDEC Association. Available at http://www.jedec.org/Catalog/catalog.cfm.

Kalbfleisch, J., Lawless, J. and Robinson, J. (1991), Methods for the Analysis and Prediction of Warranty Claims. *Technometrics*, **33**(3) August.

Kalbfleisch, J. and Prentice, R. (2002) *The Statistical Analysis of Failure Time Data*, 2nd edn, Wiley.

Kleyner, A. (2010a) *In the Twilight Zone of Humidity Testing*. TEST Engineering & Management, August/September issue.

Kleyner A. (2010b) Discussion Warranted. *Quality Progress, International Monthly Journal of American Society for Quality (ASQ)*. May 2010 issue, pp. 22–27.

Kleyner, A. and Sandborn, P. (2008) Minimizing life cycle cost by managing product reliability via validation plan and warranty return cost. *International Journal of Production Economics*, **112**, 796–807.

Lawson, R. (1984) A review of the status of plastic encapsulated semiconductor component reliability. *British Telecommunication Technology Journal*, **2**(2), 95–111.

Moltoft, J. (1994) Reliability Engineering Based on Field Information: The Way Ahead. *Quality and Reliability Engineering International*, **10**, 399–409.

Nelson, W. (1989) *Accelerated Testing: Statistical Models, Test Plans and Data Analysis*, Wiley.

NIST (2006) *Eyring*. National Institute of Standards and Technology, online Handbook section 8.1.5.2 Available at: http://www.itl.nist.gov/div898/handbook/apr/section1/apr152.htm.

Norris, K. and Landzberg, A. (1969) Reliability of Controlled Collapse Interconnections. *IBM Journal of Research and Development*, **13**(3), 266–271.

Ohring, M. (1998) *Reliability and Failure of Electronic Materials and Devices*, Academic Press.

Pan, N., Henshall, G.A., Billaut, F. *et al.* (2005) *An Acceleration Model for Sn-Ag-Cu Solder Joint Reliability under Various Thermal Cycle Conditions*, SMTA International, pp. 876–883.

Peck, D. (1986) Comprehensive Model for Humidity Testing Correlation. *IEEE IRPS Proceedings*, 44–50.

ReliaSoft (2010) *Accelerated Life Testing Reference*, ReliaSoft Publishing.

ReliaSoft (2010) *ALTA-7 User's Guide*, ReliaSoft Publishing.

ReliaSoft (2011) *Accelerated Life Testing Reference*. Online Reliability Engineering Resources. Available at http://www.weibull.com/acceltestwebcontents.htm.

Salmela, O. (2007) Acceleration Factors for Lead-Free Solder Materials. *IEEE Transactions on Components and Packaging Technologies*, **30**(4), December 2007.

SEMATECH (2000) *Semiconductor Device Reliability Failure Models*. Technology Transfer # 00053955A-XFR. International SEMATECH Report. Available at: http://www.sematech.org/docubase/document/3955axfr.pdf.

Steinberg, D. (2000), *Vibration Analysis for Electronic Equipment*, 3rd edn, Wiley.

Steven E., Rigdon, S. and Basu, A. (2000) *Statistical Methods for the Reliability of Repairable Systems*, Wiley.

Trindade, D. and Nathan, S. (2005) Simple Plots for Monitoring the Field Reliability of Repairable Systems. Proceedings of the Annual Reliability and Maintainability Symposium (RAMS).

US MIL-HDBK-781. *Reliability Testing for Equipment Development, Qualification and Production*. Available from the National Technical Information Service, Springfield, Virginia.

14

Reliability Demonstration and Growth

14.1 Introduction

As described in Chapter 12, reliability testing is the cornerstone of a reliability engineering programme. Chapter 12 emphasized the importance of reliability testing being planned primarily to improve reliability by showing up potential weaknesses. However, a properly designed series of tests can also generate data that would be useful in determining if the product meets specified requirements to operate without failure during its mission life, or to achieve a specified level of reliability. Many product development programmes require a series of environmental tests to be completed to demonstrate that the reliability requirements are met by the manufacturer and then demonstrated to the customer.

Reliability demonstration testing is usually performed at the stage where hardware (and software when applicable) is available for test and is either fully functional or can perform all or most of the intended product functions. While it is desirable to be able to test a large population of units to failure in order to obtain information on a product's or design's reliability, time and resource constraints sometimes make this impossible. In cases such as these, a test can be run on a specified number of units, or for a specified amount of time, that will demonstrate that the product has met or exceeded a given reliability at a given confidence level. In the final analysis, the actual reliability of the units will of course remain unknown, but the reliability engineer will be able to state that certain specifications have been met. This chapter will discuss those requirements and different ways of achieving them in the industrial setting.

14.2 Reliability Metrics

Most product design specifications come with some form of reliability requirements, which are expressed as reliability metrics. If the system is complex those requirements may come from reliability apportionment activities (Chapter 6) or other requirements generated at the system, subsystem or component level.

Probably the most common reliability metric is the simple reliability function R(t). For example a product specification may state that the expected reliability over a 5-yr life should be no less than 98.0 % meaning $R(5 \text{ yrs}) = 0.98$. A reliability requirement often comes with a specified confidence level based on a test sample size. This will be covered in the next section.

Another popular metric is mean time between failures (MTBF) (Section 2.6.3). MTBF is a characteristic of a repairable system with constant failure rate (see the exponential distribution Eq. (2.27)). MTBF is

Practical Reliability Engineering, Fifth Edition. Patrick D. T. O'Connor and Andre Kleyner.

often misinterpreted by non-reliability engineers as an average time between consecutive failures in all of the product population even though those could be failures of different units. Therefore it is recommended to use other reliability demonstration metrics whenever possible. For non-repairable systems with constant failure rate mean time to failure (MTTF) is used in place of MTBF (see Chapters 1 and 2). Consequently, the failure rate λ can also be used as a reliability metric for both repairable and non repairable systems.

B_X-life (Section 3.4.5) is another common reliability metric, which is often used as B_{10} specification, the product life at which 10 % of the population is expected to fail (i.e. 90 % reliability).

PPM (parts per million) can also be used as a reliability metric, although it is more often used to measure manufacturing quality in production. In the case of reliability, PPM would be a function of time, which would represent the number of parts per million failed during the time interval [0; t], therefore:

$$R(t) = 1 - \frac{PPM(t)}{10^6} \tag{14.1}$$

One form of reliability metric can often be converted to the other and used interchangeably.

Example 14.1

The initial product requirement defines B_{10} life of 5 years. Convert this requirement to other reliability metrics. Under the assumption of exponentially distributed failures:

$$R(5yrs) = 0.90 = \exp\left(-\frac{5yrs}{MTTF}\right) \tag{14.2}$$

Solving (14.2) we obtain MTTF = 47.5 years. Failure rate λ = 1/MTTF = 0.021 failures per year or assuming 24 hours per day operation $\lambda = 0.0210/(365 \times 24) = 2.4 \times 10^{-6}$ failures per hour. Also, based on (14.1) R(5yrs) = 0.9 would mean 100 000 PPM at 5 yrs.

14.3 Test to Success (Success Run Method)

Test to success, where failures are undesirable has been covered in Chapter 12. Industry dependent, it is also referred as *success run testing, attribute test*, *zero failure substantiation test* or *mission life test*. Under those conditions a product is subject to a test, often accelerated representing an equivalent to one field life (*test to a bogey*), which is expected to be completed without failure by all the units in the test sample. Methods of estimating the test equivalent of one field or mission life were discussed in Section 13.6.

14.3.1 Binomial Distribution Approach

Success run test statistics are most often based on the binomial distribution presented in Section 2.10.1. The binomial pdf is described by Chapter 2 Eq. (2.37) which can be applied to the test situations with only two possible outcomes: *pass* or *fail*. Therefore, assuming that reliability $R = p$ in (2.37) the probability of a product to survive (based on the binomial cdf) can be presented in the form of:

$$C = 1 - \sum_{i=0}^{k} \frac{N!}{i!(N-i)!} R^{N-i}(1-R)^i \tag{14.3}$$

where: R = unknown reliability.
C = confidence level.
N = total number of test samples.
k = number of failed items.

Table 14.1 *Required test sample sizes for reliability demonstration at 50 and 90 % confidence.*

Reliability, R	Sample size N at $C = 50\,\%$	Sample size N at $C = 90\,\%$
90 %	7	22
95 %	14	45
97 %	23	76
99 %	69	230
99.9 %	693	2,301
99.99	6,932	23,025

If $k = 0$ (no failures) (14.3) turns into a simple equation for success run testing:

$$C = 1 - R^N \tag{14.4}$$

(14.4) can be solved for the test sample size N as:

$$N = \frac{\ln(1 - C)}{\ln R} \tag{14.5}$$

When demonstrated reliability R approaches 1.0 the required sample size N based on Eq. (14.5) approaches infinity. Table 14.1 shows the required test sample sizes for statistical confidence of 50 and 90 % based on (14.5).

Product design and validation requirements often explicitly specify reliability and confidence level for their reliability demonstration programmes. For example, a common requirement in the automotive industry is to demonstrate reliability of 97.0 % with 50 % confidence. According to Table 14.1 this would require 23 units to be tested without failure to an equivalent of one field life.

14.3.2 Success Run Test with Undesirable Failures

When undesirable failures occur during test to success, reliability can still be estimated using Eq. (14.3). However (14.3) can be difficult to solve for R, especially when $k > 2$, therefore an approximation chi-square formula can be used instead:

$$R = \exp\left(-\frac{\chi^2_{(1-C,2k+2)}}{2N}\right) \tag{14.6}$$

The derivation of (14.6) is based on the estimates for the MTBF confidence intervals covered later in Section 14.6. χ^2 values can be found in Appendix 2 or calculated using the Excel function CHIINV(*1-C*, *2k + 2*).

14.4 Test to Failure Method

Testing to demonstrate reliability when failures occur during test can be analysed using the life data analysis methods described in detail in Chapter 3. Based on results of life data analysis we can model the reliability function $R(t)$ based on the chosen distribution and generate confidence bounds (Section 3.6) corresponding to the confidence level required for the demonstrated reliability. The two or three-parameter Weibull distribution (Section 3.4) is probably the most common choice for reliability practitioners to model the $R(t)$ function. The value of the Weibull slope β also gives insight into the product's bathtub curve (infant mortality, useful life, or wearout mode).

The downside of the test to failure method is longer test times compared to equivalent success run tests. Whereas success run testing requires test durations equivalent to one product field life (or its accelerated equivalent), a test to failure requires at least twice that in order to allow time to generate enough failures for the life data analysis. Additionally, test to failure requires some form of monitoring equipment to record failure times. Therefore, due to ever-increasing pressure to reduce development cycles, project managers often opt for success run testing instead of testing to failure.

14.5 Extended Life Test

Cost consideration is always an important part of a test planning process. Test sample size carries the cost of producing each test sample (which can be quite high in some industries), equipping each sample with monitoring equipment and providing an adequate test capacity to accommodate all the required samples. A large sample may also require additional floor space for additional test equipment, such as temperature/humidity chambers or vibration shakers, which can be quite costly.

Test duration can be used as a factor to reduce the cost of reliability demonstration. In the cases where test samples are expensive it may be advantageous to test fewer units, though for longer periods of time.

14.5.1 Parametric Binomial Method

Combining the success run formulae (14.4) with 2-parameter Weibull reliability, Eq. (2.31) Lipson and Sheth (1973) developed the relationship between two test sets with (N_1, t_1) and (N_2, t_2) characteristics needed to demonstrate the same reliability and confidence level.

$$\frac{N_2}{N_1} = \left(\frac{t_1}{t_2}\right)^{\beta} \tag{14.7}$$

where: $\beta =$ Weibull slope for primary failure mode (known or assumed).
$N_1, N_2 =$ test sample sizes.
$t_1, t_2 =$ test durations.

Therefore it is possible to extend test duration in order to reduce test sample size. Therefore if t_1 is equivalent to one mission life $t_2 = Lt_{1,}$ where L is the life test ratio. Thus combining this with (14.7).

$$N_1 = L^{\beta} N_2 \tag{14.8}$$

With the use of (14.8) the success run formula (14.4) will transform into:

$$C = 1 - R^{NL^{\beta}} or\, R = (1 - C)^{\frac{1}{NL^{\beta}}} \tag{14.9}$$

Relationship (14.9) is often referred as the *parametric binomial model*. Test sample size reduction at the expense of longer testing can also be beneficial in the cases with limited test capacities. For example, if an environmental chamber can accommodate only 18 test samples, while 23 are required for the reliability demonstration, test time can be extended to demonstrate the same reliability with only 18 units.

The required number of test samples can be reduced L^β times in the cases of extended life testing ($L > 1$). Therefore this approach allows additional flexibility to minimize the cost of testing by adjusting test sample size up or down to match the equipment capacity.

Example 14.2

The design specification requirement is 97.0 % reliability with 50 % confidence for the 1000 hours temperature test corresponding to one product field life. A temperature chamber equipped with a test monitoring rack has a capacity of only 15 units. Calculate the test time required to meet the above reliability requirement, assuming the Weibull slope $\beta = 2.0$.

According to (14.5) and Table 14.1 the sample size of 23 would be required to demonstrate 97 % reliability with 50 % confidence. Thus solving (14.9) for L and substituting $R = 97.0\,\%$, C $= 50\,\%$, $N = 15$, and $\beta = 2.0$ produces:

$$L = \left(\frac{\ln(1 - C)}{N \ln R}\right)^{\frac{1}{\beta}} = \left(\frac{\ln(1 - 0.5)}{15 \ln 0.97}\right)^{\frac{1}{2.0}} = 1.232 \qquad (14.10)$$

Therefore the original test duration of 1000 hours should be extended to $Lt = 1.232 \times 1000 = 1232$ hours without failure in order to demonstrate the required reliability with 15 samples instead of 23 (see also Figure 14.2).

14.5.2 Limitations of the Parametric Binomial Model

This method has been successfully applied to both extended and reduced life testing. However it is not recommended to change the test to a bogey time by more than ±50 %, because it may violate the assumptions of the parametric binomial model. Since (14.9) uses Weibull slope β it assumes a particular trend in the hazard rate. For example if $\beta = 3.0$ the model reflects the wear-out rate corresponding to the Weibull slope of 3.0. Significantly extending test time may accelerate the wear out process where $\beta > 3.0$, thus increasing the probability of failure beyond the model's assumptions. The reverse is also true. Shortening test time may shift the failure pattern from the wear-out phase on the bathtub curve to the useful life. That would effectively reduce the β -value to 1.0, again violating the assumptions of the parametric binomial model.

14.6 Continuous Testing

During continuous testing, the mean time between failures can be calculated as the total test time T amongst all tested units divided by the number of failures k ($MTBF = T/k$). When a product's failure rate is considered constant, the χ^2 distribution may be used to calculate confidence intervals around MTBF, and therefore demonstrated reliability.

The lower and upper confidence limits for data which are generated by a homogeneous Poisson process are given in Table 14.2. It shows the one-sided and two-sided limits for the conditions where the test is stopped at the kth failure, that is a *failure truncated test*, and for a *time truncated test*, (i.e. test is stopped after a predetermined time). Values of χ^2 for different risk factors α and degrees of freedom are given in Appendix 2 or can be calculated using Excel statistical function CHIINV.

In the case of non-repairable systems MTBF becomes MTTF.

Table 14.2 *MTBF confidence limits.*

	Time truncated test	Failure truncated test
One sided confidence interval	$MTBF \leq \frac{2T}{\chi^2_{(\alpha,2k+2)}}$	$MTBF \leq \frac{2T}{\chi^2_{(\alpha,2k)}}$
Two sided confidence interval	$\frac{2T}{\chi^2_{(\frac{\alpha}{2},2k+2)}} \leq MTBF \leq \frac{2T}{\chi^2_{(1-\frac{\alpha}{2},2k+2)}}$	$\frac{2T}{\chi^2_{(\frac{\alpha}{2},2k)}} \leq MTBF \leq \frac{2T}{\chi^2_{(1-\frac{\alpha}{2},2k)}}$

Where: T = total test time,
α = the acceptable risk of error (1- C),
k = the number of failures.

Example 14.3

Ten units were tested for a total of 2000 h and 3 failures occurred. The test was time-truncated. Assuming a CFR, what is the demonstrated 90 % lower confidence limit on reliability at 100 operating hours?

Using the appropriate equation from Table 14.2 and the Excel function CHIINV (or Appendix 2):

$$MTBF \leq \frac{2T}{\chi^2_{(\alpha,2k+2)}} = \frac{2 \times 2000}{CHIINV(0.1, 2 \times 3 + 2)} = \frac{4000}{13.362} = 299.4\,h$$

Therefore, based on the exponential distribution:

$$R(100\,h) = \exp\left(-\frac{100h}{299.36\,h}\right) = 0.716\ (71.6\%)$$

14.7 Degradation Analysis

Product testing takes time and it may take too long for the product to fail operating under normal or even accelerated stress conditions. Test to success may also take a prohibitively long time when simulating long field lives (e.g. 15–30 years). Degradation analysis introduced in Chapter 7 is one way of demonstrating reliability within a relatively short period of time. Many failure mechanisms can be directly linked to the degradation of part of the product, and degradation analysis allows the user to extrapolate to a failure time based on the measurements of degradation or change in performance over time. Examples of product degradation include the wear of brake pads, crack propagation due to material fatigue, decrease in conductivity, loss of product performance such as generated power, and so on.

For effective degradation analysis it is necessary to be able to define a level of degradation or performance which constitutes a failure. Once this *failure threshold* is established it is a relatively simple matter to use basic mathematical models to extrapolate the performance measurements over time to the point where the failure is expected to occur. Once the extrapolated failure times have been determined, it is a matter of conducting life data analysis to model and analyse the demonstrated reliability of the tested items.

Figure 14.1 demonstrates the concept of degradation analysis, where the measurements were taken at 0, 250 and 500 hours. This data was then extrapolated to estimate the point in time where the degradation parameter crosses the 'failure threshold' and becomes a failure. Most commonly used extrapolation models include linear ($y = bx+c$), exponential ($y = be^{axe}$) and power ($y = bx^a$). Other models are also available in commercial data analysis software packages (see ReliaSoft, 2006).

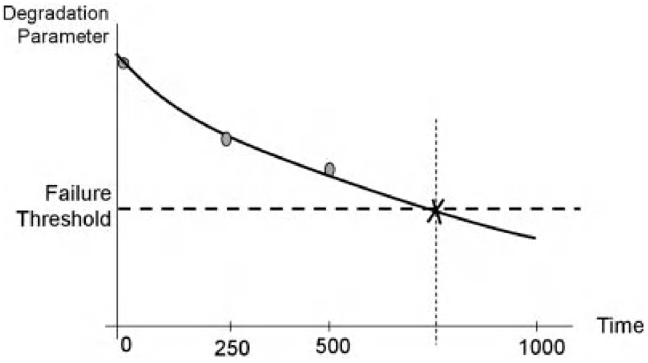


Figure 14.1 Degradation analysis diagram.

Before conducting reliability demonstration based on degradation analysis the following should be considered:

- The chosen degradation parameter has to be critical to the product failure and represent the dominant failure mechanism.
- Verify that the chosen parameter is affected by the test. This parameter may in reality degrade due to other causes, such as poor quality or be unrelated to the test altogether.
- The chosen parameter should exhibit a clear degradation trend. It would invalidate the test if the product exhibits fluctuations in the parameter value over time.
- As with any sort of extrapolation, one must be careful not to extrapolate too far beyond the actual range of data in order to avoid large inaccuracies.

Simple degradation analysis can be done with an Excel spreadsheet by extrapolating data points and estimating the time when the degradation value reaches the threshold. However it is always more efficient to use software specifically designed for that purpose.

14.8 Combining Results Using Bayesian Statistics

It can be argued that the result of a reliability demonstration test is not the only information available on a product, but that information is available prior to the start of the test, from component and subassembly tests, previous tests on the product and even intuition based upon experience. Why should this information not be used to supplement the formal test result? Bayes theorem (Chapter 2) states (Eq. 2.9):

$$P(\mathrm{B} \mid \mathrm{A}) = \frac{P(\mathrm{A} \mid \mathrm{B})P(\mathrm{B})}{P(\mathrm{A})}$$

enabling us to combine such probabilities. Eq. (2.9) can be extended to cover probability distributions:

$$p(\lambda|\phi) = \frac{f(\phi|\lambda)p(\lambda)}{f(\phi)} \tag{14.11}$$

where λ is a continuous variable and ϕ represents the new observed data: $p(\lambda)$ is the *prior* distribution of λ; $p(\lambda|\phi)$ is the *posterior* distribution of λ, given ϕ; and $f(\phi|\lambda)$ is the sampling distribution of ϕ, given λ.

Let λ denote failure rate and t denote successful test time. Let the density function for λ be gamma distributed with

$$p(\lambda) = \frac{t}{\Gamma(a)}(\lambda t)^{a-1}\exp(-\lambda t)$$

If the prior parameters are a_0, t_0, then the prior mean failure rate is $\mu = a_0/t_0$ and the prior variance is $\sigma^2 = a_0/t_0^2$ (appropriate symbol changes in Eq. 2.28). The posterior will also be gamma distributed with parameters a_1 and t_1 where $a_1 = a_0 + n$ and $t_1 = t_0 + t$ and n is the number of events in the interval from 0 to t. The confidence limits on the posterior mean are $\chi^2_{(\alpha,2a_1)}/2t_1$.

Example 14.4

The prior estimate of the failure rate of an item is 0.02, with a standard deviation of 0.01. A reliability demonstration test results in $n = 14$ failures in $t = 500$ h. What is the posterior estimate of failure rate and the 90 % lower confidence limit?

The prior mean failure rate is

$$\mu = \frac{a_0}{t_0} = 0.02\,\text{h}^{-1}$$
$$\sigma^2 = \frac{a_0}{t_0^2} = 10^{-4}\,\text{h}^{-2}$$

Therefore,

$$a_0 = \frac{\mu^2}{\sigma^2} = 4.0 \text{ failures}$$
$$t_0 = \frac{\mu}{\sigma^2} = \frac{0.02}{10^{-4}} = 200\,\text{h}$$
$$a_1 = 4 + 14 = 18 \text{ failures}$$
$$t_1 = 200 + 500 = 700\,\text{h}$$

The posterior estimate for failure rate is

$$\lambda_1 = \frac{a_1}{t_1} = \frac{18}{700} = 0.0257\,\text{h}^{-1}$$

This compares with the traditional estimate of the failure rate from the test result of $14/500 = 0.028\,\text{h}^{-1}$. The 90 % lower confidence limit on the mean is

$$\frac{\chi^2_{(0.1,\ 2\times 18)}}{2t_1} = \frac{CHIINV(0.1, 36)}{2\times 700} = \frac{47.2}{2\times 700} = 0.0337\,\text{h}^{-1}$$

compared with the traditional estimate (from Table 14.2) of

$$\lambda_1 = \frac{\chi^2_{(0.10,\ 2\times 14+2)}}{2 \times 500} = \frac{CHIINV(0.1, 30)}{2 \times 500} = \frac{40.3}{2 \times 500} = 0.0403\text{h}^{-1}$$

In Example 14.4 use of the prior information has resulted in a failure rate estimate lower than that given by the test, and closer confidence limits.

The Bayesian approach is somewhat controversial in reliability engineering, particularly as it can provide a justification for less reliability testing. For example, Kleyner *et al.* (1997) proposed a method to reduce sample size required for a success run test in order to demonstrate target reliability with a specified confidence. Choosing a prior distribution based on subjective judgement, expert opinion, or other test or field experience can also be contentious. Combining subassembly test results in this way also ignores the possibility of interface problems.

Another downside of the method is that an unfavourable prior can actually have the opposite effect on reliability demonstration, that is increase the required test time or sample size. The Bayesian approach is not normally part of a product validation program; however there have been efforts to incorporate it in reliability standards. For example Yates (2008) describes the work on an Australian defense standard for Bayesian reliability demonstration.

14.9 Non-Parametric Methods

Non-parametric statistical techniques (Section 2.13) can be applied to reliability measurement. They are arithmetically very simple and so can be useful as quick tests in advance of more detailed analysis, particularly when no assumption is made of the underlying failure distribution.

14.9.1 The C-Rank Method

If n items are tested and k fail, the reliability of the sample is

$$R_C \approx 1 - [C\text{– rank of the } (k+1)\text{th ordered value in } (n+1)] \tag{14.12}$$

where C denotes the confidence level required, using the appropriate rank (for median ranks see Chapter 3, Eq. (3.5) and Appendix 4 for 5 % and 95 % ranks).

Example 14.5

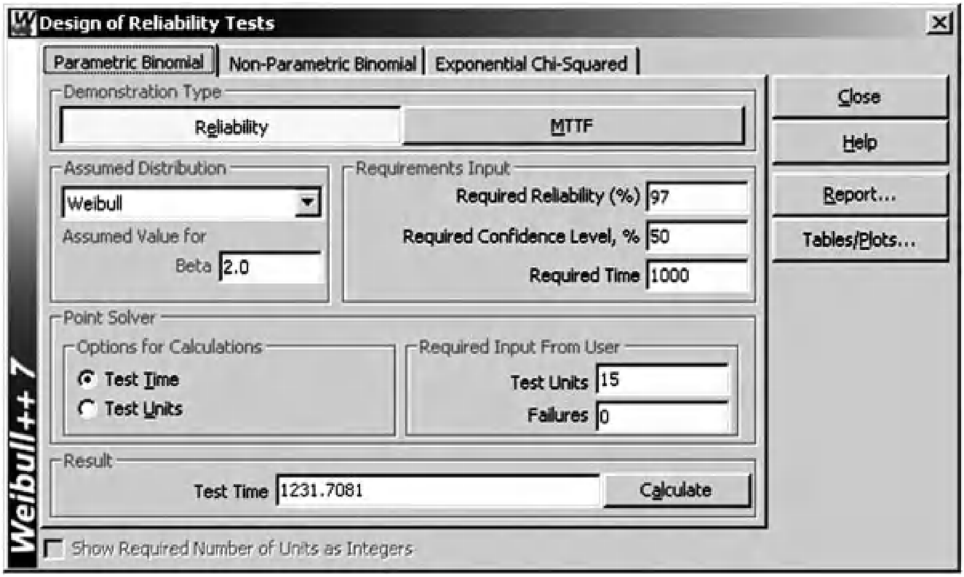
Twenty items were subjected to a 100 h test in which three failed. What is the reliability at the 50 % and 95 % lower confidence levels?

$$k + 1 = 3 + 1 = 4$$
$$n + 1 = 20 + 1 = 21$$

From Eq. (3.5) and Appendix 4, the median and 95 % rank tables, the C-rank of four items in a sample of 21 is:

$$\text{At } 50\%\text{: } 0.172 \text{ (i.e. } R_{50} \approx 0.828)$$
$$\text{At } 95\%\text{: } 0.329 \text{ (i.e. } R_{95} \approx 0.671)$$

$$(\text{cf.} 17/20 = 0.85 \text{ for } \hat{R})$$



Demonstration Type: Reliability | MTTF

Assumed Distribution: Weibull

Assumed Value for Beta 2.0

Requirements Input

Required Reliability (%) 97

Required Confidence Level, % 50

Required Time 1000

Point Solver

Options for Calculations

● Test Time

○ Test Units

Required Input From User

Test Units 15

Failures 0

Result

Test Time 1231.7081

□ Show Required Number of Units as Integers

Figure 14.2 *Solution to Example 14.2 using Weibull++® reliability demonstration calculator DRT (reproduced by permission of ReliaSoft).*

14.10 Reliability Demonstration Software

Various software packages are available to conduct quick reliability demonstration calculations according to the methods covered above. For example, Weibull++® has a built-in calculator called DRT (Design of Reliability Tests) allowing the user to expeditiously estimate the required test durations, sample sizes, confidence limits, and so on based on the available information. Figure 14.2 shows the solution to Example 14.2 using the DRT option. It calculates the extended test time using the parametric binomial model. Some of those calculation features are available in Minitab®, CARE® by BQR, WinSMITH®, Reliass® and others.

14.11 Practical Aspects of Reliability Demonstration

In the industrial setting involving suppliers and customers, test and validation programmes including reliability demonstration often require customer approvals, which sometimes become a source of contention. In some instances the customer may set the reliability targets to a level which is not easy to demonstrate by a reasonable amount of testing.

It is important to remember that demonstrating high reliability is severely limited by the test sample size required (see Table 14.1) no matter which method is chosen and therefore by the amount of money available to spend on testing activities. Moreover, the issue of reliability demonstration gets even more confused when the customer directly links it with reliability prediction (see Kleyner and Boyle (2004) on reliability demonstration vs. reliability prediction). As mentioned in Chapter 6 most reliability prediction methods are based on generic component failure rates and therefore can generate values that have large uncertainty. Also, there is often uncertainty with regard to demonstrated reliability. Therefore discrepancies between the two can be expected. In order to clarify some of those issues, following are the arguments against attempting to demonstrate high reliability by testing a statistically significant number of units.

Firstly, for example a reliability demonstration of $R = 99.9\,\%$ implies 0.1 % accuracy, which cannot possibly be obtained economically or practically with the methods described. Most of the tests performed by reliability engineers are accelerated tests with all the uncertainties associated with testing under conditions different from those in the field, the greatest contributor to which would be the field to test correlation. In other words, based on a test lasting from several hours to several weeks, we are trying to draw conclusions about the behaviour of the product in the field for the next 10–15 years. With so many unknown factors the overall uncertainty well exceeds 0.1 %.

Secondly, system interaction problems contribute heavily to warranty claims. The analysis of warranties (Chapter 13, Figure 13.9) shows that reliability related problems comprise only part of the field problems, thus even if such a high reliability is demonstrated, it would not nearly guarantee that kind of performance in the field, since many other failure factors would be present.

Thirdly, calculations involving reliability demonstration are usually based on lower confidence bounds, therefore by demonstrating 95 % reliability at a lower confidence bound we are in fact demonstrating $\mathrm{R} \geq 0.95$. Therefore the actual field reliability will most likely be higher than the demonstrated number. Demonstration testing of more units will not generally make the product more reliable, but designing for reliability will.

Fourthly, test requirements are often developed based on stresses corresponding to environmental conditions and user profiles well above average severity (see Section 7.3.2.1). Therefore the demonstrated reliability will be much higher when test results are applied to the population of users from all segments of the usage distribution.

And lastly, when very high reliability is specified (e.g. $\mathrm{R} > 0.999$ in safety-related applications), reliability demonstration by test may be totally impracticable. Due to the test sample size limitations, analysis methods such as reliability modelling and simulation, finite element analysis, FMECA and others should be emphasized.

All of these points are reasons why quantitative reliability demonstrations should be used with care. It is essential that possible points of contention, such as definitions of failures, what can and cannot be demonstrated by test, and others, are agreed in advance.

14.12 Standard Methods for Repairable Equipment

This section describes standard methods of test and analysis which are used to demonstrate compliance with reliability requirements.

The standard methods are not substitutes for the statistical analysis methods described earlier in this chapter. They may be referenced in procurement contracts, particularly for government equipment, but they may not provide the statistical engineering insights given, for example, by life data analysis. Therefore the standards should be seen as complementary to the statistical engineering methods and useful (or mandatory) for demonstrating and monitoring reliability of products which are into or past the development phase.

14.12.1 Probability Ratio Sequential Test (PRST) (US MIL-HDBK-781)

The best known standard method for formal reliability demonstration testing for repairable equipment which operates for periods of time, such as electronic equipment, motors, and so on, is US MIL-HDBK-781:

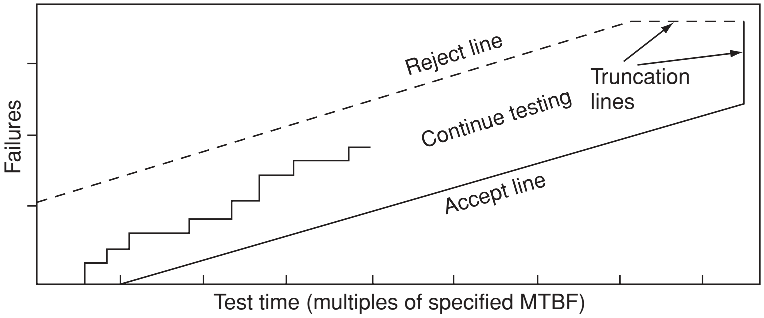


Figure 14.3 *Typical probability ratio sequential test (PRST) plan.*

Reliability Testing for Engineering Development, Qualification and Production (see Bibliography). It provides details of test methods and environments, as well as reliability growth monitoring methods (see Section 14.13).

MIL HDBK-781 testing is based on *probability ratio sequential testing* (PRST), the results of which (failures and test time) are plotted as in Figure 14.3. Testing continues until the 'staircase' plot of failures versus time crosses a decision line. The reject line (dotted) indicates the boundary beyond which the equipment will have failed to meet the test criteria. Crossing the accept line denotes that the test criteria have been met. The decision lines are truncated to provide a reasonable maximum test time. Test time is stated as multiples of the specified MTBF. British and international standards for reliability demonstration are based on MIL-HDBK-781 (see Bibliography).

14.12.2 Test Plans

MIL-HDBK-781 contains a number of test plans, which allow a choice to be made between the statistical risks involved (i.e. the risk of rejecting an equipment with true reliability higher than specified or of accepting an equipment with true reliability lower than specified) and the ratio of minimum acceptable to target reliability. The risk of a good equipment (or batch) being rejected is called the *producer's risk* and is denoted by α. The risk of a bad equipment (or batch) being accepted is called the *consumer's risk*, denoted by β. MIL-HDBK-781 test plans are based upon the assumption of a constant failure rate, so MTBF is used as the reliability index. Therefore MIL-HDBK-781 tests are appropriate for equipment where a constant failure rate is likely to be encountered, such as fairly complex maintained electronic equipment, after an initial burn-in period. Such equipment in fact was the original justification for the development of the precursor to MIL-HDBK-781, the AGREE report (Section 1.8). If predominating failure modes do not occur at a constant rate, tests based on the methods described in Chapter 2 should be used. In any case it is a good idea to test the failure data for trend, as described in Section 2.15.1.

The criteria used in MIL-HDBK-781 are:

1 Upper test MTBF, θ_0. This is the MTBF level considered 'acceptable'.
2 Lower test MTBF, θ_1. This is the specified, or contractually agreed, minimum MTBF to be demonstrated.
3 Design ratio, $d = \theta_0/\theta_1$.
4 Producer's risk, α (the probability that equipment with MTBF higher than θ_1 will be rejected).
5 Consumer's risk β (the probability that equipment with MTBF lower than θ_0 will be accepted).

Table 14.3 MIL-HDBK-781 PRST plans.

	Decision risks (%)		
Test plan	α	β	Design ratio, $d = \theta_0/\theta_1$
I	10	10	1.5
II	20	20	1.5
III	10	10	2.0
IV	20	20	2.0
V	10	10	3.0
VI	20	20	3.0
VII[a]	30	30	1.5
VIII[a]	30	30	2.0

[a] Test plans VII and VIII are known as short-run high risk PRST plans.

The PRST plans available in MIL-HDBK-781 are shown in Table 14.3. In addition, a number of fixed-length test plans are included (plans IX–XVI), in which testing is required to be continued for a fixed multiple of design MTBF. These are listed in Table 14.4. A further test plan (XVII) is provided, for production reliability acceptance testing (PRAT), when all production items are to be tested. The plan is based on test plan III. The test time is not truncated by a multiple of MTBF but depends upon the number of equipments produced.

14.12.3 Statistical Basis for PRST Plans

PRST is based on the statistical principles described in Section 14.6 and the assumption of a constant failure rate. The decision risks are based upon the risks that the estimated MTBF will not be more than the upper test MTBF (for rejection), or not less than the lower test MTBF (for acceptance).

We thus set up two null hypotheses:

$$\text{For } H_0 : \hat{\theta} \leq \theta_0$$
$$\text{For } H_1 : \hat{\theta} \geq \theta_1$$

Table 14.4 MIL-HDBK-781 fixed length test plans.

	Decision risks (%)					
Test plan	α	β	Design ratio, d	Test duration, X θ_1	Reject > failures	Accept < failures
IX	10	10	1.5	45.0	37	36
X	20	20	1.5	21.1	18	17
XI	10	10	2.0	18.8	14	13
XII	20	20	2.0	7.8	6	5
XIII	30	30	2.0	3.7	3	2
XIV	10	10	3.0	9.3	6	5
XV	20	20	3.0	4.3	3	2
XVI	30	30	3.0	1.1	1	0

The probability of accepting H_0 is $(1 - \alpha)$, if $\hat{\theta} = \theta_1$; the probability of accepting H_1 is β, if $\hat{\theta} = \theta_0$. The time at which the i^{th} failure occurs is given by the exponential distribution function $f(t_i) = (1/\theta)\exp(-t_i/\theta)$. The *sequential probability ratio*, or ratio of the expected number of failures given $\theta = \theta_0$ or θ_1, is

$$\prod_{i=1}^{n} \frac{(1/\theta_1)\exp(-\ t_i/\theta_1)}{(1\big/\theta_0 \exp(-\ t_i\big/\theta_0)} \tag{14.13}$$

where n is the number of failures.

The upper and lower boundaries of any sequential test plan specified in terms of θ_0, θ_1, α and β can be derived from the sequential probability ratio. However, arbitrary truncation rules are set in MIL-HDBK-781 to ensure that test decisions will be made in a reasonable time. The truncation alters the accept and reject probabilities somewhat compared with the values for a non-truncated test, and the α and β rules given in MIL-HDBK-781 are therefore approximations. The exact values can be determined from the *operating characteristic* (OC) curve appropriate to the test plan and are given in MIL-HDBK-781. The OC curves are described in the next section.

14.12.4 Operating Characteristic Curves and Expected Test Time Curves

An operating characteristic (OC) curve can be derived for any sequential test plan to show the probability of acceptance (or rejection) for different values of true MTBF. Similarly, the expected test time (ETT – time to reach an accept or reject decision) for any value of θ can be derived. OC and ETT curves are given in MIL-HDBK-781 for the specified test plans. Typical curves are shown in Figure 14.4 and Figure 14.5.

14.12.5 Selection of Test Criteria

Selection of which test plans to use depends upon the degrees of risk which are acceptable and upon the cost of testing. For example, during development of new equipment, when the MTBF likely to be achieved might be uncertain, a test plan with 20 % risks may be selected. Later testing, such as production batch acceptance testing, may use 10 % risks. A higher design ratio would also be appropriate for early development reliability

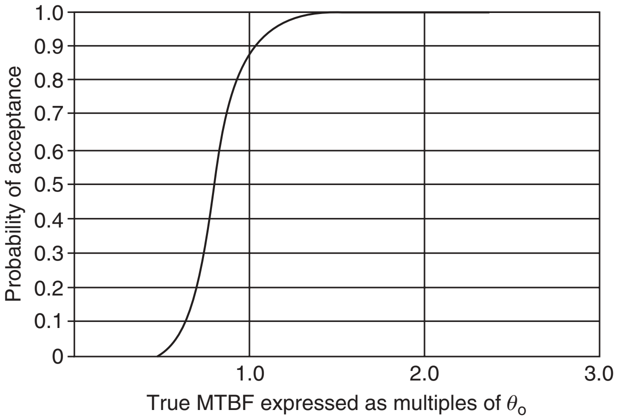


Figure 14.4 *Operating characteristic (OC) curve. Test plan 1: $\alpha = 10\ \%$, $\beta = 10\%$ and $d = 1.5$.*

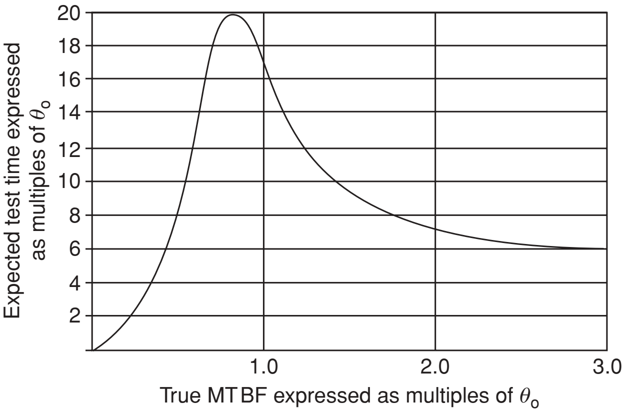


Figure 14.5 *Expected test time (ETT) curve. Test plan 1: $\alpha = 10\,\%$, $\beta = 10\%$ and $d = 1.5$.*

testing. The higher the risks (i.e. the higher the values of α and β) and the lower the design ratio, the longer will be the expected test duration and therefore the expected cost.

The design MTBF should be based upon reliability prediction, development testing and previous experience. MIL-HDBK-781 requires that a reliability prediction be performed at an early stage in the development programme and updated as development proceeds. We discussed the uncertainties of reliability prediction in Chapter 6. However, this should only apply to the first reliability test, since the results of this can be used for setting criteria for subsequent tests.

The lower test MTBF may be a figure specified in a contract, as is often the case with military equipment, or it may be an internally generated target, based upon past experience or an assessment of market requirements.

14.12.6 Test Sample Size

MIL-HDBK-781 provides recommended sample sizes for reliability testing. For a normal development programme, early reliability testing (qualification testing) should be carried out on at least two equipments. For production reliability acceptance testing the sample size should be based upon the production rate, the complexity of the equipment to be tested and the cost of testing. Normally at least three equipments per production lot should be tested.

14.12.7 Burn-In

If equipment is burned-in prior to being submitted to a production reliability acceptance test, MIL-HDBK-781 requires that all production equipments are given the same burn-in prior to delivery.

14.12.8 Practical Problems of PRST

Reliability demonstration testing using PRST is subject to severe practical problems and limitations, which cause it to be a controversial method. We have already covered one fundamental limitation: the assumption of a constant failure rate. However, it is also based upon the implication that MTBF is an inherent parameter of a system which can be experimentally demonstrated, albeit within confidence limits. In fact, reliability measurement is subject to the same fundamental constraint as reliability prediction: *reliability is not an* *inherent physical property of a system, as is mass or electric current*. The mass or power consumption of a system is measurable (also within statistical bounds, if necessary). Anyone could repeat the measurement with any copy of the system and would expect to measure the same values. However, if we measure the MTBF of a system in one test, *it is unlikely* that another test will demonstrate the same MTBF, quite apart from considerations of purely statistical variability. In fact there is no logical or physical reason to expect repeatability of such experiments. This can be illustrated by an example.

Suppose a system is subjected to a PRST reliability demonstration. Four systems are tested for 400 hours and show failure patterns as follows:

No. 1	2 memory device failures (20 h, 48 h)
	1 connector intermittent (150 h)
	1 capacitor short circuit (60 h)
No. 2	1 open-circuit PCB track (40 h)
	1 IC socket failure (200 h)
No. 3	No failures
No. 4	1 shorted connector (trapped on assembly) (0 h)

Total failures: 6
Total running time: 1600 h.

Observed MTBF $\hat{\theta} = 267$ h

Note that these failures are quite typical. However, if the experiment were repeated with another four systems, there would be no reason to expect the same number or pattern of failures. If the same four systems were tested for another 1600 hours the pattern of failures would almost certainly be different. The pattern of failures and their likelihood can be influenced by quality control of manufacture and repair. Therefore the MTBF measured in this test is really no more than historical data, related to those four systems over that period of their lives. It does not predict the MTBF of other systems or of those four over a subsequent period, any more than four sales in one day would be a prediction of the next day's sales. If any design or process changes are made as a result of the test, forecasting becomes even more uncertain.

Of course, if a large number of systems were tested we would be able to extrapolate the results with rather greater credibility and to monitor trends (e.g. average failures per system). However, PRST can seldom be extended to such large quantities because of the costs involved.

PRST is often criticized on the grounds that in-service experience of MTBF is very different to the demonstrated figure. From the discussion above this should not surprise anyone. In addition, in-service conditions are almost always very different to the environments of MIL-HDBK-781 testing, despite attempts to simulate realistic conditions in CERT.

PRST is not consistent with the reliability test philosophy described in Chapter 12, since the objective is to count failures and to hope that few occur. An effective reliability test programme should aim at generating failures, *since they provide information on how to improve the product*. Failure-counting should be a secondary consideration to failure analysis and corrective action. Also, a reliability test should not be terminated solely because more than a predetermined number of failures occur. PRST is very expensive, and the benefit to the product in terms of improved reliability is sometimes questionable.

14.12.9 Reliability Demonstration for One-Shot Items

For equipment which operate only once, or cyclically, such as pyrotechnic devices, missiles, fire warning systems and switchgear, the sequential method of testing based on operating time may be inappropriate.

Statistical acceptance sampling methods can be used for such items, as described in Chapter 15. Alternatively, a MIL-HDBK-781 test could be adapted for items which operate cyclically, using a baseline of mean cycles to failure, or MTBF assuming a given cycling rate.

14.13 Reliability Growth Monitoring

14.13.1 The Duane Method

It is common for new products to be less reliable during early development than later in the programme, when improvements have been incorporated as a result of failures observed and corrected. Similarly, products in service often display reliability growth. This was first analysed by J. T. Duane, who derived an empirical relationship based upon observation of the MTBF improvement of a range of items used on aircraft. Duane observed that the cumulative MTBF θ_c (total time divided by total failures) plotted against total time on log–log paper gave a straight line. The slope (α) gave an indication of reliability (MTBF) growth, that is

$$\log \theta_c = log\theta_0 + \alpha(\log T - \log T_0)$$

where θ_0 is the cumulative MTBF at the start of the monitoring period T_0. Therefore,

$$\theta_c = \theta_0 \left(\frac{T}{T_0}\right)^{\alpha} \tag{14.14}$$

The relationship is shown plotted in Figure 14.6.

The slope α gives an indication of the rate of MTBF growth and hence the effectiveness of the reliability programme in correcting failure modes. Duane observed that typically α ranged between 0.2 and 0.4, and that the value was correlated with the intensity of the effort on reliability improvement.

The Duane method is applicable to a population with a number of failure modes which are progressively corrected, and in which a number of items contribute different running times to the total time. Therefore it is not appropriate for monitoring early development testing, and it is common for early test results to show a poor fit to the Duane model. The method is also not consistent with the use of accelerated tests during

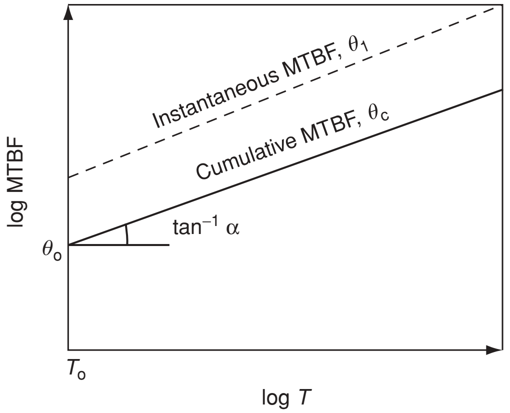


Figure 14.6 Duane reliability growth.

development, since the objective of these is to force failures, not to generate reliability statistics, as described in Chapter 12.

We can derive the instantaneous MTBF θ_i of the population by differentiation of Eq. (14.14)

$$\theta_\mathrm{c} = \frac{T}{k}$$

where k is the number of failures. Therefore,

$$\begin{aligned} k &= \frac{T}{\theta_\mathrm{c}} = \frac{T}{\theta_0 (T/T_0)^\alpha} \\ &= T^{(1-\alpha)} \left(\frac{T_0^\alpha}{\theta_0} \right) \end{aligned}$$

(T_0^α / θ_0) is a constant. Differentiation gives

$$\begin{aligned} \frac{\mathrm{d}k}{\mathrm{d}T} &= (1-\alpha) T^{-\alpha} \left(\frac{T_0^\alpha}{\theta_0} \right) = (1-\alpha) \left(\frac{T_0}{T} \right)^\alpha \frac{1}{\theta_0} \\ &= \frac{1-\alpha}{\theta_\mathrm{c}} \\ \frac{\mathrm{d}k}{\mathrm{d}T} &= \frac{1}{\theta_\mathrm{i}} \end{aligned}$$

So

$$\theta_\mathrm{i} = \frac{\theta_\mathrm{c}}{1-\alpha} \tag{14.15}$$

θ_i is shown in Figure 14.6. The plot of θ_i is parallel to that for θ_c. A reliability monitoring programme may be directed towards a target either for cumulative or instantaneous MTBF.

After the end of a development programme in which MTBF growth is being managed, the anticipated MTBF of production items is θ_i, measured at the end of the programme. This assumes that the development testing accurately simulated the expected in-use stresses of the production items and that the standard of items being tested at the end of the development programme fully represents production items. Of course, these assumptions are often not valid and extrapolations of reliability values from one set of conditions to another must always be considered to be tentative approximations. Nevertheless, the empirical Duane method provides a reasonable approach to monitoring and planning MTBF growth for complex systems.

The Duane method can also be used in principle to assess the amount of test time required to attain a target MTBF. If the MTBF is known at some early stage, the test time required can be estimated if a value is assumed for α. The value chosen must be related to the expected effectiveness of the programme in detecting and correcting causes of failure. Knowledge of the effectiveness of past reliability improvement programmes operated by the organization can provide guidance in selecting a value for α. The following may be used as a guide:

- $\alpha = 0.4 - 0.6$. Programme dedicated to the elimination of failure modes as a top priority. Use of accelerated (overstress) tests. Immediate analysis and effective corrective action for all failures.
- $\alpha = 0.3 - 0.4$. Priority attention to reliability improvement. Normal (typical expected stresses) environment test. Well-managed analysis and corrective action for important failure modes.

- $\alpha = 0.2$. Routine attention to reliability improvement. Testing without applied environmental stress. Corrective action taken for important failure modes.
- $\alpha = 0.2$–0. No priority given to reliability improvement. Failure data not analysed. Corrective action taken for important failure modes, but with low priority.

Example 14.6

The first reliability qualification test on a new electronic test equipment generates 11 failures in 600 h, with no one type of failure predominating. The requirement set for the production standard equipment is an MTBF of not less than 500 h in service. How much more testing should be planned, assuming values for α of 0.3 and 0.5?

$$\hat{\theta}_0 = \frac{600}{11} = 54.4\,\text{h}$$

When $\theta_i = 500$,

$$\theta_\text{c} = 500(1 - \alpha)$$

$$\begin{cases} = 350\ (\text{for}\,\alpha = 0.3) \\ = 250\ (\text{for}\,\alpha = 0.5) \end{cases}$$

Using $\theta_0 = 54.4$, from Eq. (14.14),

$$\begin{aligned} \theta_\text{c} &= \theta_0 \left(\frac{T}{T_0}\right)^\alpha \\ T &= T_0 \left(\frac{\theta_\text{c}}{\theta_0}\right)^{1/\alpha} \\ &= 600 \left(\frac{350}{54.4}\right)^{1/0.3} = 297\,200\,\text{h} \quad (\text{for}\,\alpha = 0.3) \\ &= 600 \left(\frac{250}{54.4}\right)^{1/0.5} = 126\,70\,\text{h} \quad (\text{for}\,\alpha = 0.5) \end{aligned}$$

Graphical construction can be used to derive the same result, as shown in Figure 14.7 [$tan^{-1}(0.3) = 17^\circ\ tan^{-1}(0.5) = 27^\circ\ \theta_\text{i} = \theta_\text{c}/(1 - \alpha)$].

Obviously nearly 300 000 h of testing is unrealistic, and therefore in this case a value for α of 0.5 would have to be the objective to achieve the MTBF requirement of 500 h in a further (12 670 − 600) ≈ 12 000 h of testing.

Example 14.6 shows that the results of a Duane analysis are very sensitive to the starting assumptions. If θ_0 was 54.4 h at $T_0 = 200$h, the test time required for a 500 h MTBF would be 4200 h. The initial reliability figure is usually uncertain, since data at the early stage of the programme are limited. It might be more appropriate to use a starting reliability based upon a combination of data and engineering judgement. If in the previous example immediate corrective action was being taken to remove some of the causes of earlier failures, a higher value of θ_0 could have been used. It is important to monitor early reliability growth and to adjust the plan accordingly as test results build up.

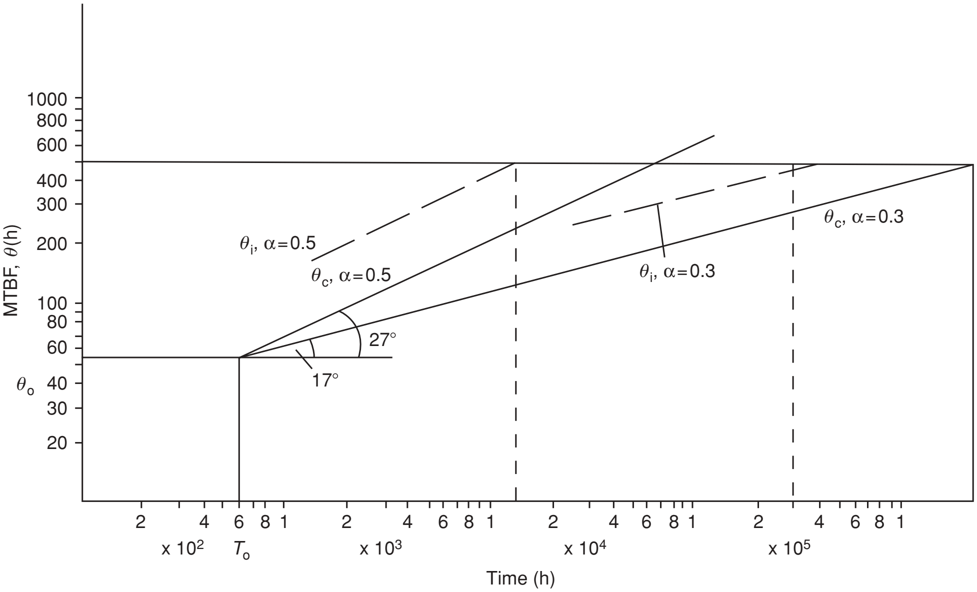


Figure 14.7 *Duane plot for Example 14.6.*

The Duane model is criticized as being empirical and subject to wide variation. It is also argued that reliability improvement in development is not usually progressive but occurs in steps as modifications are made. However, the model is simple to use and it can provide a useful planning and monitoring method for reliability growth. Difficulties can arise when results from different types of test must be included, or when corrective action is designed but not applied to all models in the test programme. These can be overcome by common-sense approaches. A more fundamental objection arises from the problem of quantifying and extrapolating reliability data. The comments made earlier about the realism of reliability demonstration testing apply equally to reliability growth measurement. As with any other failure data, trend tests as described in Chapter 2 should be performed to ascertain whether the assumption of a constant failure rate is valid.

Other reliability growth models are also used, some of which are described in MIL-HDBK-781, which also describes the management aspects of reliability growth monitoring. Reliability growth monitoring for one-shot items can be performed similarly by plotting cumulative success rate. Statistical tests for MTBF or success rate changes can also be used to confirm reliability growth, as described in Chapter 2. Example 14.7 shows a typical reliability growth plan and record of achievement.

Example 14.7

Reliability growth plan:

Office Copier Mk 4

Specification:

In-use call rate: 2 per year max. (at end of development)
1 per year max. (after first year)
Average copies per machine per year: 40 000

Assumptions:

$$\theta_0 = 1000 \text{ copies per failure at 10 000 copies on prototypes}$$
$$\alpha \begin{cases} = \ 0.5 \text{ during development} \\ = \ 0.3 \text{ in service} \end{cases}$$

Notes to Duane plot (Figure 14.8)

1 Prototype reliability demonstration: models 2 and 3:10 000 copies each.
2 First interim reliability demonstration: models 6–8, 10:10 000 copies each.
3 Accelerated and ageing tests (data not included in θ_{c}).
4 Second interim reliability demonstration: models 8, 10, 12, 13:10 000 copies each.
5 Accelerated and ageing tests (data not included in θ_{c}).
6 Final reliability demonstration: models 12, 13:10 000 copies each. Models 8, 10:20 000 copies each.

Reliability demonstration test results give values for θ_{i}.

Note that in Example 14.7 the accelerated stress test results are plotted separately, so that the failures during these tests will not be accumulated with those encountered during tests in the normal operating environment. Therefore, accelerated stress failure data will be obtained to enable potential in-use failure modes to be highlighted, without confusing the picture as far as measured reliability achievement is concerned. The improvement from the first to the second accelerated stress test has a higher Duane slope than the main reliability growth line, indicating effective improvement. The example also includes a longevity test, to show up potential wearout failure modes, by having two of the first reliability demonstration units continue to undergo test in the second interim and final reliability demonstrations. These units will have been modified between tests to include all design improvements shown to be necessary. A lower value for α is assumed for the in-service phase, as improvements are more difficult to implement once production has started.

14.13.2 The M(t) Method

The $M(t)$ method of plotting failure data is a simple and effective way of monitoring reliability changes over time. It is most suitable for analysing the reliability performance of equipment in service. Figure 14.9(a) shows a typical situation in which new equipment is introduced to service over a period (calendar time), and suffer failures which incur down time (repair, delays, etc.). Figure 14.9(b) shows the same population, but now with operating time (calendar time minus down time, or operating hours, as appropriate) as the horizontal scale. The *population graph* (Figure 14.9(c)) shows the systems at risk as a function of operating time.

$M(t)$ is calculated at each failure from the formula

$$M(t_i) = M(t_{i-1}) + 1/N(t_i) \tag{14.16}$$

where: $M(t_i)$ is the value of M at operating time t_i.
$M(t_{i-1})$ is the preceding value of M.
$N(t_i)$ is the number of equipments in service at operating time t_i.

$M(t)$ is the mean accumulated number of failures as a function of operating time.

$M(t_i)$ can also be calculated for groups of failures occurring over intervals, by replacing $1/N(t_i)$ with $\Delta r(t_i)/N(t_i)$, where $\Delta r(t_i)$ is the failures that have occurred in the time interval between t_{i-1} and t_i.

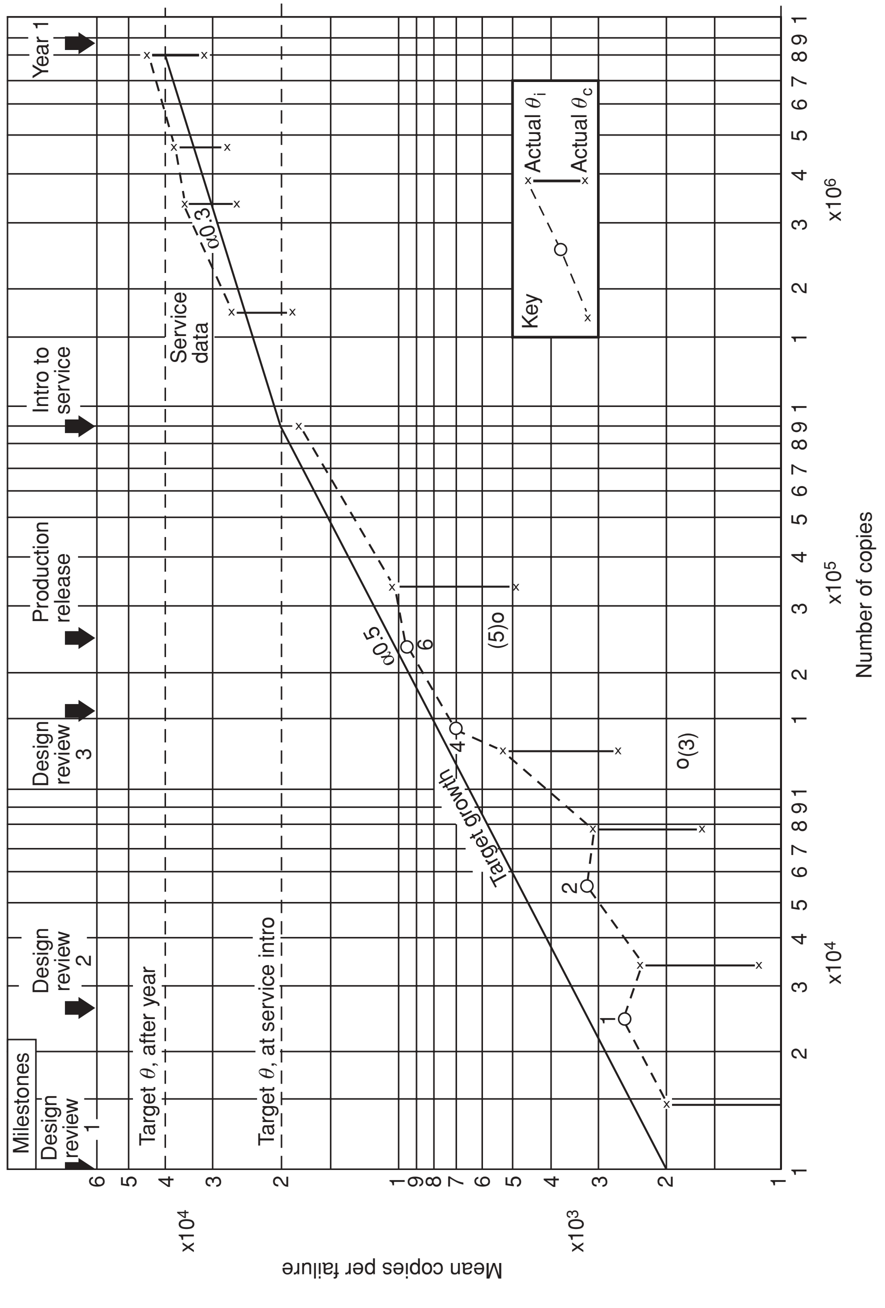


Figure 14.8 Duane plot for Example 14.7.

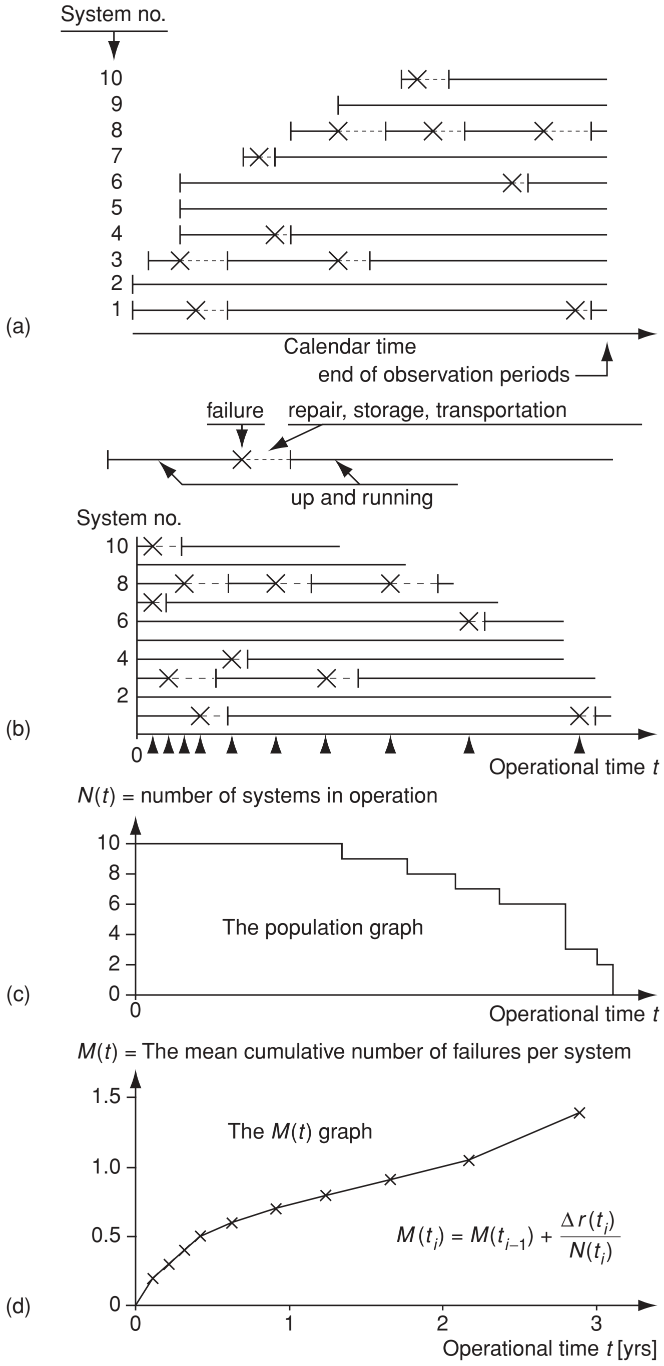


Figure 14.9 *The M(t) analysis method (Courtesy J. Møltoft).*

Figure 14.9(d) is the $M(t)$ graph. The slope of the line indicates the proportion per time unit failing at that time, or the failure intensity. Reliability improvement (either due to a decreasing instantaneous failure rate pattern or to reliability improvement actions) will reduce the slope. A straight line indicates a constant (random) pattern. An increasing slope indicates an increasing pattern, and vice versa Changes in slope indicate changing trends: for example, Figure 14.9(d) is typical of an early 'infant mortality' period, caused probably by manufacturing problems, followed by a constant failure intensity. The failure intensity over any period can be calculated by measuring the slope. The proportions failing over a period can be determined by reading against the $M(t)$ scale.

The $M(t)$ method can be used to monitor reliability trends such as the effectiveness of improvement actions. Example 14.8 illustrates this.

Example 14.8

Table 14.5 shows data on systems in service. Figure 14.10a and Figure 14.10b show $\Delta r(t)$ and $N(t)$ respectively, and Figure 14.10c is the $M(t)$ plot.

From the $M(t)$ graph we can determine that:

1 Over the first 400 h no failures occur, but then failures occur at a fairly high and increasing intensity thereafter. This could indicate a failure-free period for the population.
2 The failure intensity continues at a roughly constant level (about 2.5 failures/unit/1000 h), until 1400 h, when it drops to about 1 failure/unit/1000 h. This might be due to the weeding out of the flaws (latent defects) in the systems, modification to units in service, improved maintenance, and so on.
3 By extrapolating the asymptotic part of the curve back to the $M(t)$ axis, we see that on average each system has failed and been repaired about twice.

The $M(t)$ method can be useful for identifying and interpreting failure trends. It can also be used for evaluating logistics and warranty policies. For example, if the horizontal scale represents calendar time, the expected number of failures over selected periods can be determined by reading the intercepts on the $M(t)$ scale (e.g. in Example 14.8 a linear extrapolation of the final slope would indicate about 1 failure/unit/1000 h of the matured systems and about 12 % flaws). The method is described in Møltoft (1994).

Table 14.5 Service data.

Units at risk ($N(t)$)	Operating time t (hours)	Failures ($\Delta r(t)$)	$\Delta r(t)/N(t)$	$M(t)$
105	0	0	0.00	0.00
105	400	0	0.00	0.00
105	600	4	0.04	0.04
85	800	10	0.12	0.16
65	1000	17	0.26	0.42
45	1200	22	0.49	0.91
35	1400	18	0.51	1.42
25	1600	5	0.20	1.62
15	1800	3	0.20	1.82
5	2000	1	0.20	2.02

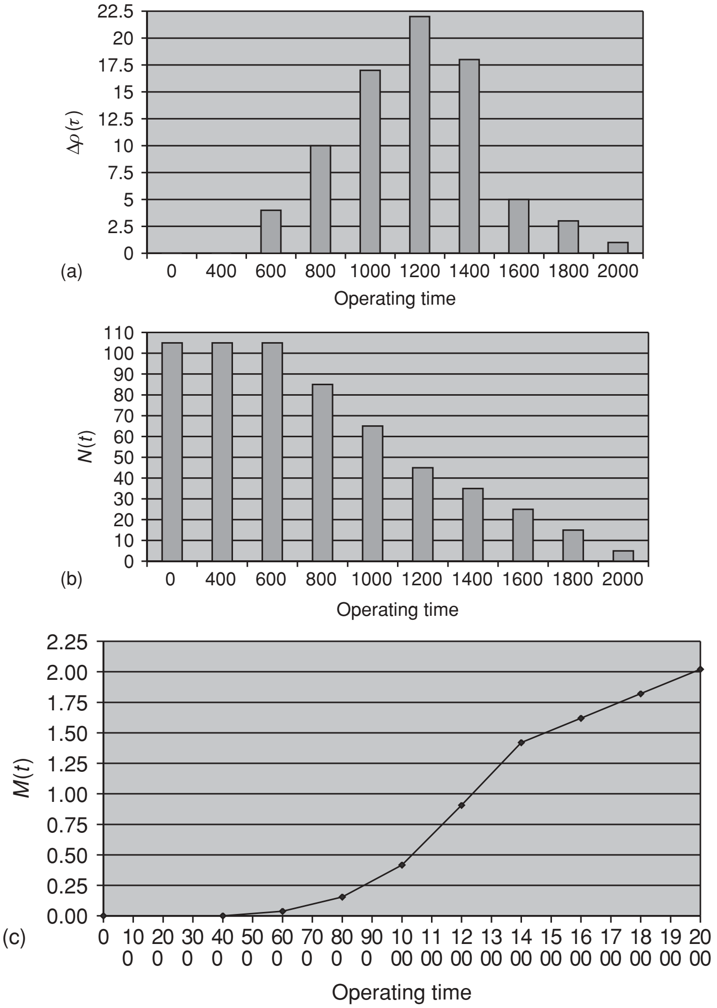


Figure 14.10 *M(t) graph of Table 14.5.*

14.13.3 Reliability Growth Estimation by Failure Data Analysis

Reliability growth can be estimated by considering the failure data and the planned corrective action. No empirical model is used and the method takes direct account of what is known and planned, so it can be easier to sell. However, it can only be applied when sufficient data are available, well into a development programme or when the product is in service.

If we know that 20 % of failures are caused by failure modes for which corrective action is planned and we are sure that the changes will be effective, we can simply estimate that the improvement in failure rate will be 20 %. Alternatively, we could assign an effectiveness value to the changes, say 80 %, in which case the failure rate improvement will be 16 %.

This approach should be used whenever failure data and failure investigations are comprehensive enough to support a Pareto analysis, as described in Section 13.2. The method can be used in conjunction with a Duane plot. If known failure modes can be corrected, reliability growth can be anticipated. However, if reliability is below target and no corrective action is planned, the reliability growth forecast will not have much meaning.

14.14 Making Reliability Grow

In this chapter we have covered methods for measuring reliability achievement and growth. Of course it is not enough just to measure performance. Effort must be directed towards maximizing reliability. In reliability engineering this means taking positive action to unearth design and production shortfalls which can lead to failure, to correct these deficiencies and to prove that the changes are effective. In earlier chapters we have covered the methods of stress analysis, design review, testing and failure data analysis which can be used to ensure a product's reliability. To make these activities as effective as possible, it is necessary to ensure that they are all directed towards the reliability achievement which has been specified.

There is a dilemma in operating such a programme. There will be a natural tendency to try to demonstrate that the reliability requirements have been met. This can lead to reluctance to induce failures by increasing the severity of tests and to a temptation to classify failures as non-relevant or unlikely to recur. On the other hand, reliability growth is maximized by deliberate and aggressive stress-testing, analysis and corrective action as described in Chapter 12. Therefore, the objective should be to stimulate failures during development testing and not to devise tests and failure-reporting methods whose sole objective is to maximize the chances of demonstrating that a specification has been satisfied. Such an open and honest programme makes high demands on teamwork and integrity, and emphasizes the importance of the project manager understanding the aims and being in control of the reliability programme.[1] The reliability milestones should be stated early in the programme and achievement should be monitored against this statement.

14.14.1 Test, Analyse and Fix

Reliability growth programmes as described above have come to be known as test, analyse and fix (TAAF). It is very important in such programmes that:

1 All failures are analysed fully, and action taken in design or production to ensure that they should not recur. No failure should be dismissed as being 'random' or 'non-relevant' during this stage, unless it can be demonstrated conclusively that such a failure cannot occur on production units in service.
2 Corrective action must be taken as soon as possible on all units in the development programme. This might mean that designs have to be altered more often, and can cause programme delays. However, if faults are not corrected reliability growth will be delayed, potential failure modes at the 'next weakest link' may not be highlighted, and the effectiveness of the corrective action will not be adequately tested.

Action on failures should be based on a disciplined FRACAS (Section 12.6) and Appendix 5.

Whenever failures occur, the investigation should refer back to the reliability predictions, stress analyses and FMECAs to determine if the analyses were correct. Discrepancies should be noted and corrected to aid future work of this type.

[1] The politics of test planning to ensure accept decisions for contractual or incentive purposes is an aspect of reliability programme management which will not be covered here.

14.14.2 Reliability Growth in Service

The same principles as described above should be applied to reliability growth in service. However, there are three main reasons why in-service reliability growth is more difficult to achieve than during the development phase:

1 Failure data are often more difficult to obtain. Warranty or service contract repair reports are a valuable source of reliability data, but they are often harder to control, investigation can be more difficult with equipment in the users' hands and data often terminate at the end of the warranty period. (Use of warranty data was covered in detail in Chapter 13). Some companies make arrangements with selected dealers to provide comprehensive service data. Military and other government customers often operate their own in-use failure data systems. However, in-use data very rarely match the needs of a reliability growth programme.
2 It is much more difficult and much more expensive to modify delivered equipment or to make changes once production has started.
3 A product's reputation is made by its early performance. Reliance on reliability growth in use can be very expensive in terms of warranty costs, reputation and markets.

Nevertheless, a new product will often have in-service reliability problems despite an intensive development programme. Therefore reliability data must be collected and analysed, and improvements designed and implemented. Most products which deserve a reliability programme have a long life in service and undergo further evolutionary development, and therefore a continuing reliability improvement programme is justifiable. When evolutionary development takes place in-use data can be a valuable supplement to further reliability test data, and can be used to help plan the follow-on development and test programme. The FRACAS described in Appendix 5 can be used for in-service failures.

Another source of reliability data is that from production test and inspection. Many products are tested at the end of the production line and this includes burn-in for many types of electronic equipment. Whilst data from production test and inspection are collected primarily to monitor production quality costs and vendor performance, they can be a useful supplement to in-use reliability data. Also, as the data collection and the product are still under the manufacturer's control, faster feedback and corrective action can be accomplished.

The manufacturer can run further tests of production equipment to verify that reliability and quality standards are being maintained. Such tests are often stipulated as necessary for batch release in government production contracts. As in-house tests are under the manufacturer's control, they can provide early warning of incipient problems and can help to verify that the reliability of production units is being maintained or improved.

Questions

1. Compare the B_5-life of 10 years and MTBF = 150 years. Which metric gives the higher reliability at 10 years?
2. How many test units are required to demonstrate 97.5 % reliability at 50 % confidence if no failures are allowed?
3. You are planning to conduct a success run test with 50 samples:
   a What reliability can you demonstrate with 90 % confidence?
   b What would be your demonstrated reliability if during the test you experience two failures?

4. The cost of product validation is \$ 1600 per test sample (includes the cost to produce, test, monitor and analyse the unit). How much extra cost will be incurred if the reliability demonstration requirement changes from R = 95.0 % to 97.0 % at the same 90 % confidence level?
5. Reliability specification calls for reliability demonstration of R = 95.0 % with the confidence C = 90 %. The accelerated thermal cycling test equivalent to one field life consists of 1000 cycles. Calculate the following:
   a How many samples need to be tested without failures in order to meet this requirement?
   b If your chamber can only fit 30 units, how many cycles you would need to run to meet this requirement with this quantity of units? From the past test experience with this product you know the Weibull slope $\beta = 2.5$.
6. Three electric transformers were tested under high temperature and humidity conditions. The test units were inspected at the beginning of the test (time = 0), at 500 hours, and at the end of the test (time = 1000 hours). The unit is considered failed if the inductance falls below 32 μH. The inductance values are presented below:

Item/Time	0 Hours	500 Hours	1000 Hours
Unit 1	38 μH	36 μH	34 μH
Unit 2	42 μH	38 μH	36 μH
Unit 3	38 μH	35 μH	33 μH

   Determine:
   a The expected failure times for each unit using the linear extrapolation.
   b Conduct life data analysis (2-parameter Weibull) and determine β and η parameters.
   c Calculate the B_{10} life.
   d Explore the exponential and power extrapolations. Compare with the results obtained in (a), (b) and (c).
7. a Explain why and under what circumstances it might be valid to assume the exponential distribution for interfailure times of a complex repairable system even though it may contain ‘wearout’ components.
   b Such a system has, on test, accumulated 1053 h of running during which there have been two failures. Estimate the MTBF of the system, and its lower 90 % confidence limit.
   c On the assumption that no more failures occur, how much more testing is required to demonstrate with 90 % confidence that the true MTBF is not less than 500 h? Comment on the implications of your answer.
8. a In a complex repairable system 1053 h of testing have been accumulated, with failures at 334 h and 891 h. Assuming constant failure rate, calculate (i) the current estimate of the system failure rate; (ii) the current estimate of the mean time between failures (MTBF); and (iii) the lower 90 % confidence limit for the MTBF.
   b For the above system, if there is a specification requirement that the MTBF shall be at least 500 h, and this must be demonstrated at 90 % confidence, how much more test running of the system, without further failure, is required?
9. Five engines are tested to failure. Failure times are 628 hours, 3444 hours, 822 hour, 846 hours, and 236 hours. Assuming a constant failure rate, what is the two-sided 90 % confidence interval for the MTBF?
10. You are running a continuous operation of five turbines for 1000 hours. During that test one turbine has failed at 825 hours and was removed from the test. What is the one-sided 90 % confidence limit on the failure rate?

11. Explain the principles of probability ratio sequential testing (PRST) to demonstrate MTBF. What are the main limitations of the method?
12. Explain what is the design ratio in PRST? How is it linked with the risks and the expected test duration?
13. If a Duane growth model has a slope of 0.4 with a cumulative MTBF of 40 000 hours, what is the instantaneous MTBF?
14. What is the major contributor to reliability growth and continuous product and process improvement?
15. A prototype of a repairable system was subjected to a test programme where engineering action is supposedly taken to eliminate causes of failure as they occur. The first 500 h of running gave failures at 12, 36, 80, 120, 200, 360, 400, 440 and 480 hours:
    a Use a Duane plot to discover whether reliability growth is occurring.
    b Calculate the trend statistic (Eq. 2.46) and see whether it gives results consistent with (a).

Bibliography

British Standard, BS 5760. *Reliability of Systems, Equipments and Components*, Part 2. British Standards Institution, London.

Kleyner, A. and Boyle, J. (2004) *The Myths of Reliability Demonstration Testing!* TEST Engineering and Management, August/ September 2004, pp. 16–17.

Kleyner, A., Bhagath, S., Gasparini, M. *et al.* (1997) *Bayesian Techniques to Reduce the Sample Size in Automotive Electronics Attribute Testing*. Microelectronics and Reliability, **37**(6), 879–883.

Lipson, C. and Sheth, N. (1973) *Statistical Design and Analysis of Engineering Experiments*, McGraw-Hill.

Møltoft, J. (1994) *Reliability Engineering Based on Field Information: The Way Ahead*. Quality and Reliability Engineering International, **10**, 399–409.

ReliaSoft (2006) Degradation Analysis. *Web reference*. Available at: http://www.weibull.com/LifeDataWeb/degradation_analysis.htm.

US MIL-HDBK-781. *Reliability Testing for Equipment Development, Qualification and Production*. Available from the National Technical Information Service, Springfield, Virginia.

Wasserman, G. (2003) *Reliability Verification, Testing, and Analysis in Engineering Design*, Marcel Dekker.

Yates, S. (2008) *Australian Defense Standard for Bayesian Reliability Demonstration*. Proceedings of the Annual Reliability and Maintainability Symposium (RAMS), pp. 103–107.

15

Reliability in Manufacture

15.1 Introduction

It is common knowledge that a well-designed product can be unreliable in service because of poor quality of production. Control of production quality is therefore an indispensable facet of an effective reliability effort. This involves controlling and minimizing variability and identifying and solving problems.

Human operations, particularly repetitive, boring or unpleasant tasks, are frequent sources of variability. Automation of such tasks therefore usually leads to quality improvement. Typical examples are paint spraying, welding by robots in automobile production, component placement and soldering in electronic production and CNC machining.

Variability can never be completely eliminated, since there will nearly always be some human operations, and automatic processes are not without variation. A reliable design should cater for expected production variation, so designers must be made aware of the variability inherent in the manufacturing processes to be used.

The production quality team should use the information provided by design analyses, FMECAs and reliability tests. A reliable and easily maintained design will be cheaper to produce, in terms of reduced costs of scrap and rework.

The integration of reliability and manufacturing quality programmes is covered in more detail in Chapter 17.

15.2 Control of Production Variability

The main cause of production-induced unreliability, as well as rework and scrap, is the variability inherent in production processes. In principle, a correct design, correctly manufactured, should not fail in normal use. However, all manual processes are variable. Automatic processes are also variable, but the variability is usually easier to control. Bought-in components and materials also have variable properties. Production quality control (QC) is primarily concerned with measuring, controlling and minimizing these variations in the most cost-effective way.

Statistical process control (SPC) is the term used for the measurement and control of production variability. In SPC, QC people rely heavily on the normal distribution. However, the comments in Sections 2.8.1 and 2.17 should be noted: conventional SPC often ignores the realities discussed.

Practical Reliability Engineering, Fifth Edition. Patrick D. T. O'Connor and Andre Kleyner.

15.2.1 Process Capability

If a product has a tolerance or specification width, and it is to be produced by a process which generates variation of the parameter of interest, it is obviously important that the process variation is less than the tolerance. The ratio of the tolerance to the process variation is called the *process capability*, and it is expressed as

$$C_p = \frac{\text{Tolerance width } (T)}{\text{Process } 3\sigma \text{ limits}}$$

A process capability index of 1 will generate, in theory for normal variation, approximately 0.15 % out of tolerance product, at each extreme (Figure 15.1). A process capability index of greater than 1.33 will theoretically generate about 0.005 % out of tolerance product, or practically 100 % yield.

C_p values assume that the specification centre and the process mean coincide. To allow for the fact that this is not necessarily the case, an alternative index, C_{pk}, is used, where

$$C_{p_k} = (1 - K)C_p$$

and

$$K = \frac{D - \bar{x}}{T/2} \quad (\text{if } D > \bar{x}, \ \text{otherwise use } \bar{x} - D)$$

D being the design centre, $\bar{x}$ the process mean, and T the tolerance width. Figure 15.2 shows examples of C_{pk}. Ideally $C_p = C_{pk}$. Modern production quality requirements typically demand C_{pk} values of 2 or even higher, to provide high assurance of consistent performance. The 'six sigma' approach (Chapter 17) extends the concept even further.

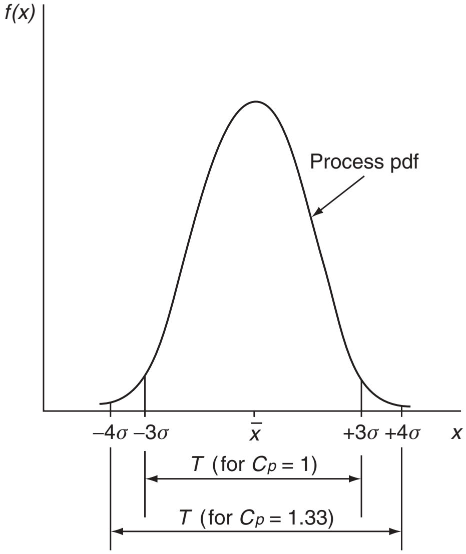


Figure 15.1 *Process capability, C_p.*

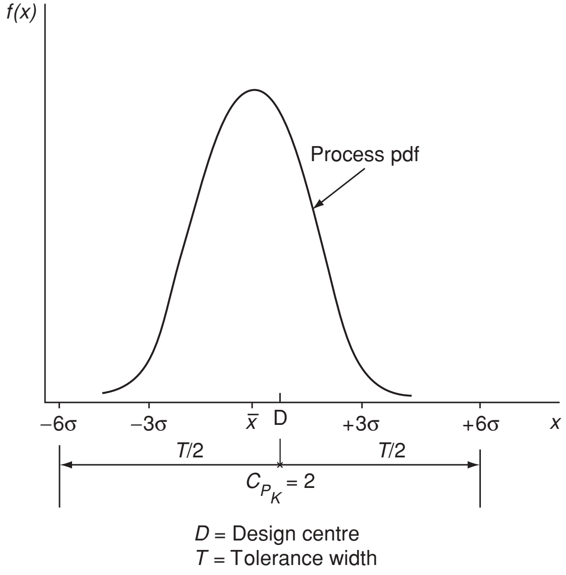


Figure 15.2 *Process capability C_{pk}.*

Use of the process capability index assumes that the process is normally distributed far into the tails and is stationary. Any systematic divergence, due, for example to set-up errors, movement of the process mean during the manufacturing cycle, or other causes, could significantly affect the output. Therefore the use of the capability index to characterize a production process is appropriate only for processes which are under statistical control, that is when there are no *special causes* of variation such as those just mentioned, only *common causes* (Section 2.8.2). Common cause variation is the random variation inherent in the process, when it is under *statistical control*.

The necessary steps to be taken when setting up a production process are:

1 Using the information from the product and process design studies and experiments, determine the required tolerance.
2 Obtain information on the process variability, either from previous production or by performing experiments.
3 Evaluate the process capability index.
4 If the process capability index is high enough, start production, and monitor using statistical control methods, as described below.
5 If C_p/C_{pk} is too low, investigate the causes of variability, and reduce them, before starting production (see Section 15.5).

15.2.2 Process Control Charts

Process control charts are used to ensure that the process is under statistical control, and to indicate when special causes of variation exist. In principle, an in-control process will generate a random fluctuation about the mean value. Any trend, or continuous performance away from the mean, indicates a special cause of variation.

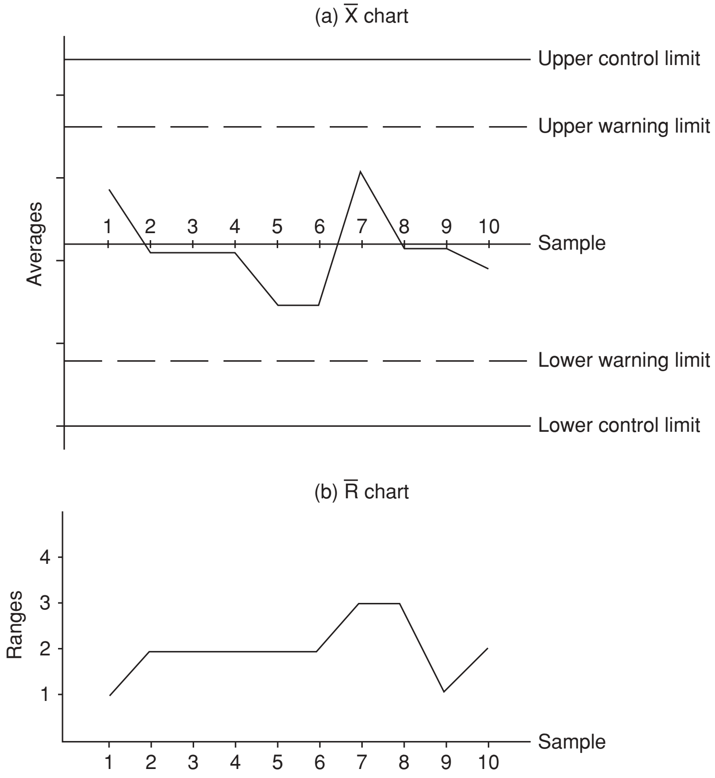


Figure 15.3 *Process control charts.*

Figure 15.3 is an example of a process control chart. As measurements are made the values are marked as points on the control chart against the sample number on the horizontal scale. The data plotted can be individual values or sample averages; when sample averages are plotted the chart is called an *$\bar{x}$ chart*. The $\bar{x}$ chart shows very clearly how the process value varies. Upper and lower control limits are drawn on the chart, as shown, to indicate when the process has exceeded preset limits.

The control limits on an $\bar{x}$ chart are based on the tolerance required of the process. Warning limits are also used. These are set within the control limits to provide a warning that the process might require adjustment. They are based on the process capability, and could be the process 3 σ values ($C_{pk} = 1.0$), or higher. Usually two or more sample points must fall outside the warning limits before action is taken. However, any point falling outside the control limit indicates a need for immediate investigation and corrective action.

Figure 15.3(b) is a *range chart* ($\bar{R}$ chart). The plotted points show the range of values within the sample. The chart indicates the repeatability of the process.

$\bar{x}$ and $\bar{R}$ charts (also called *Shewhart charts*) are the basic tools of SPC for control of manufacturing processes. Their ease of use and effectiveness make them very suitable for use by production operators for controlling their own work, and therefore they are commonly used in operator control and quality circles (see later in this chapter). Computer programs are available which produce $\bar{x}$ and $\bar{R}$ charts automatically when process data are input. Integrated measurement and control systems provide for direct input of measured values to the control chart program, or include SPC capabilities (analysis and graphics).

Statistical process control is applicable to relatively long, stable production runs, so that the process capability can be evaluated and monitored with reasonable statistical and engineering confidence. Methods have, however, been developed for batch production involving smaller quantities. Other types of control chart have also been developed, including variations on the basic Shewhart charts, and non-statistical graphical methods. These are all described in Montgomery (2008), Oakland and Followell (2003), and several other books on SPC.

The most effective application of SPC is the detection of special causes of variation, to enable process improvements to be made. Statistical finesse and precision are not usually important or essential. The methods must be applied carefully, and selected and adapted for the particular processes. Personnel involved must be adequately trained and motivated, and the methods and criteria must be refined as experience develops.

It is important to apply SPC to the processes that influence the quality of the product, not just to the final output parameter, whenever such upstream process variables can be controlled. For example, if the final dimension of an item is affected by the variation in more than one process, these should be statistically monitored and controlled, not just the final dimension. Applying SPC only to the final parameter might not indicate the causes of variation, and so might not provide effective and timely control of the processes.

15.3 Control of Human Variation

Several methods have been developed for controlling the variability inherent in human operations in manufacturing, and these are well documented in the references on quality assurance. Psychological approaches, such as improving motivation by better work organization, exhortation and training, have been used since early industrialization, particularly since the 1940s. These were supported by the development of statistical methods, described earlier.

15.3.1 Inspection

One way of monitoring and controlling human performance is by independent inspection. This was the standard QC approach until the 1950s, and is still used to some extent. An inspector can be made independent of production requirements and can be given authority to reject work for correction or scrap. However, inspection is subject to three major drawbacks:

1 Inspectors are not perfect; they can fail to detect defects. On some tasks, such as inspecting large numbers of solder joints or small assemblies, inspection can be a very poor screen, missing 10 to 50 % of defects. Inspector performance is also as variable as any other human process.
2 Independent inspection can reduce the motivation of production people to produce high quality work. They will be concerned with production quantity, relying on inspection to detect defects.
3 Inspection is expensive. It is essentially non-productive, the staff employed are often more highly paid than production people and output is delayed while inspection takes place. Probably worse, independent inspection can result in an overlarge QC department, unresponsive to the needs of the organization.

These drawbacks, particularly the last, have led increasingly to the introduction of operator control of quality, described below. Automatic inspection aids and systems have also been developed, including computerized optical comparators and automatic gauging systems.

15.3.2 Operator Control

Under operator control, the production worker is responsible for monitoring and controlling his or her own performance. For example, in a machining operation the operator will measure the finished article, log the results and monitor performance on SPC charts. Inspection becomes part of the production operation and worker motivation is increased. The production people must obviously be trained in inspection, measurement and SPC methods, but this is usually found to present few problems. Operator control becomes even more relevant with the increasing use of production machinery which includes measuring facilities, such as self-monitoring computer numerically controlled (CNC) machines.

A variation of operator control is to have production workers inspect the work of preceding workers in the manufacturing sequence, before starting their production task. This provides the advantages of independent inspection, whilst maintaining the advantages of an integrated approach.

15.4 Acceptance Sampling

Acceptance sampling provides a method for deciding whether to accept a particular production lot, based upon measurements of samples drawn at random from the lot. Sampling can be by *attributes* or by *variables*. Criteria are set for the allowable proportion defective, and the sampling risks.

15.4.1 Sampling by Attributes

Sampling by attributes is applicable to go/no-go tests, using binomial and Poisson statistics. Sampling by attributes is covered in standard plans such as ANSI/ASQZ1-4 and BS 6001. These give accept and reject criteria for various sampling plans, based upon sample size and risk levels. The main criterion is the *acceptable quality level* (AQL), defined as the maximum percentage defective which can be accepted as a process average. The tables in the standards give accept and reject criteria for stated AQLs, related to sample size, and for ‘tightened’, ‘normal’ and ‘reduced’ inspection. These inspection levels relate to the consumer’s risk in accepting a lot with a percentage defective higher than the AQL. Table 15.1 shows a typical sampling plan. Some sampling plans are based upon the *lot tolerance percentage defective* (LTPD). The plans provide the minimum sample size to assure, with given risk, that a lot with a percentage defective equal to or more than the specified LTPD will be rejected. LTPD tests give lower consumers’ risks that substandard lots will be accepted. LTPD sampling plans are shown in Table 15.2.

For any attribute sampling plan, an *operating characteristic* curve can be derived. The OC curve shows the power of the sampling plan in rejecting lots with a given percentage defective. For example, Figure 15.4 shows OC curves for single sampling plans for 10 % samples drawn from lots of 100, 200 and 1000, when one or more defectives in the sample will lead to rejection (acceptance number = 0). If the lot contains, say, 2 % defective the probability of acceptance will be 10 % for a lot of 1000, 65 % for a lot of 200, and 80 % for a lot of 100. Therefore the lot size is very important in selecting a sampling plan.

Double sampling plans are also used. In these the reject decision can be deferred pending the inspection of a second sample. Inspection of the second sample is required if the number of defectives in the first sample is greater than allowable for immediate acceptance but less than the value set for immediate rejection. Tables and OC curves for double (and multiple) sampling plans are also provided in the references quoted above.

15.4.2 Sampling by Variables

Sampling by variables involves using actual measured values rather than individual attribute (‘good or bad’) data. The methods are based upon use of the normal distribution.

Sampling by variables is not as popular as sampling by attributes, since it is a more complex method. However, it can be useful when a particular production variable is important enough to warrant extra control.

Table 15.1 *Master table for normal inspection-single sampling (MIL-STD-105D, Table II-A).*

Sample size code letter	Sample size	Acceptable quality levels (normal inspection)																									
		0.010	0.015	0.025	0.040	0.065	0.10	0.15	0.25	0.40	0.65	1.0	1.5	2.5	4.0	6.5	10	15	25	40	65	100	125	250	400	650	1000
		Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re
A	2	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	0 1	↓	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	30 31
B	3														0 1	↑		1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	30 31	44 45
C	5													0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	30 31	44 45	↑
D	8												0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	30 31	44 45	↑	
E	13											0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	30 31	44 45	↑		
F	20										0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	↑	↑	↑			
G	32									0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	↑						
H	50								0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	↑							
I	80							0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	↑								
K	125						0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	↑									
L	200					0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	↑										
M	315				0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	↑											
N	500			0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	↑												
P	800		0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	↑													
Q	1250	0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	↑														
R	2000	↑		1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	↑															

↓ = Use first sampling plan below arrow. If sample size equals, or exceeds, lot or batch size, do 100 percent inspection.

↑ = Use first sampling plan above arrow.

Ac = Acceptance number.

Re = Rejection number.

Table 15.2 *LTPD sampling plans.*[a] *Minimum size of sample to be tested to assure, with 90 % confidence, that a lot having percentage defective equal to the specified LTPD will not be accepted (single sample).*

Max. percentage defective (LTPD) or λ	30	20	15	10	7	5	3	2	1.5	1	0.7	0.5	0.3
Acceptance number (c) ($r = c + 1$)	Minimum sample sizes (for device-hours required for life test, multiply by 1000)												
0	8	11	15	22	32	45	76	116	153	231	328	461	767
(0.64)	(0.46)	(0.34)	(0.23)	(0.16)	(0.11)	(0.07)	(0.04)	(0.03)	(0.02)	(0.02)	(0.01)	(0.007)	
1	13	18	25	38	55	77	129	195	258	390	555	778	1296
	(2.7)	(2.0)	(1.4)	(0.94)	(0.65)	(0.46)	(0.28)	(0.18)	(0.14)	(0.09)	(0.06)	(0.045)	(0.027)
2	18	25	34	52	75	105	176	266	354	533	759	1065	1773
	(4.5)	(3.4)	(2.24)	(1.6)	(1.1)	(0.78)	(0.47)	(0.31)	(0.23)	(0.15)	(0.11)	(0.080)	(0.045)
3	22	32	43	65	94	132	221	333	444	668	953	1337	2226
	(6.2)	(4.4)	(3.2)	(2.1)	(1.5)	(1.0)	(0.62)	(0.41)	(0.31)	(0.20)	(0.14)	(0.10)	(0.062)
4	27	38	52	78	113	158	265	398	531	798	1140	1599	2663
	(7.3)	(5.3)	(3.9)	(2.6)	(1.8)	(1.3)	(0.75)	(0.50)	(0.37)	(0.25)	(0.17)	(0.12)	(0.074)
5	31	45	60	91	131	184	308	462	617	927	1323	1855	3090
	(8.4)	(6.0)	(4.4)	(2.9)	(2.0)	(1.4)	(0.85)	(0.57)	(0.42)	(0.28)	(0.20)	(0.14)	(0.085)
6	35	51	68	104	149	209	349	528	700	1054	1503	2107	3509
	(9.4)	(6.6)	(4.9)	(3.2)	(2.2)	(1.6)	(0.94)	(0.62)	(0.47)	(0.31)	(0.22)	(0.155)	(0.093)
7	39	57	77	116	186	234	390	589	783	1178	1680	2355	3922
	(10.2)	(7.2)	(5.3)	(3.5)	(2.4)	(1.7)	(1.0)	(0.57)	(0.51)	(0.34)	(0.24)	(0.17)	(0.101)

[a] MIL-S-19500 and MIL-M-38510. Sample sizes are based on the Poisson exponential binomial limit. The minimum quality (approximate AQL) required to accept (on the average) 19 of 20 lots is shown in parentheses for information only.

The life test failure rate, λ, shall be defined as the LTPD per 1000 h.

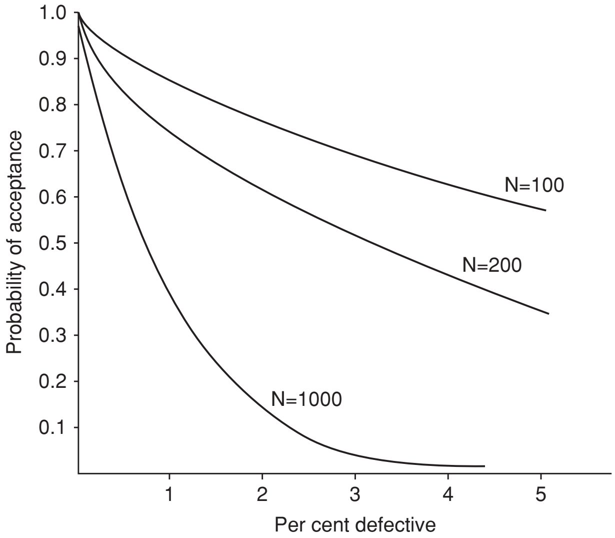


Figure 15.4 *Operating characteristic (OC) curves for single sampling plans (10 % sample, acceptance number = 0) (See Section 15.4.1.).*

15.4.3 General Comments on Sampling

Whilst standard sampling plans can provide some assurance that the proportion defective is below a specified figure, they do not provide the high assurance necessary for many modern products. For example, if an electronic assembly consists of 100 components, all with an AQL of 0.1 %, there is on average a probability of about 0.9 that an assembly will be free of defective components. If 10 000 assemblies are produced, about 1000 on average will be defective. The costs involved in diagnosis and repair or scrap during manufacture would obviously be very high. With such quantities typical of much modern manufacturing, higher assurance than can realistically be obtained from statistical sampling plans is obviously necessary. Also, standard QC sampling plans often do not provide assurance of long-term reliability.

Electronic component manufacturers sometimes quote quality levels in parts defective per million (ppm). Typical figures for integrated circuits are 5 to 30 ppm and lower for simpler components such as transistors, resistors and passive components.

All statistical sampling methods rely on the inspection or testing of samples which are drawn at random from the manufacturing batches, then using the mathematics of probability theory to make assertions about the quality of the batches. The sampling plans are based upon the idea of balancing the cost of test or inspection against minimizing the probability of the batch being accepted with an actual defective proportion higher than the AQL or LTPD.

However, optimizing the cost of test or inspection is not an appropriate objective. The logically correct objective in any test and inspection situation is to minimize the total cost of manufacture and support. When analysed from this viewpoint, the only correct decision is either to perform 100 % or zero test/inspection. There is no theoretical sample size between these extremes that will satisfy the criterion of total cost minimization. In addition, most modern manufacturing processes, particularly at the level of components, generate such small proportions defective (typically a few per million) that the standard statistical sampling methods such as AQL and LTPD cannot discriminate whether or not a batch is 'acceptable'.

The fundamental illogic of statistical acceptance sampling was first explained by Deming (1987). If the cost of test or inspection of one item is k_1 the cost of a later failure caused by not inspecting or testing is k_2, and the average proportion defective is p, then if p is less than k_1/k_2 the correct (lowest total cost) strategy is not to test any. If p is greater than k_1/k_2 the correct strategy is to test all. This explanation represents the simplest case, but the principle is applicable generally: there is no alternative theoretically optimum sample size to test or inspect. The logic holds for inspection or test at any stage, whether of components entering a factory or of assembled products leaving.

For example, if an item costs \$ 50 to test at the end of production, and the average cost of failure in service is \$ 1000 (warranty, repair, spares, reputation), then $k_1/k_2 = 0.05$. So long as we can be confident that the production processes can ensure that fewer than 5 % will have defects that will cause failures in service, then the lowest cost policy is not to test or inspect any.

The logic of 0 or 100 % test or inspection is correct in stable conditions, that is, p, k_1 and k_2 are known and are relatively constant. Of course this is often not the case, but so long as we know that p is either much larger or much smaller than k_1/k_2 we can still base our test and inspection decisions on this criterion. If we are not sure, and particularly if the expected value of p can approach k_1/k_2, we should test/inspect 100 %.

There are some situations in which sampling is appropriate. In any production operation where the value of p is lower than the breakeven point but is uncertain, or might vary so that it approaches or exceeds this, then by testing or inspecting samples we might be able to detect such deviations and take corrective action or switch to 100 % inspection/test. It is important to note, however, that there can be no calculated optimum statistical sampling plan, since we do not know whether p changes or by how much it might. The amount and frequency of sampling can be determined only by practical thinking in relation to the processes, costs and risks involved. For example, if the production line in the example above produces items that are on average only 0.01 % defective, at a rate of 1000/week, we might decide to inspect or test 10/week as an audit, because 10 items can be dealt with or fitted into a test chamber with minimum interruption to production and delivery.

Items that operate or are used only once, such as rivets or locking fasteners, airbag deployment systems and pressure bursting discs can be tested only on a sample basis, since 100 % testing of production items is obviously not feasible. The optimum sample plan is still not statistically calculable, however, since the proportions defective are usually much lower than can be detected by any sample, and it will nearly always be highly uncertain and variable.

For these reasons statistical sampling is very little used nowadays.

15.5 Improving the Process

When a production process has been started, and is under statistical control, it is likely still to produce an output with some variation, even if this is well within the allowable tolerance. Also, occasional special causes might lead to out-of-tolerance or otherwise defective items. It is important that steps are taken to improve variation and yield, even when these appear to be at satisfactory levels. Continuous improvement nearly always leads to reduced costs, higher productivity, and higher reliability. The concept of continuous improvement was first put forward by W. E. Deming, and taken up enthusiastically in Japan, where it is called *Kaizen*.

The idea of the quality loss function, due to Taguchi (Section 11.5), also provides economic justification for continuous process improvement.

Methods that are available to generate process improvement are described below, and in Defeo and Juran (2010), Feigenbaum (1991), Deming (1987), Imai (1997), and Hutchins (1985).

15.5.1 Simple Charts

A variety of simple charting techniques can be used to help to identify and solve process variability problems. The Pareto chart (Section 13.2) is often used as the starting point to identify the most important problems and

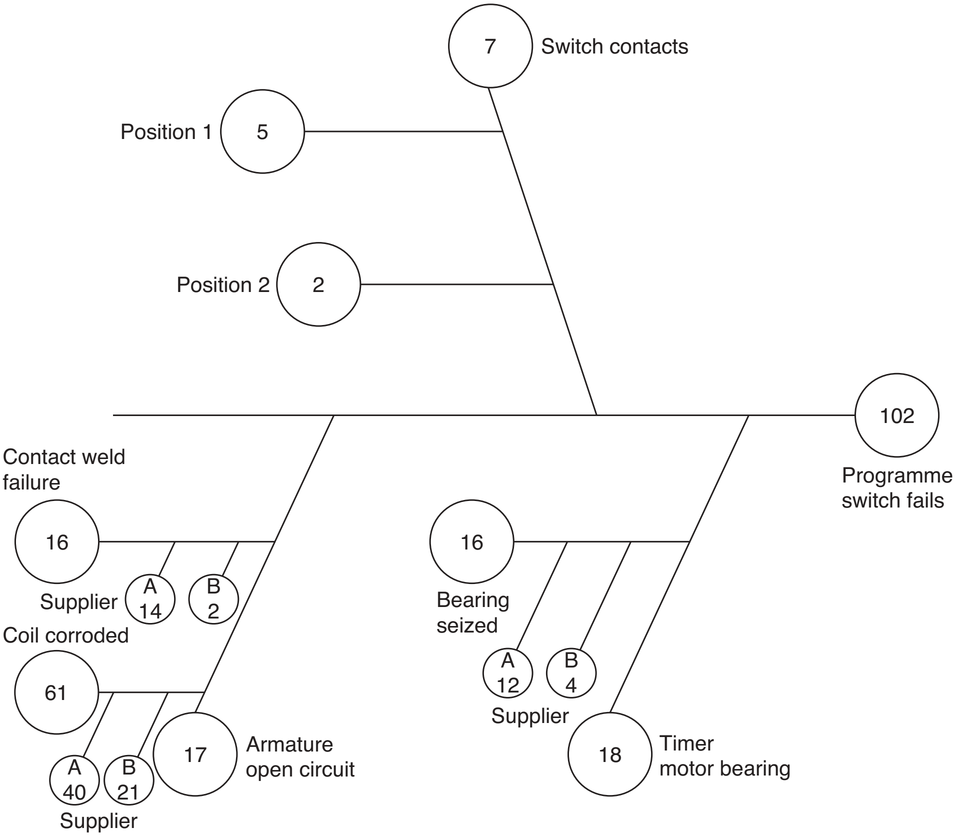


Figure 15.5 *Cause and effect diagram.*

the most likely causes. Where problems can be distributed over an area, for example defective solder joints on electronic assemblies, or defects in surface treatments, the *measles chart* is a useful aid. This consists simply of a diagram of the item, on which the locations of defects are marked as they are identified. Eventually a pattern of defect locations builds up, and this can indicate appropriate corrective action. For example, if solder defects cluster at one part of a PCB, this might be due to incorrect adjustment of the solder system.

The *cause-and-effect diagram* was invented by K. Ishikawa (Ishikawa, 1991) as an aid to structuring and recording problem-solving and process improvement efforts. The diagram is also called a *fishbone*, or Ishikawa, diagram. The main problem is indicated on a horizontal line, and possible causes are shown as branches, which in turn can have sub-causes, indicated by sub-branches, and so on. An example is shown in Figure 15.5; the numbers in the circles indicate the number of failures attributable to that cause.

15.5.2 Control Charts

When process control charts are in use, they should be monitored continuously for trends that might indicate special causes of variation, so that the causes can be eliminated. Trends can be a continuous run high or low on the chart, or any cyclic pattern. A continuous high or low trend indicates a need for process or measuring adjustment. A cyclic trend might be caused by temperature fluctuations, process drifts between settings, operator changeover, change of material, and so on. Therefore it is important to record supporting data on the SPC chart, such as time and date, to help with the identification of causes. When a process is being run on different machines the SPC charts of the separate processes should be compared, and all statistically significant differences investigated.

15.5.3 Multi-Vari Charts

A *multi-vari chart* is a graphical method for identifying the major causes of variation in a process. Multi-vari charts can be used for process development and for problem solving, and they can be very effective in reducing the number of variables to include in a statistical experiment.

Multi-vari charts show whether the major causes of variation are spatial, cyclic or temporal. The principle of operation is very simple: the parameter being monitored is measured in different positions (e.g. locations for measurement of a dimension, hardness, etc.), at different points in the production cycle (e.g. batch number from tool change), and at different times. The results are plotted as shown in Figure 15.6, which shows a machined dimension plotted against two measurement locations, for example diameters at each end of a shaft, plotted against batch number from set-up. It shows that batch-to-batch variation is the most significant cause, with a significant pattern of end-to-end variation (taper). This information would then be used to seek the reasons for the major cause, if necessary by running further experiments. Finally, statistical experiments can be run to refine the process further, particularly if interactions are statistically significant. The multi-vari method is described in Bhote (1998).

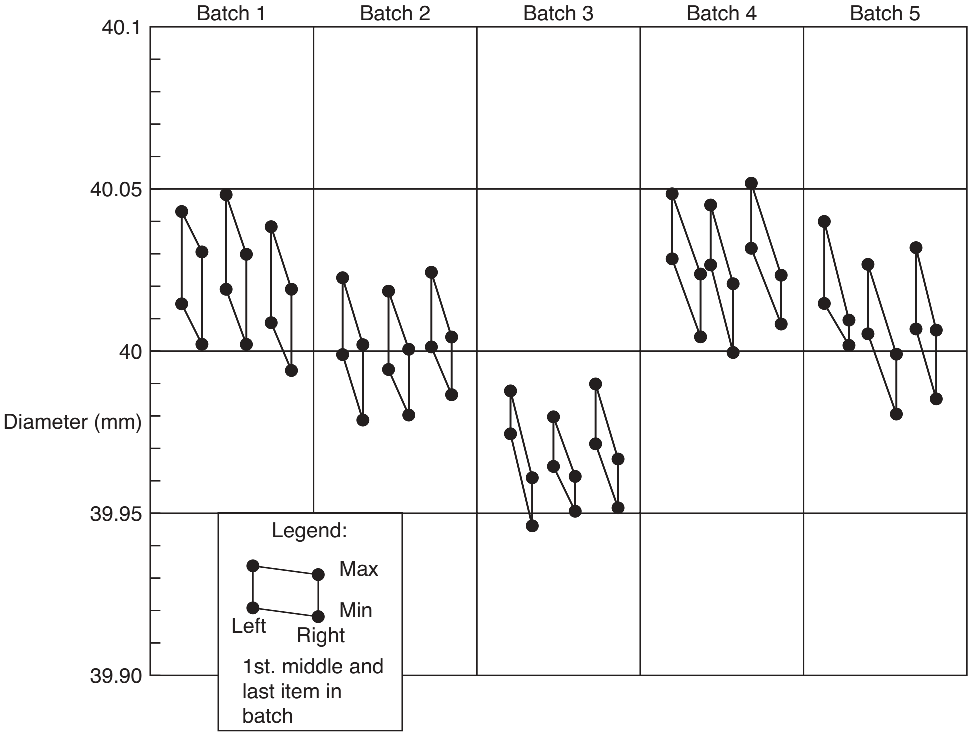


Figure 15.6 *Multi-vari chart.*

15.5.4 Statistical Methods

The methods for analysis of variation, described in Chapter 11, can be used just as effectively for variation reduction in production processes. They should be used for process improvement, in the same way as for product and process initial design. If a particular process has been the subject of such experiments during development, then the results can be used to guide studies for further improvement.

The methods described above can also be used to identify the major causes of variation, prior to setting up statistical experiments. In this way the number of variables to be investigated in the statistical experiment can be reduced, leading to cost savings.

15.5.5 'Zero Defects'

The '*zero defects*' (ZD) approach to QC was developed in the United States in the 1960s. ZD is based very much upon setting QC targets, publicizing results and exhortation by award presentations and poster campaigns. Successes were claimed for ZD but it has always been controversial. The motivational basis is evangelical and emotional, and the initial enthusiasm is hard to sustain. There are few managers who can set up and maintain a ZD programme as originally publicized, and consequently the approach is now seldom used.

15.5.6 Quality Circles

The quality circles movement started in Japan in the 1950s, and is now used worldwide. The idea is largely based on Drucker's management teaching, developed and taught for QC application by W. E. Deming and K. Ishikawa. It uses the methods of operator control, consistent with Drucker's teaching that the most effective management is that nearest to the action; this is combined with basic SPC and problem solving methods to identify and correct problems at their sources. The operator is often the person most likely to understand the problems of the process he or she operates, and how to solve them. However, the individual operator does not usually have the authority or the motivation to make changes. Also, he or she might not be able to influence problems elsewhere in the manufacturing system. The quality circles system gives workers this knowledge and influence, by organizing them into small work groups, trained to monitor quality performance, to analyse problems and to recommend solutions to management.

Quality circle teams manage themselves, select their leaders and members, and the problems to be addressed. They introduce the improvements if the methods are under their control. If not, they recommend the solutions to management, who must respond positively.

It is therefore a very different approach to that of ZD, since it introduces quality motivation as a normal working practice at the individual and team level, rather than by exhortation from above. Whilst management must be closely involved and supportive, it does not have to be continually active and visible in the way ZD requires.

Quality circles are taught to use analytical techniques to help to identify problems and generate solutions. These are called the *seven tools of quality*. The seven tools are:

1 Brainstorm, to identify and prioritize problems.
2 Data collection.
3 Data analysis methods, including measles charts, trend charts and regression analysis.
4 Pareto chart.
5 Histogram.
6 Cause-and-effect (or Ishikawa) diagram.
7 Statistical process control (SPC) chart.

For example, the team would be trained to interpret SPC charts to identify special causes of variation, and to use cause-and-effect diagrams. The cause-and-effect diagram is used by the team leader, usually on a flip chart, to put on view the problem being addressed and the ideas and solutions that are generated by the team, during the 'brainstorming' stage.

Quality circles must be organized with care and with the right training, and must have full support from senior and middle management. In particular, quality circles recommendations must be carefully assessed and actioned whenever they will be effective, or good reasons given for not following the recommendation.

The concept is really straightforward, enlightened management applied to the quality control problem, and in retrospect it might seem surprising that it has taken so long to be developed. It has proved to be highly successful in motivating people to produce better quality, and has been part of the foundation of the Japanese industrial revolution since the Second World War. The quality circles approach is closely associated with the concept of *kaizen*, the Japanese word meaning continuous improvement. The quality circles approach can be very effective when there is no formal quality control organization, for example in a small company.

The quality circles and kaizen approach to quality improvement is described fully in Hutchins (1985) and Imai (1997).

15.6 Quality Control in Electronics Production

15.6.1 Test Methods

Electronic equipment production is characterized by very distinct assembly stages, and specialist test equipment has been developed for each of these. Since electronic production is so ubiquitous, and since test methods can greatly affect quality costs and reliability, it is appropriate to consider test methods for electronics in this context. Test equipment for electronic production falls into the following main categories:

1 Manual test equipment, which includes basic instruments such as digital multimeters (DMMs), oscilloscopes, spectrum analysers, waveform generators, and logic analysers, as well as special instruments, such as radio frequency testers, distortion meters, high voltage testers, optical signal testers, and so on. Computer-based testing uses software that enables PCs to emulate test equipment.
2 Automatic Test Equipment.

Automatic test equipment (ATE) is used for testing manufactured circuits, and also for in-service fault-finding and for testing repaired units. The main types of ATE for assembly testing are:

15.6.1.1 Vision Systems

Vision systems refer generically to inspection systems that acquire an image and then analyse it. They do not actually test circuits, but they have become part of many production test sequences because of the great difficulty of human inspection of the large numbers of components, solder connections and tracks on modern circuits. *Automatic optical inspection* (AOI) machines are capable of scanning a manufactured circuit and identifying anomalies such as damaged, misplaced or missing components, faulty solder joints, solder spills across conductors, and so on. X-ray systems (AXI) are also used, to enable inspection of otherwise hidden aspects such as solder joints and internal component problems. Other technologies, such as infra-red and laser scanning, are also used.

15.6.1.2 In-Circuit Testers (ICT), Manufacturing Defects Analysers (MDA)

ICT tests the functions of components within circuits, on loaded circuit boards. It does not test the circuit function. The ICT machine accesses the components, one at a time, via a test fixture (sometimes referred to as a 'bed of nails' fixture), which consists of a large number of spring loaded contact pins, spaced to make contact with the appropriate test points for each component, for example the two ends of a resistor or the pin connections for an IC. ICT does not test circuit-level aspects such as tolerance mismatch, timing, interference, and so on. MDAs are similar but lower cost machines, with capabilities to detect only manufacturing-induced defects such as opens, shorts, and missing components: justification for their use instead of ICT is the fact that, in most modern electronics assembly, such defects are relatively more common than faulty components. *Flying probe testers* (also called *fixtureless testers)* perform the same functions, but access test points on the circuit using probes that are rapidly moved between points, driven by a programmed high-speed positioning system. The advantage over ICT/MDA is the fact that the probe programming is much less expensive and more adaptable to circuit changes than are expensive multipin ICT/MDA adaptors, which must be designed and built for each circuit to be tested. They can also gain access to test points that are difficult to access via a bed-of-nails adaptor.

15.6.1.3 Functional Testers (FT)

Functional testers access the circuit, at the circuit board or assembly level, via the input and output connectors or via a smaller number of spring-loaded probes. They perform a functional test, as driven by the test software. Functional testers usually include facilities for diagnosing the location of causes of incorrect function. There is a wide range, from low-cost bench-top ATE for use in development labs, in relatively low complexity/low production rate manufacture, in-service tests and in repair shops, to very large, high-speed high-capability systems. The modern trend is for production ATE to be specialized, and focused at defined technology areas, such as computing, signal processing, and so on. Some ATE for circuit testing during manufacture includes combined ICT and FT.

Electronics testing is a very fast-moving technology, driven by the advances in circuit performance, packaging and connection technology and production economics. O'Connor (2001) provides an introduction.

It is very important to design circuits to be testable by providing test access points and by including additional circuitry, or it will not be possible to test all functions or to diagnose the causes of certain failures. Design for testability can have a major impact on quality costs and on future maintenance costs. These aspects are covered in more detail in Chapters 9 and 16 and in Davis (1994).

The optimum test strategy for a particular electronic product will depend upon factors such as component quality, quality of the assembly processes, test equipment capability, costs of diagnosis and repair at each stage, and production throughput. Generally, the earlier a defect is detected, the lower will be the cost. A common rule of thumb states that the cost of detecting and correcting a defect increases by a factor of 10 at each test stage (component/PCB, ICT, functional test stages, in-service). Therefore the test strategy must be based on detecting and correcting defects as early as practicable in the production cycle, but defect probability, detection probability and test cost must also be considered.

The many variables involved in electronic equipment testing can be assessed by using computer models of the test options, under the range of assumed input conditions. Figure 15.7 shows a typical test flow arrangement.

Test results must be monitored continuously to ensure that the process is optimized in relation to total costs, including in-service costs. There will inevitably be variation over time. It is important to analyse the causes of failures at later test stages, to determine whether they could have been detected and corrected earlier.

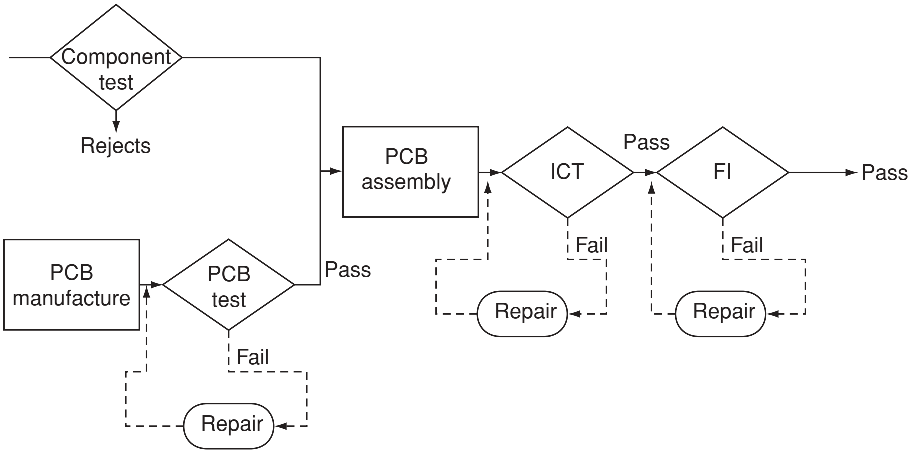


Figure 15.7 *Electronic equipment test strategy.*

Throughout, the causes of defects must be rapidly detected and eliminated, and this requires a very responsive data collection and analysis system (Section 15.8). Some ATE systems provide data logging and analysis, and data networking between test stations is also available.

Davis (1994) and O'Connor (2001) provide introductions to electronic test methods and economics.

15.6.2 Reliability of Connections

Complete opens and shorts will nearly always be detected during functional test, but open-circuit and intermittent failures can occur in use due to corrosion or fatigue of joints which pass initial visual and functional tests. The main points to watch to ensure solder joint reliability are:

1 *Process control*. Solder temperature, solder mix (when this is variable, e.g. in wave solder machines), soldering time, fluxes, PCB cleaning.
2 *Component mounting*. The solder joint should not be used to provide structural support, particularly in equipment subject to vibration or shock, or for components of relatively large mass, such as trimpots and some capacitors.
3 *Component preparation*. Component connections must be clean and wettable by the solder. Components which have been stored unpackaged for more than a few days or packaged for more than six months require special attention, as oxide formation may inhibit solderability, particularly on automatic soldering systems, where all joints are subject to the same time in the solder wave or oven. If necessary, such components should have their leads cleaned and retinned prior to assembly. Components should be subject to sampling tests for solderability, as near in time to the assembly stage as practicable.
4 *Solder joint inspection*. Inspectors performing visual inspection of solder joints are typically about 80 % effective in seeing joints which do not pass visual inspection standards. Also, it is possible to have unreliable joints which meet appearance criteria. If automatic testing for opens and shorts is used instead of 100 % visual inspection, it will not show up marginal joints which can fail later.

Surface mounted devices, as described in Chapter 9, present particular problems, since the solder connections are so much smaller and more closely spaced, and in many cases not visible. Manual soldering is not practicable, so automatic placement and soldering systems must be used. Visual inspection is difficult or impossible, and this has resulted in the development of semi-automatic and automated optical inspection systems, though these cannot be considered to be totally reliable. SMD solderability and soldering must be very carefully controlled to minimize the creation of defective joints.

15.7 Stress Screening

Stress screening, or *environmental stress screening* (ESS), is the application of stresses that will cause defective production items which pass other tests to fail on test, while not damaging or reducing the useful life of good ones. It is therefore a method for improving reliability and durability in service. Other terms are sometimes used for the process, the commonest being *burn-in*, particularly for electronic components and systems, for which the stresses usually applied are high temperature and electrical stress (current, voltage). The stress levels and durations to be applied must be determined in relation to the main failure-generating processes and the manufacturing processes that could cause weaknesses. Stress screening is normally a 100 % test, that is all manufactured items are tested. Stress screening is applied mainly to electronic components and assemblies, but it should be considered for non-electronic items also, for example precision mechanisms (temperature, vibration) and high pressure tests for pneumatics and hydraulics to check for leaks or other weaknesses.

ESS guidelines have been developed for electronic components and systems. The US Navy has published guidelines (NAVMAT P-9492), and the US DOD published MIL-STD-2164 (ESS Guidelines for Electronics Production), but these were inflexible and the stress levels specified were not severe (typically temperature cycling between 20 and 60 °C for 8 h, with random or fixed frequency vibration in the range 20–2000 Hz, and the equipment not powered or monitored). The US Institute for Environmental Sciences and Technology (IEST) developed more detailed guidelines in 1990 (Environmental Stress Screening of Electronic Hardware (ESSEH)), to cover both development and manufacturing tests. These recommended stress regimes similar to the military ESS guidelines. The details were based to a large extent on industry feedback of the perceived effectiveness of the methods that had been used up to the preparation of the guidelines, so they represented past experience, particularly of military equipment manufacture.

If ESS shows up very few defects, it is either insufficiently severe or the product being screened is already highly reliable. Obviously the latter situation is preferable, and all failures during screening should be analysed to determine if QC methods should have prevented them or discovered them earlier, particularly prior to assembly. At this stage, repairs are expensive, so eliminating the need for them by using high quality components and processes during production is a worthwhile objective.

Screening is expensive, so its application must be carefully monitored. Failure data should be analysed continuously to enable the process to be optimized in terms of operating conditions and duration. Times between failures should be logged to enable the process to be optimized.

Screening must be considered in the development of the production test strategy. The costs and effects in terms of reduced failure costs in service and the relationships with other test methods and QC methods must be assessed as part of the integrated quality assurance approach. Much depends upon the reliability requirement and the costs of failures in service. Also, screening should ensure that manufacturing quality problems are detected before shipment, so that they can be corrected before they affect much production output. With large-scale production, particularly of commercial and domestic equipment, screening of samples is sometimes applied for this purpose.

Jensen and Peterson (1983) describe methods and analytical techniques for screening of electronic components and assemblies.

15.7.1 Highly Accelerated Stress Screening

Highly accelerated stress screening (HASS) is an extension of the HALT principle, as described in Chapter 12, that makes use of very high combined stresses. No 'guidelines' are published to recommend particular stresses and durations. Instead, the stresses, cycles and durations are determined separately for each product (or group of similar products) by applying HALT during development. HALT shows up the product weak points, which are then strengthened as far as practicable so that failures will occur only well beyond the envelope of expected in-service combined stresses. The stresses that are then applied during HASS are higher than the operating limit, and extend into the lower tail of the distribution of the permanent failure limit. It is essential that the equipment under test is operated and monitored throughout.

The stresses applied in HASS, like those applied in HALT, are not designed to represent worst-case service conditions. They are designed to precipitate failures which would otherwise occur in service. This is why they can be developed only by applying HALT in development, and why they must be specific to each product design. Because the stress levels are so high, they cannot be applied safely to any design that has not been ruggedized through HALT.

Obviously the determination of this stress–time combination cannot be exact, because of the uncertainty inherent in the distribution of strength. However, by exploring the product's behaviour on test, we can determine appropriate stress levels and durations. The durations will be short, since the stress application rates are very high and there is usually no benefit to be gained by, for example, operating at constant high or low temperatures for longer than it takes for them to stabilize. Also, only a few stress cycles will be necessary, typically one to four.

When stresses above the operating limit are applied, it will not be possible to perform functional tests. Therefore the stresses must then be reduced to levels below the operating limit. The functional test will then show which items have been caused to fail, and which have survived. The screening process is therefore in two stages: the *precipitation screen* followed by the *detection screen*, as shown in Figure 15.8.

The total test time for HASS is much less than for conventional ESS: a few minutes vs. many hours. HASS is also much more effective at detecting defects. Therefore it is far more cost effective, by orders of magnitude. HASS is applied using the same facilities, particularly environmental chambers, as used for

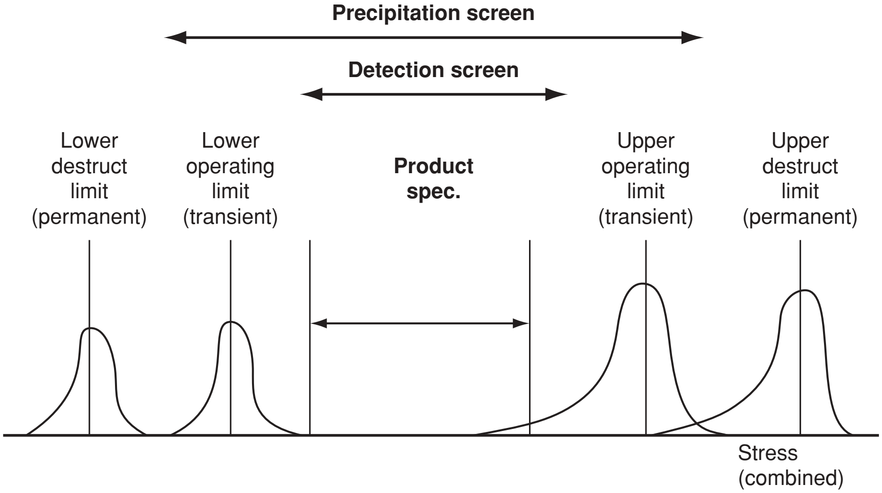

Precipitation screen
Detection screen
Lower destruct limit (permanent)
Lower operating limit (transient)
Product spec.
Upper operating limit (transient)
Upper destruct limit (permanent)
Stress (combined)

Figure 15.8 HASS.

HALT. Since the test times are so short, it is often convenient to use the same facilities during development and in production, leading to further savings in relation to test and monitoring equipment, interfaces, and so on. The HASS concept can be applied to any type of product or technology. It is by no means limited to electronic assemblies. If the design can be improved by HALT, as described in Chapter 12, then in principle manufacturing quality can be improved by HASS. The HASS approach to manufacturing stress screening is described fully in McLean (2009).

15.8 Production Failure Reporting Analysis and Corrective Action System (FRACAS)

Failure reporting and analysis is an important part of the QA function. The system must provide for:

1 Reporting of all production test and inspection failures with sufficient detail to enable investigation and corrective action to be taken.
2 Reporting the results of investigation and action.
3 Analysis of failure patterns and trends, and reporting on these.
4 Continuous improvement (*kaizen*) by removal of causes of failures.

The FRACAS principles (Section 12.6) are equally applicable to production failures.

The data system should be computerized for economy and accuracy. Modern ATE sometimes includes direct test data recording and inputting to a central system by networking. The data analysis system should provide Pareto analysis, probability plots and trend analyses for management reporting.

Production defect data reporting and analysis must be very quick to be effective. Trends should be analysed daily, or weekly at most, particularly for high rates of production, to enable timely corrective action to be taken. Production problems usually arise quickly, and the system must be able to detect a bad batch of components or an improperly adjusted process as soon as possible. Many problems become immediately apparent without the aid of a data analysis system, but a change of, say, 50 % in the proportion of a particular component in a system failing on test might not be noticed otherwise. The data analysis system is also necessary for indicating areas for priority action, using the Pareto principle of concentrating action on the few problem areas that contribute the most to quality costs. For this purpose longer term analysis, say monthly, is necessary.

Defective components should not be scrapped immediately, but should be labelled and stored for a period, say one to two months, so that they are available for more detailed investigation if necessary.

Production defect data should not be analysed in isolation by people whose task is primarily data management. The people involved (production, supervisors, QC engineers, test operators, etc.) must participate to ensure that the data are interpreted by those involved and that practical results are derived. The quality circles approach provides very effectively for this.

Production defect data are important for highlighting possible in-service reliability problems. Many in-service failure modes manifest themselves during production inspection and test. For example, if a component or process generates failures on the final functional test, and these are corrected before delivery, it is possible that the failure mechanism exists in products which pass test and are shipped. Metal surface protection and soldering processes present such risks, as can almost any component in electronic production. Therefore production defects should always be analysed to determine the likely effects on reliability and on external failure costs, as well as on internal production quality costs.

15.9 Conclusions

The modern approach to production quality control and improvement is based on the use of statistical methods and on organizing, motivating and training production people at all levels to work for continuously improving performance, of people and of processes. A very close link must exist between design and development of the product and of the production processes, and the criteria and methods to be used to control the processes. This integrated approach to management of the design and production processes is described in more detail in Chapter 17. The journals of the main professional societies for quality (see the Bibliography) also provide information on new developments.

Questions

1. A machined dimension on a component is specified as 12.50 mm ± 0.10 mm. A preliminary series of ten samples, each of five components, is taken from the process, with measurements of the dimension as follows:

Sample	Dimensions (mm)				
1	12.55	12.51	12.48	12.55	*12.46*
2	12.54	12.56	12.51	12.54	*12.47*
3	12.53	12.46	12.49	12.45	*12.50*
4	12.55	12.55	12.49	12.55	*12.47*
5	12.49	12.52	12.49	12.48	*12.48*
6	12.51	12.54	12.51	12.52	*12.45*
7	12.53	12.52	12.49	12.46	*12.50*
8	12.50	12.55	12.52	12.44	*12.46*
9	12.48	12.52	12.54	12.49	*12.50*
10	12.50	12.50	12.49	12.54	*12.54*

   a Use the averages and ranges for each sample to assess the capability of the process (calculated C_p and C_{pk}).
   b Use the relationship (taken from BS 5700) that, for samples of size 5, standard deviation = average range × 2.326.
   c Suggest any action that may be necessary.
2. Sketch the two charts used for statistical process control. What is the primary objective of the method?
3. Explain why statistical acceptance sampling is not an effective method for monitoring processes quality.
4. A particular component is used in large quantities in an assembly. It costs very little, but some protection against defective components is needed as they can be easily detected only after they have been built-in to an expensive assembly, which is then scrap if it contains a defective component. 100 % inspection of the components is not possible as testing at the component level would be destructive, but the idea of acceptance sampling is attractive.
   a How would you decide on the AQL to use?
   b If the AQL was 0.4 % defective and these components were supplied in batches of 2500, what sampling plan would you select from Table 15.1? (Batches of 2500 require sample size code K for general inspection, level II.)

c What would you do if you detected a single defective component in the sample?

d What would you do if this plan caused a batch to be rejected?

5. A manufacturing line produces car radios. The cost of the final test is $ 15, and the average proportion found to be not working is 0.007, though on some batches it is as high as 0.02. If the total cost of selling a non-working radio (replacement, administration, etc.) is $ 400, comment on the continued application of the test.
6. List the '7 tools of quality' as applied in the Quality Circles approach. Explain how they are applied by a Quality Circles team.
7. Briefly describe the three main approaches used for automatic inspection and test of modern electronic circuits. Sketch a typical inspection/test flow for a line producing circuit assemblies.
8. Describe the stresses that are typically applied to electronic hardware during environmental stress screening (ESS) in manufacturing.
9. a Define the main advantages and disadvantages of applying environmental stress screening (ESS) to electronic assemblies in production.

   b Discuss why it would be wrong to impose an ESS using MIL-HDBK-781 type environmental profiles.
10. How does highly accelerated stress screening (HASS) differ from conventional ESS? How is it related to HALT in development?
11. a Explain the difference between reducing failure rate by 'burn-in' and by reliability growth in service.

    b Assuming that you are responsible for the reliability of a complex electronic system, that is about to go into production, outline the way that you would set up 'burn-in' testing for purchased components, sub-assemblies and the completed product, in each case stating the purpose of the test and the criteria you would consider in deciding its duration.
12. What would be the Cp and $\mathrm{Cp_k}$ values for a '6-sigma process', which shifts $\pm 1.5\sigma$?
13. As mentioned at the beginning of Section 15.2, SPC is based on the assumption that the process follows the normal distribution, which sometimes is not the case. How would you analyse a process which can be modelled by a skewed distribution, for example lognormal?
14. Compare the Fault Tree Analysis and the Ishikawa diagram methods. What are the similarities and the differences of the two methods?

Bibliography

ANSI/ASQ Z1–4. Sampling Procedures and Tables for Inspection by Attributes.

Bergman B. and Klefsjö, B. (2003) *Quality: from Customer Needs to Customer Satisfaction*, 3rd edn, McGraw-Hill.

Bhote, K.R. (1998) *World Class Quality*, American Management Association.

British Standard BS 6001. *Sampling Procedures and Tables for Inspection by Attributes*.

Davis, B. (1994) *The Economics of Automatic Test Equipment*, 2nd edn, McGraw-Hill.

Defeo, J. and Juran, J. (2010) *Juran's Quality Handbook*, 6th. edn, McGraw-Hill.

Deming, W.E. (1987) *Out of the Crisis*, MIT University Press.

Feigenbaum, A.V. (1991) *Total Quality Control*, 3rd edn, McGraw-Hill.

Grant, E.L. and Leavenworth, R.S. (1996) *Statistical Quality Control*, 6th edn, McGraw-Hill.

Hutchins, D.C. (1985) *Quality Circles Handbook*. Pitman.

Imai, M. (1997) *Gemba Kaizen*. McGraw-Hill.

Ishikawa, K. (1991) *Guide to Quality Control*. Chapman and Hall.

Jensen, F., Peterson, N.E. (1983) *Burn-in: An Engineering Approach to the Design and Analysis of Burn-in Procedures*. Wiley.

McLean, H. (2009) *Halt, Hass, and Hasa Explained: Accelerated Reliability Techniques*, American Society for Quality (ASQ) Publishing.

Montgomery, D.C. (2008) *Introduction to Statistical Quality Control*, Wiley.

Oakland, J.S. and Followell, R.F. (2003) *Statistical Process Control, a Practical Guide*, 5th edn, Butterworth-Heinemann.

O'Connor, P.D.T. (2001) *Test Engineering*, Wiley.

Quality Assurance. *Journal of the Institute of Quality Assurance (UK).*

Quality Progress. *Journal of the American Society for Quality*.

Thomas, B. (1995) *The Human Dimension of Quality*, McGraw-Hill.

16

Maintainability, Maintenance and Availability

16.1 Introduction

Most engineered systems are maintained, that is they are repaired when they fail, and work is performed on them to keep them operating. The ease with which repairs and other maintenance work can be carried out determines a system's maintainability.

Maintained systems may be subject to *corrective* and *preventive* maintenance (CM and PM). Corrective maintenance includes all action to return a system from a failed to an operating or available state. The amount of corrective maintenance is therefore determined by reliability. Corrective maintenance action usually cannot be planned; we must repair failures when they occur, though sometimes repairs can be deferred.

Corrective maintenance can be quantified as the *mean time to repair* (MTTR). The time to repair, however, includes several activities, usually divided into three groups:

1 Preparation time: finding the person for the job, travel, obtaining tools and test equipment, and so on.
2 Active maintenance time: actually doing the job.
3 Delay time (logistics time): waiting for spares, and so on, once the job has been started.

Active maintenance time includes time for studying repair charts, and so on, before the actual repair is started, and time spent in verifying that the repair is satisfactory. It might also include time for post-repair documentation when this must be completed before the equipment can be made available, for example on aircraft. Corrective maintenance is also specified as a *mean active repair time* (MART) or *mean active corrective maintenance time* (MACMT), since it is only the active time (excluding documentation) that the designer can influence.

Preventive maintenance seeks to retain the system in an operational or available state by preventing failures from occurring. This can be by servicing, such as cleaning and lubrication, or by inspection to find and rectify incipient failures, for example by crack detection or calibration. Preventive maintenance affects reliability directly. It is planned and should be performed when we want it to be. Preventive maintenance is measured by the time taken to perform the specified maintenance tasks and their specified frequency.

Practical Reliability Engineering, Fifth Edition. Patrick D. T. O'Connor and Andre Kleyner.

Maintainability affects availability directly. The time taken to repair failures and to carry out routine preventive maintenance removes the system from the available state. There is thus a close relationship between reliability and maintainability, one affecting the other and both affecting availability and costs.

The maintainability of a system is clearly governed by the design. The design determines features such as accessibility, ease of test, diagnosis and repair and requirements for calibration, lubrication and other preventive maintenance actions.

This chapter describes how maintainability can be optimized by design, and how it can be predicted and measured. It also shows how plans for preventive maintenance can be optimized in relation to reliability, to minimize downtime and costs.

16.2 Availability Measures

In order to analyse a system's availability it needs to be measured. Depending on the available data and the objectives of the analysis availability can be expressed in several different ways.

16.2.1 Inherent Availability

Inherent availability is the steady state availability which considers only the corrective maintenance (CM) (covered in Section 6.7). Assuming that CM actions occur at a constant rate, it can be estimated as:

$$A_I = \frac{MTBF}{MTBF + MTTR} \tag{16.1}$$

MTBF is the mean time between failure and *MTTR* is the mean time to repair (Chapter 6) which is the same as mean corrective maintenance time.

16.2.2 Achieved Availability

Achieved availability is very similar to inherent availability with the exception that PM downtimes are also included. Specifically, it is the steady state availability in an ideal support environment (i.e. readily available tools, spares, personnel, etc.) The achieved availability is sometimes referred to as the availability seen by the maintenance department (does not include logistic delays, supply delays or administrative delays).

Achieved availability can be computed by looking at the *mean time between maintenance actions MTBMA* (both preventive and corrective) and the *mean maintenance time*, *MMT*:

$$A_A = \frac{MTBMA}{MTBMA + MMT} \tag{16.2}$$

Assuming constant failure rate *MTBMA* can be calculated as:

$$MTBMA = \frac{1}{\lambda + f_{PM}} \tag{16.3}$$

where: $\lambda =$ the failure rate (assuming all failures are repaired).
$f_{PM} =$ the frequency of preventive maintenance, the reciprocal of PM Cycle.

The mean maintenance time *MMT* can be further decomposed into the effects of preventive and corrective maintenance as:

$$MMT = \frac{\lambda\, MTTR + f_{PM} MPMT}{\lambda + f_{PM}} \tag{16.4}$$

Where *MTTR* is the mean CM time and *MPMT* is the mean PM time.

16.2.3 Operational Availability

Operational availability is a measure of the 'real' average availability over a period of time in an actual operational environment. It includes all experienced sources of downtime, such as administrative downtime, logistic downtime, and so on

$$A_{Op} = \frac{MTBMA}{MTBMA + MDT} \tag{16.5}$$

where: $MDT =$ Mean maintenance downtime.
$MDT = MMT +$ (logistics delay time) $+$ (administrative delay time).

For more on availability measures see Blanchard and Fabrycky (2011).

Example 16.1

Calculate the achieved availability for a system with the demonstrated failure rate of 1 failure per 1000 hours of operation. Preventive maintenance is scheduled every 500 hours of operation and on average takes 6 hours. To repair a failed system takes on average 16 hours.

The frequency of preventive maintenance is $f_{PM} = 1/500\,\text{h} = 0.002\,\text{h}^{-1}$ and the failure rate is $\lambda = 1/1000\,\text{h} = 0.001\,\text{h}^{-1}$. Next step is to calculate *MTBMA*, *MMT*, and A_A based on (16.2)–(16.4).

$$\begin{aligned} MTBMA &= \frac{1}{0.001 + 0.002} = 333.33\ h \\ MMT &= \frac{0.001h^{-1} \times 16h\ + 0.002h^{-1} \times 6h}{0.001h^{-1} + 0.002h^{-1}} = 9.33\,h \\ A_A &= \frac{333.33}{333.33 + 9.33} = 0.973 \end{aligned}$$

16.3 Maintenance Time Distributions

Maintenance times tend to be lognormally distributed (see Figure 16.1). This has been shown by analysis of data. It also fits our experience and intuition that, for a task or group of tasks, there are occasions when the work is performed rather quickly, but it is relatively unlikely that the work will be done in much less time than usual, whereas it is relatively more likely that problems will occur which will cause the work to take much longer than usual. This skews time-to-repair distributions to the right.

In addition to the job-to-job variability, leading typically to a lognormal distribution of repair times, there is also variability due to learning. Depending upon how data are collected, this variability might be included

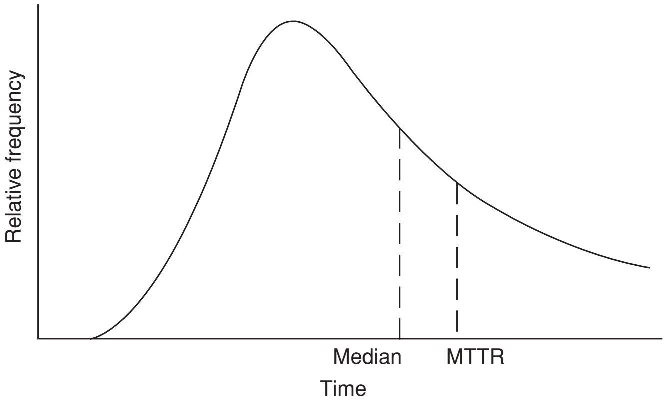


Figure 16.1 *The lognormal distribution of maintenance times.*

in the job-to-job variability, for example if technicians of different experience are being used simultaneously. However, both the mean time and the variance should reduce with experience and training.

The properties of the lognormal distribution are described in Chapter 2.

16.4 Preventive Maintenance Strategy

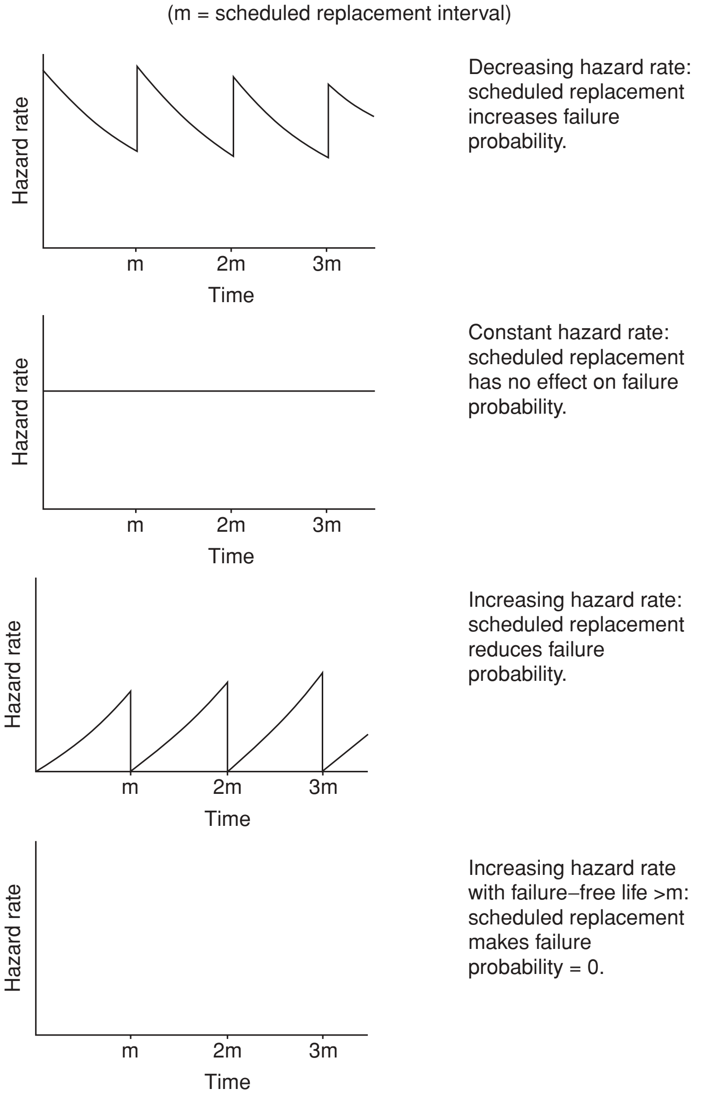
The effectiveness and economy of preventive maintenance can be maximized by taking account of the time-to-failure distributions of the maintained parts and of the failure rate trend of the system.

In general, if a part has a decreasing hazard rate, any replacement will increase the probability of failure. If the hazard rate is constant, replacement will make no difference to the failure probability. If a part has an increasing hazard rate, then scheduled replacement at any time will in theory improve reliability of the system. However, if the part has a failure-free life (Weibull $\gamma < 0$), then replacement before this time will ensure that failures do not occur. These situations are shown in Figure 16.2.

These are theoretical considerations. They assume that the replacement action does not introduce any other defects and that the time-to-failure distributions are exactly defined. As explained in Chapters 2 and 6, these assumptions must not be made without question. However, it is obviously of prime importance to take account of the time-to-failure distributions of parts in planning a preventive maintenance strategy.

In addition to the effect of replacement on reliability as theoretically determined by considering the time-to-failure distributions of the replaced parts, we must also take account of the effects of the maintenance action on reliability. For example, data might show that a high pressure hydraulic hose has an increasing hazard rate after a failure-free life, in terms of hose leaks. A sensible maintenance policy might therefore be to replace the hose after, say, 80 % of the failure-free life. However, if the replacement action increases the probability of hydraulic leaks from the hose end connectors, it might be more economical to replace hoses on failure.

The effects of failures, both in terms of effects on the system and of costs of downtime and repair, must also be considered. In the hydraulic hose example, for instance, a hose leak might be serious if severe loss of fluid results, but end connector leaks might generally be only slight, not affecting performance or safety. A good example of replacement strategy being optimized from the cost point of view is the replacement of incandescent and fluorescent light units. It is cheaper to replace all units at a scheduled time before an



Decreasing hazard rate: scheduled replacement increases failure probability.

Constant hazard rate: scheduled replacement has no effect on failure probability.

Increasing hazard rate: scheduled replacement reduces failure probability.

Increasing hazard rate with failure–free life >m: scheduled replacement makes failure probability = 0.

Figure 16.2 *Theoretical reliability and scheduled replacement relationships.*

expected proportion will have failed, rather than to replace each unit on failure, in large installations such as offices and street lights. However, at home we would replace only on failure.

In order to optimize preventive replacement, it is therefore necessary to know the following for each part:

1 The time-to-failure distribution parameters for the main failure modes.
2 The effects of all failure modes.
3 The cost of failure.
4 The cost of scheduled replacement.
5 The likely effect of maintenance on reliability.

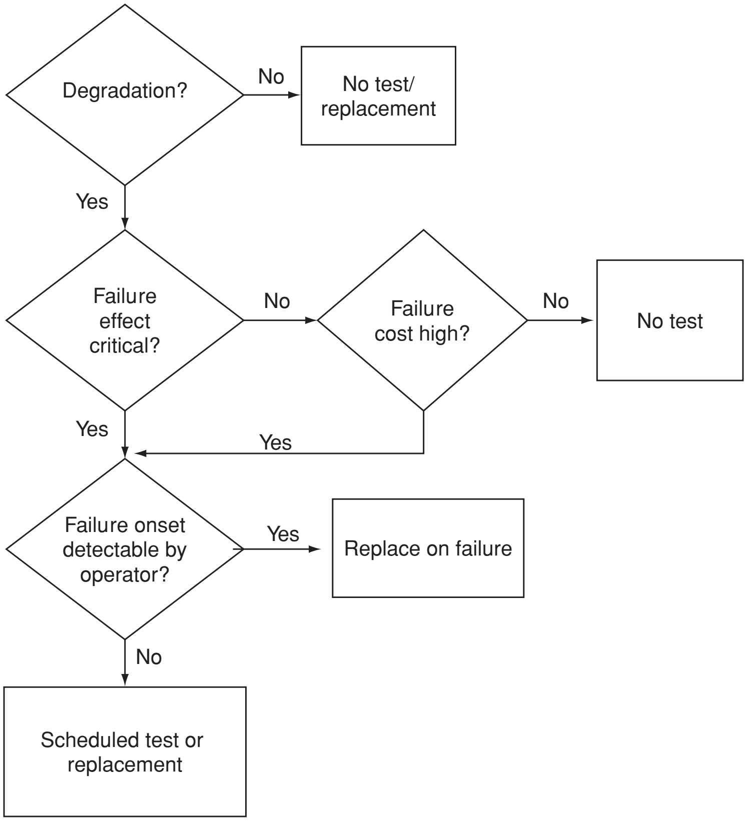


Figure 16.3 *RCM logic.*

We have considered so far parts which do not give any warning of the onset of failure. If incipient failure can be detected, for example by inspection, non-destructive testing, and so on, we must also consider:

6 The rate at which defects propagate to cause failure.
7 The cost of inspection or test.

Note that, from 2, an FMECA is therefore an essential input to maintenance planning.

This systematic approach to maintenance planning, taking account of reliability aspects, is called *reliability centred maintenance* (RCM). Figure 16.3 shows the basic logic of the method. RCM is widely applied, for example on aircraft, factory systems, and so on. It is described in Moubray (1999), Bloom (2005) and other books. RCM software is also available.

Example 16.2

A flexible cable on a robot assembly line has a time-to-failure distribution which is Weibull, with $\beta = 1.7$, $\eta = 300\,\text{h}$ and $\gamma = 150\,\text{h}$. If failure occurs whilst in use the cost of stopping the line and replacing the cable is \$ 5000. The cost of replacement during scheduled maintenance is \$ 500. If the line runs for 5000 hours a year and scheduled maintenance takes place every week (100 hours), what would be the annual expected cost of replacement at one-weekly or two-weekly intervals?

With no scheduled replacement the probability of a failure occurring in *t* hours will be

$$1 - \exp\left[-\left(\frac{t-150}{300}\right)^{1.7}\right] \tag{16.6}$$

With scheduled replacement after *m* hours, the scheduled maintenance cost in 5000 h will be

$$\frac{5000}{m} \times 500 = \frac{2.5 \times 10^6}{m}$$

and the expected failure cost in each scheduled replacement interval will be (assuming not more than one failure in any replacement interval):

$$5000\left\{1 - \exp\left[1 - \left(\frac{m-150}{300}\right)^{1.7}\right]\right\}$$

Then the total cost per year

$$C = \frac{2.5 \times 10^6}{m} + \frac{5000 \times 5000}{m}\left\{1 - \exp\left[-\left(\frac{m-150}{300}\right)^{1.7}\right]\right\}$$

Results are as follows:

m	No. of scheduled replacements	Expected failures	C
100	50	0	$ 25 000
200	25	1.2	$ 18 304
400	12	6.5	$ 38 735

Therefore the optimum policy might be to replace the cables at alternate scheduled maintenance intervals, taking a slight risk of failure. (Note that the example assumes that not more than one failure occurs in any scheduled maintenance interval. If *m* is only a little more than γ this is a reasonable assumption.)

A more complete analysis could be performed using Monte Carlo simulation (Chapter 4). We could then take account of more detailed maintenance policies; e.g. it might be decided that if a cable had been replaced due to failure shortly before a scheduled maintenance period, it would not be replaced at that time.

16.4.1 Practical Implications

The time-to-failure patterns of components in a system therefore largely dictate the optimum maintenance policy. Generally, since most electronic components do not wear out, scheduled tests and replacements do not improve reliability. Indeed they are more likely to induce failures (real or reported). Electronic equipment should only be subjected to periodic test or calibration when drifts in parameters or other failures can cause the equipment to operate outside specification without the user being aware. Built-in test and auto-calibration can reduce or eliminate the need for periodic test.

Mechanical equipment subject to wear, corrosion, fatigue, and so on, should be considered for preventive maintenance.

16.5 FMECA and FTA in Maintenance Planning

The FMECA is an important prerequisite to effective maintenance planning and maintainability analysis. As shown earlier, the effects of failure modes (costs, safety implications, detectability) must be considered in determining scheduled maintenance requirements. The FMECA is also a very useful input for preparation of diagnostic procedures and checklists, since the likely causes of failure symptoms can be traced back using the FMECA results. When a fault tree analysis (FTA) has been performed it can also be used for this purpose.

16.6 Maintenance Schedules

When any scheduled maintenance activity is determined to be necessary, we must also determine the most suitable intervals between its performance. Maintenance schedules should be based upon the most appropriate time or other base. These can include:

- Road and rail vehicles: distance travelled.
- Aircraft: hours flown, numbers of takeoff/landing cycles.
- Electronic equipment: hours run, numbers of on/off cycles.
- Fixed systems (factory equipment, communications networks, rail infrastructure, etc.): calendar time.

The most appropriate base is the one that best accounts for the equipment's utilization in terms of the causes of degradation (wear, fatigue, parameter change, etc.), and is measured. For example, we measure the distance travelled by our cars, and most degradation is related to this. On the other hand, there is no point in setting a calibration schedule based upon running time for a measuring instrument unless an automatic or manual record of its utilization were maintained, so these are usually calendar-based.

16.7 Technology Aspects

16.7.1 Mechanical

Monitoring methods are used to provide periodic or continuous indications of the condition of mechanical components and systems. These include:

- Non-destructive test (NDT) for detection of fatigue cracks.
- Temperature and vibration monitors on bearings, gears, engines, and so on.
- Oil analysis, to detect signs of wear or breakup in lubrication and hydraulic systems.

16.7.2 Electronic and Electrical

Most electronic components and assemblies generally do not degrade in service, so long as they are protected from environments such as corrosion. Most electronic components and connections do not suffer from wear or fatigue, except as discussed in Chapter 9, so there is very seldom a pronounced 'wearout' phase in which failures become more likely or frequent. Therefore, apart from calibration for items like measuring instruments, scheduled tests are seldom appropriate.

16.7.3 'No Fault Found'

A large proportion of the reported failures of many electronic systems are not confirmed on later test. These occurrences are called *no fault found* (NFF) or *re-test OK* (RTOK) faults. Other terms used include No Trouble Found (NTF) and Customer Complaint Not Verified (CCNV). There are several causes of these, including:

- Intermittent failures, such as components that fail under certain conditions (temperature, etc.), intermittent opens on conductor tracks or solder connections, and so on.
- Tolerance effects, which can cause a unit to operate correctly in one system or environment but not in another.
- Connector troubles. The failure seems to be cleared by replacing a unit, when in fact it was caused by the connector, that is disturbed to replace it.
- Built-in test (BIT) systems which falsely indicate failures that have not actually occurred (see below).
- Failures that have not been correctly diagnosed and repaired, so that the symptoms recur.
- Inconsistent test criteria between the in-service test and the test applied during diagnosis elsewhere such as the repair depot.
- Human error or inexperience.
- In some systems, the diagnosis of which item (card, box) has failed might be ambiguous, so more than one is replaced even though only one has failed. Sometimes it is quicker and easier for the technician to replace multiple items, rather than trying to find out which has failed. In these cases multiple units are sent for diagnosis and repair, resulting in a proportion being classified NFF. (Returning multiple units can sometimes be justified economically. e.g., it might be appropriate to spend the minimum time diagnosing the cause of a problem on a system such as an aircraft or oil rig, in order to return the system to operation as soon as possible.).

The proportion of reported failures that are caused in these ways can be very high, often over 50 % and sometimes up to 80 %. This can generate high costs, in relation to warranty, spares, support, test facilities, and so on. NFF rates can be minimized by effective management of the design in relation to in-service test, and of the diagnosis and repair operations. Stress screening of repaired items can also reduce the proportion of failures that are not correctly diagnosed and repaired.

16.7.4 Software

As discussed in Chapter 10 software does not fail in the ways that hardware can, so there is no 'maintenance'. If it is found to be necessary to change a program for any reason (change in system requirements, correction of a software error), this is really redesign of the program, not repair. So long as the change is made to all copies in use, they will all work identically, and will continue to do so.

16.7.5 Built-in Test (BIT)

Complex electronic systems such as laboratory instruments, avionics, communications networks and process control systems now frequently include built-in test (BIT) facilities. BIT consists of additional hardware and software which is used for carrying out functional test on the system. BIT might be designed to be activated by the operator, or it might monitor the system continuously or at set intervals.

BIT can be very effective in increasing system availability and user confidence in the system. However, BIT inevitably adds complexity and cost and can therefore increase the probability of failure. Additional sensors might be needed as well as BIT circuitry and displays. In microprocessor-controlled systems BIT can be largely implemented in software.

BIT can also adversely affect apparent reliability by falsely indicating that the system is in a failed condition. This can be caused by failures within the BIT, such as failures of sensors, connections, or other components. BIT should therefore be kept simple, and limited to monitoring of essential functions which cannot otherwise be easily monitored.

It is important to optimize the design of BIT in relation to reliability, availability and cost. Sometimes BIT performance is specified (e.g. ‘90 % of failures must be detected and correctly diagnosed by BIT’). An FMECA can be useful in checking designs against BIT requirements since BIT detection can be assessed against all the important failure modes identified.

16.8 Calibration

Calibration is the regular check or test of equipment used for measuring physical parameters, by making comparisons against standard sources. Calibration is applied to basic measurement tools such as micrometers, gauges, weights and torque wrenches, and to transducers and instruments that measure parameters such as flow rate, electrical potential, current and resistance, frequency, and so on. Therefore calibration may involve simple comparisons (weight, timekeeping, etc.) or more complex testing (radio frequency, engine torque, etc.).

Whether an item needs to be calibrated or not depends primarily upon its application, and also upon whether or not inaccuracy would be apparent during normal use. Any instrument used in manufacturing should be calibrated, to ensure that correct measurements are being made. Calibration is often a legal requirement, for example of measuring systems in food or pharmaceutical production and packaging, retailing, safety–critical processes, and so on.

Calibration requirements and systems are described in Morris (1997).

16.9 Maintainability Prediction

Maintainability prediction is the estimation of the maintenance workload which will be imposed by scheduled and unscheduled maintenance. A standard method used for this work is US MIL-HDBK-472 which contains four methods for predicting the mean time to repair (MTTR) of a system. Method II is the most frequently used. This is based simply on summing the products of the expected repair times of the individual failure modes, t_r and dividing by the sum of the individual failure rates, that is:

$$\mathrm{MTTR} = \frac{\Sigma(\lambda t_\mathrm{r})}{\Sigma\lambda} \tag{16.7}$$

The same approach is used for predicting the mean preventive maintenance time, with λ replaced by the frequency of occurrence of the preventive maintenance action.

MIL-HDBK-472 describes the methods to be used for predicting individual task times, based upon design considerations such as accessibility, skill levels required, and so on. It also describes the procedures for calculating and documenting the analysis, and for selection of maintenance tasks when a sampling basis is to be used (method III), rather than by considering all maintenance activities, which is impracticable on complex systems.

16.10 Maintainability Demonstration

A standard approach to maintainability demonstration is MIL-HDBK-470. The technique is the same as for maintainability prediction using method III of MIL-HDBK-472, except that the individual task times are measured rather than estimated from the design. Selection of task times to be demonstrated might be by agreement or by random selection from a list of maintenance activities. For more on maintainability demonstration see Blanchard and Fabrycky (2011).

16.11 Design for Maintainability

It is obviously important that maintained systems are designed so that maintenance tasks are easily performed, and that the skill levels required for diagnosis, repair and scheduled maintenance are not too high, considering the experience and training of likely maintenance personnel and users. Features such as ease of access and handling, the use of standard tools and equipment rather than specials, and the elimination of the need for delicate adjustment or calibration are desirable in maintained systems. As far as is practicable, the need for scheduled maintenance should be eliminated. Whilst the designer has no control over the performance of maintenance people, he or she can directly affect the inherent maintainability of a system.

Design rules and checklists should include guidance, based on experience of the relevant systems, to aid design for maintainability and to guide design review teams.

Design for maintainability is closely related to design for ease of production. If a product is easy to assemble and test maintenance will usually be easier. Design for testability of electronic circuits is particularly important in this respect, since circuit testability can greatly affect the ease and accuracy of failure diagnosis, and thus maintenance and logistics costs. Design of electronic equipment for testability is covered in more detail in Chapter 9.

Interchangeability is another important aspect of design for ease of maintenance of repairable systems. Replaceable components and assemblies must be designed so that no adjustment or recalibration is necessary after replacement. Interface tolerances must be specified to ensure that replacement units are interchangeable.

16.12 Integrated Logistic Support

Integrated logistic support (ILS) is a concept developed by the military, in which all aspects of design and of support and maintenance planning are brought together, to ensure that the design and the support system are optimized. Operational effectiveness, availability, and total costs of deployment and support are all considered. The approach is described in US-MIL-HDBK 1388, and in Jones (2006).

ILS, and the associated logistic support analysis (LSA), require inputs of reliability and maintainability data and forecasts, as well as data on costs, weights, special tools and test equipment, training requirements, and so on. MIL-HDBK-1388 requires that all analyses are computerized, and lays down standard input and output formats. Several commercial computer programs have been developed for the tasks.

ILS/LSA outputs are obviously very sensitive to the accuracy of the inputs. In particular, reliability forecasts can be highly uncertain, as explained in Chapter 6. Therefore such analyses, and decisions based upon them, should take full account of these uncertainties.

Questions

1. Define and explain the following terms: (a) availability; (b) maintainability; (c) condition monitoring.
2. a Explain circumstances in which it is possible to improve maintainability to counteract poor reliability, and also identify circumstances where this approach is unrewarding.

   b The expression for steady-state availability of a single element is:

$$A(t) = \frac{\mu}{\lambda + \mu}$$

   Explain what is meant by a steady state. State the meanings of the symbols μ and λ, and the assumptions inherent in the expression.
3. Describe the concept of reliability centred maintenance (RCM), including the key inputs to the analysis.
4. Recommend a planned replacement policy for the pumps in question 4 in Chapter 3.
5. Detail the time-related activities you may consider when analysing a maintenance-related task. (Hint: break the time down into active (or uptime) and inactive (or downtime)).
6. The median (50th percentile) active time to restore/repair a system after failure, using specified procedures and resources, is not to be more than 4.5 h. The maximum 15 % active restore/repair time should not be more than 13.5 h (i.e. mean repair time = 5.7 h). Criticize the statement given and deduce what would be a realistic set of consistent numbers. You should use the following equations related to the log-normal distribution:

$$\sigma = [2\,(\ln t_{\mathrm{MART}} - \ln t_m]^{1/2} \text{ and } \sigma Z_\alpha = \ln t_\alpha - \ln t_{\mathrm{m}},$$

   Where t_{MART} is the mean time to repair, t_{m} the median time to repair, and t_α the 'maximum' time to repair, evaluated at the $(1-\alpha)$ percentile point on the distribution. Z_α is the standardized normal deviate found in Appendix 1, and σ is the standard deviation of the distribution.
7. Given the removal and replacement time data in the table, calculate t_{MART} (mean time to repair)

Part identity	Quantity	Failure rate × 10^{-4} (h)	Total M task time (h)
Bolts	3	0.46	0.20
Earth strap	1	0.12	0.10
Power lead	1	0.36	0.36
Signal lead	1	1.16	0.10
Cover plate	1	1.05	0.26
Brushes	2	23.6	0.35

8. Calculate the operational availability for a power generator with 1 scheduled maintenance per 1000 hours of operation and failure rate of 0.6 failures per 1000 hours. Preventive maintenance takes on average 8 hours, however if the engine fails, it typically takes 26 hours to repair it. Logistics and administrative delays typically add extra 12 hours to the maintenance time.
9. Solve problem 8 with maintenance times distributed lognormally. For corrective maintenance times $\mu = 3.3$ and $\sigma = 0.6$ and for preventive maintenance $\mu = 2.2$ and $\sigma = 0.4$.
10. How will the routine replacement of parts, which have a constant hazard rate, affect their field failure rate? Justify your answer.

11. A wave soldering machine failures follow 3-parameter Weibull distribution with $\beta = 1.6$, $\eta = 800$ hours and $\gamma = 500$ hours. The cost of one PM is \$ 6000 and the cost of an unexpected failure is \$ 16 000. The machine operates 24 hours/day and PMs are done every three months. Determine the yearly cost for this maintenance schedule.

Bibliography

Blanchard, B. and Fabrycky, W. (2011) *Systems Engineering and Analysis*, 5th edn, Prentice Hall.

Bloom, N. (2005) *Reliability Centered Maintenance (RCM): Implementation Made Simple*, McGraw-Hill.

Dhillon, B. (1999) *Engineering Maintainability; How to Design for Reliability and Easy Maintenance*, Gulf Publishing.

Jardine, A. and Tsang, A. (2005) *Maintenance, Replacement and Reliability*. Dekker.

Jones, J. (2006) *Integrated Logistics Support Handbook*, McGraw-Hill Logistics Series.

Morris, A. (1997) *Measurement and Calibration Requirements for Quality Assurance to ISO9000*, Wiley.

Moubray, J. (1999) *Reliability Centred Maintenance*, 2nd edn, Butterworth-Heinemann.

Smith, D. (2005) *Reliability, Maintainability, and Risk*, 7th edn, Elsevier.

UK Defence Standard 00–40. *The Management of Reliability and Maintainability*. HMSO.

US MIL-HDBK-1388. *Integrated Logistic Support*. Available from the National Technical Information Service, Springfield, Virginia.

US MIL-HDBK-470. *Maintainability Program Requirements*. Available from the National Technical Information Service, Springfield, Virginia.

US MIL-HDBK-472. *Maintainability Prediction*. Available from the National Technical Information Service, Springfield, Virginia.

17
Reliability Management

17.1 Corporate Policy for Reliability

A really effective reliability function can exist only in an organization where the achievement of high reliability is recognized as part of the corporate strategy and is given top management attention. If these conditions are not fulfilled, and if it receives only lip service, reliability effort will be cut back whenever cost or time pressures arise. Reliability staff will suffer low morale and will not be accepted as part of project teams. Therefore, quality and reliability awareness and direction must start at the top and must permeate all functions and levels where reliability can be affected.

Several factors of modern industrial business make such high level awareness essential. The high costs of repairs under warranty, and of those borne by the user, even for relatively simple items such as domestic electronic and electrical equipment, make reliability a high value property. Other less easily quantifiable effects of reliability, such as customer goodwill and product reputation, and the relative reliability of competing products, are also important in determining market penetration and retention.

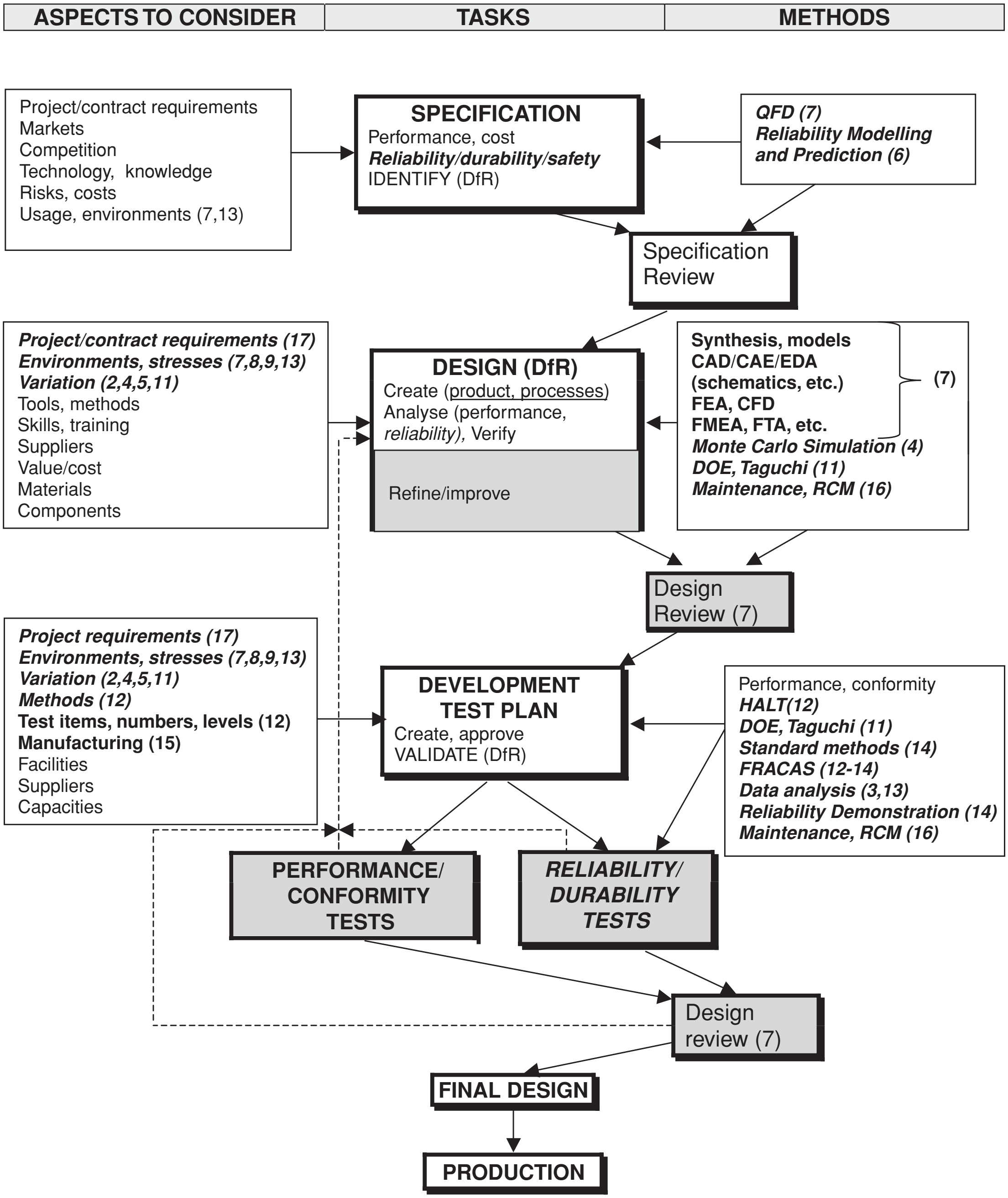
17.2 Integrated Reliability Programmes

The reliability effort should always be treated as an integral part of the product development and not as a parallel activity unresponsive to the rest of the development programme. This is the major justification for placing responsibility for reliability with the project manager. Whilst specialist reliability services and support can be provided from a central department in a matrix management structure, the responsibility for reliability achievement must not be taken away from the project manager, who is the only person who can ensure that the right balance is struck in allocating resources and time between the various competing aspects of product development.

The elements of a comprehensive and integrated reliability programme are shown, related to the overall development, production and in-service programme, in Figure 17.1 and Figure 17.2. These show the continuous feedback of information, so that design iteration can be most effective. Most of the design for reliability (DfR) tools and activities (Chapter 7) feature in the Figure 17.1 flow.

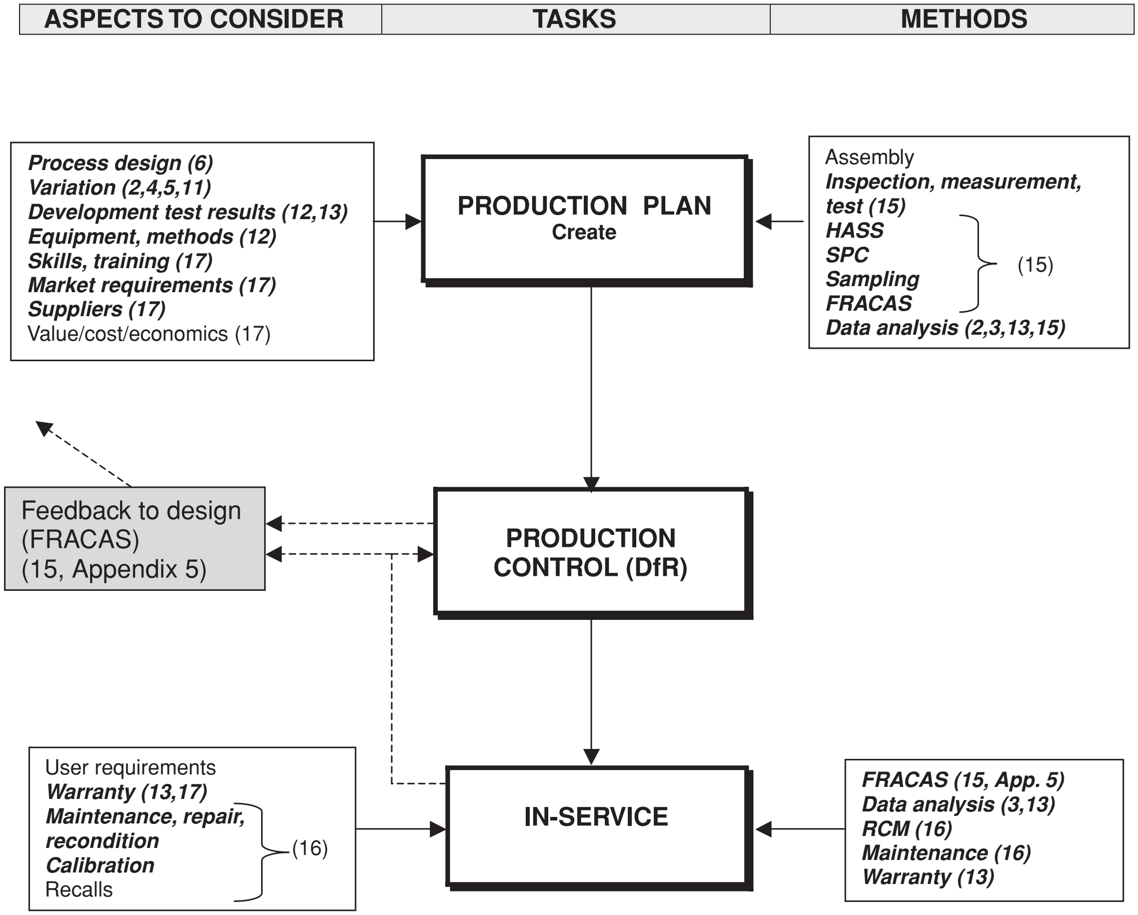
Since production quality will affect reliability, quality control is an integral part of the reliability programme. Quality control cannot make up for design shortfalls, but poor quality can negate much of the reliability

Practical Reliability Engineering, Fifth Edition. Patrick D. T. O'Connor and Andre Kleyner.



Notes:
1. Items in italics are reliability specific aspects.
2. Figures in brackets indicate relevant chapters.
3. Shaded boxes indicate processes that are usually iterative.
4. Dotted lines indicate data feedback (FRACAS).

Figure 17.1 *Reliability Programme Flow (design/development).*



Notes:

1. Items in italics directly influence reliability.
2. Figures in brackets indicate relevant chapters.
3. Shaded boxes indicate processes that are usually iterative.
4. Dotted lines indicate data feedback (FRACAS).

Figure 17.2 *Reliability Programme Flow (production, in-service).*

effort. The quality control effort must be responsive to the reliability requirement and must not be directed only at reducing production costs and the passing of a final test or inspection. Quality control can be made to contribute most effectively to the reliability effort if:

1 Quality procedures, such as test and inspection criteria, are related to factors which can affect reliability, and not only to form and function. Examples are tolerances, inspection for flaws which can cause weakening, and the need for adequate screening when appropriate.

2 Quality control test and inspection data are integrated with the other reliability data.
3 Quality control personnel are trained to recognize the relevance of their work to reliability, and trained and motivated to contribute.

An integrated reliability programme must be disciplined. Whilst creative work such as design is usually most effective when not constrained by too many rules and guidelines, the reliability (and quality) effort must be tightly controlled and supported by mandatory procedures. The disciplines of design analysis, test, reporting, failure analysis and corrective action must be strictly imposed, since any relaxation can result in a reduction of reliability, without any reduction in the cost of the programme. There will always be pressure to relax the severity of design analyses or to classify a failure as non-relevant if doubt exists, but this must be resisted. The most effective way to ensure this is to have the agreed reliability programme activities written down as mandatory procedures, with defined responsibilities for completing and reporting all tasks, and to check by audit and during programme reviews that they have been carried out. More material on integrating reliability programmes can be found in Silverman (2010).

17.3 Reliability and Costs

Achieving high reliability is expensive, particularly when the product is complex or involves relatively untried technology. The techniques described in earlier chapters require the resources of trained engineers, management time, test equipment and products for testing, and it often appears difficult to justify the necessary expenditure in the quest for an inexact quantity such as reliability. It can be tempting to trust to good basic design and production and to dispense with specific reliability effort, or to provide just sufficient effort to placate a customer who insists upon a visible programme without interfering with the 'real' development activity. However, experience points to the fact that all well-managed reliability efforts pay off.

17.3.1 Costs of Reliability

There are usually practical limits to how much can be spent on reliability during a development programme. However, the authors are unaware of any programme in which experience indicated that too much effort was devoted to reliability or that the law of diminishing returns was observed to be operating to a degree which indicated that the programme was saturated. This is mainly due to the fact that nearly every failure mode experienced in service is worth discovering and correcting during development, owing to the very large cost disparity between corrective action taken during development and similar action (or the cost of living with the failure mode) once the equipment is in service. The earlier in a development programme that the failure mode is identified and corrected the cheaper it will be, and so the reliability effort must be instituted at the outset and as many failure modes as possible eliminated during the early analysis, review and test phases. Likewise, it is nearly always less costly to correct causes of production defects than to live with the consequences in terms of production costs and unreliability.

It is dangerous to generalize about the cost of achieving a given reliability value, or of the effect on reliability of stated levels of expenditure on reliability programme activities. Some texts show a relationship as in Figure 1.7 (Chapter 1) with the lowest total cost (life cycle cost – LCC) indicating the 'optimum reliability' (or quality) point. However, this can be a misleading picture unless all cost factors contributing to the reliability (and unreliability) are accounted for. The direct failure costs can usually be estimated fairly accurately, related to assumed reliability levels and yields of production processes, but the cost of achieving these levels is much more difficult to forecast. There are different models accounting for the cost of reliability. For example Kleyner and Sandborn (2008) propose a comprehensive life cycle cost model accounting for the various aspects of achieving and demonstrating reliability as well as the consequent warranty costs. As described above, the relationship is more likely to be a decreasing one, so that the optimum quality and reliability is in fact closer to 100 %, as shown in Figure 1.9 (Chapter 1).

Several standard references on quality management suggest considering costs under three headings, so that they can be identified, measured and controlled. These *quality costs* are the costs of all activities specifically directed at reliability and quality control, and the costs of failure. Quality costs are usually considered in three categories:

1 Prevention costs.
2 Appraisal costs.
3 Failure costs.

Prevention costs are those related to activities which prevent failures occurring. These include reliability efforts, quality control of bought-in components and materials, training and management.

Appraisal costs are those related to test and measurement, process control and quality audit.

Failure costs are the actual costs of failure. Internal failure costs are those incurred during manufacture. These cover scrap and rework costs (including costs of related work in progress, space requirements for scrap and rework, associated documentation, and related overheads). Failure costs also include external or post-delivery failure costs, such as warranty costs; these are the costs of unreliability.

Obviously it is necessary to minimize the sum of quality and reliability costs over a suitably long period. Therefore the immediate costs of prevention and appraisal must be related to the anticipated effects on failure costs, which might be affected over several years. Investment analysis related to quality and reliability is an uncertain business, because of the impossibility of accurately predicting and quantifying the results. Therefore the analysis should be performed using a range of assumptions to determine the sensitivity of the results to assumed effects, such as the yield at test stages and reliability in service.

For example, two similar products, developed with similar budgets, may have markedly different reliabilities, due to differences in quality control in production, differences in the quality of the initial design or differences in the way the reliability aspects of the development programme were managed. It is even harder to say by how much a particular reliability activity will affect reliability. $ 20 000 spent on FMECA might make a large or a negligible difference to achieved reliability, depending upon whether the failure modes uncovered would have manifested themselves and been corrected during the development phase, or the extent to which the initial design was free of shortcomings.

The value gained from a reliability programme must, to a large extent, be a subjective judgement based upon experience and related to the way the programme is managed. The reliability programme will usually be constrained by the resources which can be usefully applied within the development time-scale. Allocation of resources to reliability programme activities should be based upon an assessment of the risks. For a complex new design, design analysis must be thorough and searching, and performed early in the programme. For a relatively simple adaptation of an existing product, less emphasis may be placed on analysis. In both cases the test programme should be related to the reliability requirement, the risks assessed in achieving it and the costs of non-achievement. The two most important features of the programme are:

1 The statement of the reliability aim in such a way that it is understood, feasible, mandatory and demonstrable.
2 Dedicated, integrated management of the programme.

Provided these two features are present, the exact balance of resources between activities will not be critical, and will also depend upon the type of product. A strong test–analyse-fix programme can make up for deficiencies in design analysis, albeit at higher cost; an excellent design team, well controlled and supported by good design rules, can reduce the need for testing. The reliability programme for an electronic equipment will not be the same as for a power station. As a general rule, all the reliability programme activities described in this book are worth applying insofar as they are appropriate to the product, and the earlier in the programme they can be applied the more cost-effective they are likely to prove.

In a well-integrated design, development and production effort, with all contributing to the achievement of high quality and reliability, and supported by effective management and training, it is not possible to isolate the costs of reliability and quality effort. The most realistic and effective approach is to consider all such effort as investments to enhance product performance and excellence, and not to try to classify or analyse them as though they were burdens.

17.3.2 Costs of Unreliability

The costs of unreliability in service should be evaluated early in the development phase, so that the effort on reliability can be justified and requirements can be set, related to expected costs. The analysis of unreliability costs takes different forms, depending on the type of development programme and how the product is maintained. The example below illustrates a typical situation.

Example 17.1

A commercial electronic communication system is to be developed as a risk venture. The product will be sold outright, with a two year parts and labour warranty. Outline the LCC analysis approach and comment on the support policy options.

The analysis must take account of direct and indirect costs. The direct costs can be related directly to failure rate (or removal rate, which is likely to be higher).

The *direct costs* are:

1 Warranty repair costs.
The annual warranty repair cost will be:

$$(\text{Number of warranted units in use}) \times (\text{annual call rate per unit}) \times (\text{cost per call}).$$

The number of warranted units will be obtained from market projections. The call rate will be related to MTBF and expected utilization.

2 Spares production and inventory costs for warranty support.
Spares costs: to be determined by analysis (e.g. Poisson model, simulation) using inputs of call rate, proportion of calls requiring spares, spares costs, probability levels of having spares in stock, repair time to have spares back in stock, repair and stockholding costs.

3 Net of profits on post-warranty repairs and spares.
Annual profit on post-warranty spares and repairs: analysis to be similar to warranty costs analysis, but related to post-warranty equipment utilization.

Indirect costs (not directly related to failure or removal rate):

1 Service organization (training manuals, overheads). (Warranty period contribution).
2 Product reputation.

These costs cannot be derived directly. A service organization will be required in any case and its performance will affect the product's reputation. However, a part of its costs will be related to servicing the warranty. A parametric estimate should be made under these headings, for example:

Service organization: 50 % of annual warranty cost in first two years, 25 % thereafter.

Product reputation: agreed function of call rate.

Since these costs will accrue at different rates during the years following launch, they must all be evaluated for, say, the first five years. The unreliability costs progression should then be plotted (Figure 17.3) to show the relationship between cost and reliability.

The net present values of unreliability cost should then be used as the basis for planning the expenditure on the reliability programme.

This situation would be worth analysing from the point of view of what support policy might show the lowest cost for varying call rates. For example, a very low call rate might make a 'direct exchange, no repair' policy cost-effective, or might make a longer warranty period worth considering, to enhance the product's reputation. Direct exchange would result in service department savings, but a higher spares cost.

The example given above involves very simple analysis and simplifying assumptions. A Monte Carlo simulation (Chapter 4) would be a more suitable approach if we needed to consider more complex dynamic effects, such as distributed repair times and costs, multi-echelon repair and progressive increase in units at risk. However, simple analysis is often sufficient to indicate the magnitude of costs, and in many cases this is all that is needed, as the input variables in logistics analysis are usually somewhat imprecise, particularly failure (removal) rate values. Simple analysis is adequate if relatively gross decisions are required. However, if it is necessary to attempt to make more precise judgments or to perform sensitivity analyses, then more powerful methods such as simulation should be considered.

There are of course other costs which can be incurred as a result of a product's unreliability. Some of these are hard to quantify, such as goodwill and market share, though these can be very large in a competitive situation and where consumer organizations are quick to publicize failure. In extreme cases unreliability can lead to litigation, especially if damage or injury results. An unreliability cost often overlooked is that due to

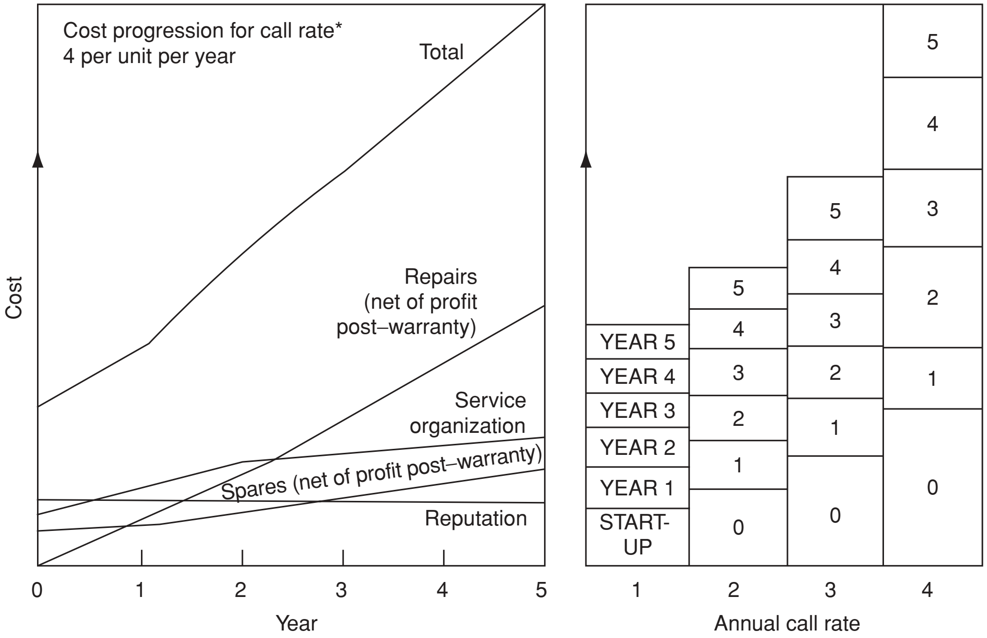


Figure 17.3 *Reliability cost progression (Example 17.1).*

failures in production due to unreliable features of the design. A reliable product will usually be cheaper to manufacture, and the production quality cost monitoring system should be able to highlight those costs which are due to design shortfalls.

17.4 Safety and Product Liability

Product liability legislation in the United States, Europe and in other countries adds a new dimension to the importance of eliminating safety-related failure modes, as well as to the total quality assurance approach in product development and manufacture. Before product liability (PL), the law relating to risks in using a product was based upon the principle of *caveat emptor* ('let the buyer beware'). PL introduced *caveat venditor* ('let the supplier beware'). PL makes the manufacturer of a product liable for injury or death sustained as a result of failure of his product. A designer can now be held liable for a failure of his design, even if the product is old and the user did not operate or maintain it correctly. Claims for death or injury in many product liability cases can only be defended successfully if the producer can demonstrate that he has taken all practical steps towards identifying and eliminating the risk, and that the injury was entirely unrelated to failure or to inadequate design or manufacture. Since these risks may extend over ten years or even indefinitely, depending upon the law in the country concerned, long-term reliability of safety-related features becomes a critical requirement. The size of the claims, liability being unlimited in the United States, necessitates top management involvement in minimizing these risks, by ensuring that the organization and resources are provided to manage and execute the quality and reliability tasks which will ensure reasonable protection. PL insurance is a business area for the insurance companies, who naturally expect to see a suitable reliability and safety programme being operated by the manufacturers they insure.

Abbot and Tyler (1997) provide an overview of this topic.

17.5 Standards for Reliability, Quality and Safety

Reliability programme standard requirements have been issued by several large agencies which place development contracts on industry. The best known of these was US MIL-STD-785 – *Reliability Programs for Systems and Equipments, Development and Production* which covers all development programmes for the US Department of Defense. MIL-STD-785 was supported by other military standards, handbooks and specifications, and these have been referenced in earlier chapters. This book has referred mainly to US military documents, since in many cases they are the most highly developed and best known. However, in 1995 the US Department of Defense cancelled most military standards and specifications, including MIL-STD-785, and downgraded others to handbooks (HDBK) for guidance only. Military suppliers are now required to apply 'best industry practices', rather than comply with mandatory standards.

In the United Kingdom, Defence Standards 00–40 and 00–41 cover reliability programme management and methods for defence equipment, and BS 5760 has been published for commercial use, and can be referenced by any organization in developing contracts. ARMP-1 is the NATO standard on reliability and maintainability. Some large agencies such as NASA, major utilities and corporations issue their own reliability standards. International standards have also been issued, and some of these are described below.

These official standards generally (but not all) tend to over-emphasize documentation, quantitative analysis, and formal test. They do not reflect the integrated approach described in this book and used by many modern engineering companies. They suffer from slow response to new ideas. Therefore they are not much used outside the defence and related industries. However, it is necessary for people involved in a reliability programme, whether from the customer or supplier side, to be familiar with the appropriate standards.

17.5.1 ISO/IEC60300 (Dependability)

ISO/IEC60300 is the international standard for '*dependability*', which is defined as covering reliability, maintainability and safety. It describes management and methods related to these aspects of product design and development. The methods covered include reliability prediction, design analysis, maintenance and support, life cycle costing, data collection, reliability demonstration tests, and mathematical/statistical techniques; most of these are described in separate standards within the ISO/IEC60000 series. Manufacturing quality aspects are not included in this standard. For more on IEC standards see Barringer (2011).

ISO/IEC60300 has not, so far, been made the subject of audits and registration in the way that ISO9000 has (see next section).

17.5.2 ISO9000 (Quality Systems)

The international standard for quality systems, IS09000, has been developed to provide a framework for assessing the quality management system which an organization operates in relation to the goods or services provided. The concept was developed from the US Military Standard for quality management, MIL-Q-9858, which was introduced in the 1950s as a means of assuring the quality of products built for the US military. However, many organizations and companies rely on ISO9000 registration to provide assurance of the quality of products and services they buy and to indicate quality of their products and services.

Registration is to the relevant standard within the ISO9000 'family'. ISO9001 is the standard applicable to organizations that design, develop and manufacture products. We will refer to 'ISO9000 registration' as a general indication.

ISO9000 does not specifically address the quality of products and services, nor does it prescribe methods for achieving quality, such as design analysis, test and quality control. It describes, in very general terms, the system that should be in place to assure quality. In principle, there is nothing in the standard to prevent an organization from producing poor quality goods or services, so long as written procedures exist and are followed. An organization with an effective quality system would normally be more likely to take corrective action and to improve processes and service, than would one which is disorganized. However, the fact of registration should not be considered as a guarantee of quality.

In the ISO9000 approach, suppliers' quality management systems (organization, procedures, etc.) are audited by independent 'third party' assessors, who assess compliance with the standard and issue certificates of registration. Certain organizations are 'accredited' as 'certification bodies' by the appropriate national accreditation services. The justification given for third party assessment is that it removes the need for every customer to perform his own assessment of all of his suppliers. However, a matter as important as quality cannot safely be left to be assessed infrequently by third parties, who are unlikely to have the appropriate specialist knowledge, and who cannot be members of the joint supplier–purchaser team. The *total quality management* (TQM) philosophy (see Section 17.17.5) demands close, ongoing partnership between suppliers and purchasers.

Since its inception, ISO9000 has generated considerable controversy. The effort and expense that must be expended to obtain and maintain registration tend to engender the attitude that optimal standards of quality have been achieved. The publicity that typically goes with initial certification of a business supports this belief. The objectives of the organization, and particularly of the staff directly involved in obtaining and maintaining registration, are directed at the maintenance of procedures and at audits to ensure that staff work to them. It can become more important to work to procedures than to develop better ways of working. However, some organizations have generated real improvements as a result of certification. So why is the approach so widely used? The answer is partly cultural and partly coercive.

The cultural pressure derives from the tendency to believe that people perform better when told what to do, rather than when they are given freedom and the necessary skills and motivation to determine the best ways to perform their work. This belief stems from the concept of *scientific management*, as described in Drucker (1955) and O'Connor (2004).

The coercion to apply the standard comes from several directions. In practice, many agencies demand that bidders for contracts must be registered. All contractors and their subcontractors supplying the UK Ministry of Defence must be registered, since the MoD decided to drop its own assessments in favour of the third party approach, and the US Department of Defense decided to apply ISO9000 in place of MIL-STD-Q9858. Several large companies, as well as public utilities, demand that their suppliers are registered to ISO9000 or to industry versions, such as QS9000 for automotive and TL9000 for telecommunications applications. Defenders of ISO9000 say that the total quality management (TQM) approach is too severe for most organizations, and that ISO9000 can provide a 'foundation' for a subsequent total quality effort. However, the foremost teachers of modern quality management all argued against this view. They point out that any organization can adopt the TQM philosophy, and that it will lead to greater benefits than will registration to the standard, and at lower costs.

There are many books in publication that describe ISO9000 and its application: Hoyle (2009) is an example.

17.5.3 IEC61508 (Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related Systems)

IEC61508 is the international standard to set requirements for design, development, operation and maintenance of 'safety-related' control and protection systems which are based on electrical, electronics and software technologies. A system is 'safety-related' if any failure to function correctly can present a hazard to people. Thus, systems such as railway signalling, vehicle braking, aircraft controls, fire detection, machine safety interlocks, process plant emergency controls and car airbag initiation systems would be included. The standard lays down criteria for the extent to which such systems must be analysed and tested, including the use of independent assessors, depending on the criticality of the system. It also describes a number of methods for analysing hardware and software designs.

The extent to which the methods are to be applied is determined by the required or desired *safety integrity level* (SIL) of the safety function, which is stated on a range from 1 to 4. SIL 4 is the highest, relating to a 'target failure measure' between 10^{-5}and10^{-4} per demand, or 10^{-9}and10^{-8} per hour. For SIL 1 the figures are $10^{-2} - 10^{-1}$and$10^{-6} - 10^{-5}$. For the quantification of failure probabilities reliability prediction methods such as those covered in Chapter 6 are recommended. The methods listed include 'use of well-tried components' (recommended for all SILs), 'simulation' (recommended for SIL 2, 3 and 4), and 'modularization' ('highly recommended' for all SILs).

In the early 2000's ISO produced an automotive version of this standard, IEC 26262. It addresses functional safety in a way similar to IEC61508 with appropriate adaptations for road vehicles.

The value and merits for both standards are somewhat questionable. The methods described are often inconsistent with accepted best industry practices. The issuing of these standards has lead to a growth of bureaucracy, auditors and consultants, and increased costs. It is unlikely to generate any improvements in safety, for the same reasons that ISO9000 does not necessarily improve quality. However, they exist and compliance is often mandatory, so system designers must be aware of them and ensure that the requirements are met.

17.6 Specifying Reliability

In order to ensure that reliability is given appropriate attention and resources during design, development and manufacture, the requirement must be specified. Before describing how to specify reliability adequately, we will cover some of the ways of how *not* to do it:

1 Do not write vague requirements, such as 'as reliable as possible', 'high reliability is to be a feature of the design', or 'the target reliability is to be 99 %'. Such statements do not provide assurance against reliability being compromised.
2 Do not write unrealistic requirements. 'Will not fail under the specified operation conditions' is a realistic requirement in many cases. However, an unrealistically high reliability requirement for, say, a complex electronic equipment will not be accepted as a credible design parameter, and could therefore be ignored.

The reliability specification *must* contain:

1 A definition of failure related to the product's function. The definition should cover all failure modes relevant to the function.
2 A full description of the environments in which the product will be stored, transported, operated and maintained.
3 A statement of the reliability requirement, and/or a statement of failure modes and effects which are particularly critical and which must therefore have a very low probability of occurrence. Examples of reliability metrics to be used are discussed in Section 14.2. GMW 3172 (2004) can be used as an example of a detailed reliability specification. Also UK Defense Standard 00-40 covers the preparation of reliability specifications in detail.

17.6.1 Definition of Failure

Care must be taken in defining failure to ensure that the failure criteria are unambiguous. Failure should always be related to a measurable parameter or to a clear indication. A seized bearing indicates itself clearly, but a leaking seal might or might not constitute a failure, depending upon the leak rate, or whether or not the leak can be rectified by a simple adjustment. An electronic equipment may have modes of failure which do not affect function in normal operation, but which may do so under other conditions. For example, the failure of a diode used to block transient voltage spikes may not be apparent during functional test, and will probably not affect normal function. Defects such as changes in appearance or minor degradation that do not affect function are not usually relevant to reliability. However, sometimes a perceived degradation is an indication that failure will occur and therefore such incidents can be classified as failures.

Inevitably there will be subjective variations in assessing failure, particularly when data are not obtained from controlled tests. For example, failure data from repairs carried out under warranty might differ from data on the same equipment after the end of the warranty period, and both will differ from data derived from a controlled reliability demonstration. The failure criteria in reliability specifications can go a long way to reducing the uncertainty of relating failure data to the specification and in helping the designer to understand the reliability requirement.

17.6.2 Environmental Specifications

The environmental specification must cover all aspects of the many loads and other effects that can influence the product's strength or probability of failure. Without a clear definition of the conditions which the product will face, the designer will not be briefed on what he is designing against. Of course, aspects of the environmental specification might sometimes be taken for granted and the designer might be expected to cater for these conditions without an explicit instruction. It is generally preferable, though, to prepare a complete environmental specification for a new product, since the discipline of considering and analysing the likely usage conditions is a worthwhile exercise if it focuses attention on any aspect which might otherwise be overlooked in the design. Environments are covered in Section 7.3.2. For most design groups only a limited number of standard environmental specifications is necessary. For example, the environmental requirements and methods of test for military equipment are covered in specifications such as US MIL-STD-810 and UK Defence Standard 07–55. Another good example is the automotive validation standard GMW3172 (2004) mentioned earlier.

The environments to be covered must include handling, transport, storage, normal use, foreseeable misuse, maintenance and any special conditions. For example, the type of test equipment likely to be used, the skill level of users and test technicians, and the conditions under which testing might be performed should be stated if these factors might affect the observed reliability.

17.6.3 Stating the Reliability Requirement

The reliability requirement should be stated in a way which can be verified, and which makes sense relative to the use of the product. For example, there is little point in specifying a time between failures if the product's operation will be measured only in distance travelled, or if it will not be measured at all (either in a reliability demonstration or during service).

Levels of reliability can be stated as a success ratio, or as a life. For 'one-shot' items the success ratio is the only relevant criterion.

Reliability specifications based on life parameters must be framed in relation to the appropriate life distributions. The examples of reliability metrics appropriate for reliability specifications are covered in Section 14.2.

Specified life parameters must clearly state the life characteristic. For example, the life of a switch, a sequence valve or a data recorder cannot be usefully stated merely as a number of hours. The life must be related to the duty cycle (in these cases switch reversals and frequency, sequencing operations and frequency, and anticipated operating cycles on record, playback and switch on/off). The life parameter may be stated as some time-dependent function, for example distance travelled, switching cycles, load reversals, or it may be stated as a time, with a stipulated operating cycle.

17.7 Contracting for Reliability Achievement

Users of equipment which can have high unreliability costs have for some time imposed contractual conditions relative to reliability. Of course, every product warranty is a type of reliability contract. However, contracts which stipulate specific incentives or penalties related to reliability achievement have been developed, mainly by the military, but also by other major equipment users such as airlines and public utilities.

The most common form of reliability contract is one which ties an incentive or penalty to a reliability demonstration. The demonstration may either be a formal test (see the methods covered in Chapter 14) or may be based upon the user's experience. In either case, careful definition of what constitutes a relevant failure is necessary, and a procedure for failure classification must be agreed. If the contract is based only on incentive payments, it can be agreed that the customer will classify failures and determine the award, since no penalty is involved. One form of reliability incentive contract is that used for spacecraft, whereby the customer pays an incentive fee for successful operation for up to, say, two years.

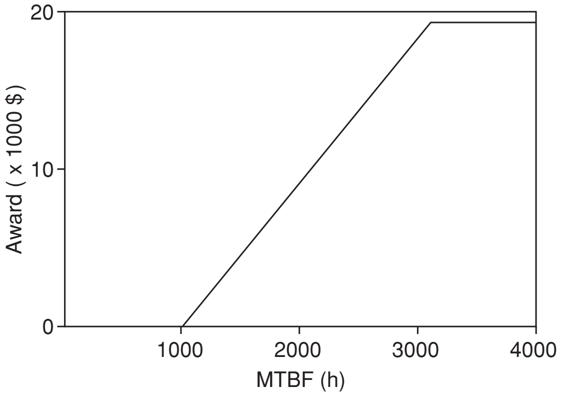


Figure 17.4 *Reliability incentive structure.*

Incentive payments have advantages over incentive/penalty arrangements. It is important to create a positive motivation, rather than a framework which can result in argument or litigation, and incentives are preferable in this respect. Also, an incentive is easier to negotiate, as it is likely to be accepted as offered. Incentive payments can be structured so that, whilst they represent a relatively small percentage of the customer's saving due to increased reliability, they provide a substantial increase in profit to the supplier. The receipt of an incentive fee has significant indirect advantages, as a morale booster and as a point worth quoting in future bid situations. A typical award fee structure is shown in Figure 17.4.

When planning incentive contracts it is necessary to ensure that other performance aspects are sufficiently well specified and, if appropriate, also covered by financial provisions such as incentives or guarantees, so that the supplier is not motivated to aim for the reliability incentive at the expense of other features. Incentive contracting requires careful planning so that the supplier's motivation is aligned with the customer's requirements. The parameter values selected must provide a realistic challenge and the fee must be high enough to make extra effort worthwhile.

17.7.1 Warranty Improvement Contracts

Billions of dollars are spent by manufacturers on warranty each year. Therefore in many industries manufacturers make the efforts to motivate their suppliers to improve their reliability and consequently reduce warranty. For example, some automotive manufacturers would cover the cost of warranty up to a certain failure rate (e.g., one part per thousand vehicles per year), and everything above that rate would be the supplier's responsibility. In recent years there have been more and more efforts on the part of manufacturers to force their suppliers to pay their 'fair share' of the warranty costs. Warranty Week (2011) newsletter publishes a comprehensive coverage of finance and management aspects of warranty for various industries and regions.

17.7.2 Total Service Contracts

A total service supply contract is one in which the supplier is required to provide the system, as well as all of the support. The purchaser does not specify a quantity of systems, but a level of availability. Railway rolling stock contracts in Europe epitomize the approach: the rail company specifies a timetable, and the supplier must determine and build the appropriate number of trains, provide all maintenance and other logistic support, including staffing and running the maintenance depots, spares provisioning, and so on.

All the train company does is operate the trains. The contracts include terms to cover failures to meet the timetable due to train failures or non-availability. Similar contracts are used for the supply of electronic instrumentation to large buyers, medical equipment to hospitals, and some military applications such as trainer aircraft.

A total supply contract places the responsibility and risk aspects of reliability firmly with the supplier, and can therefore be highly motivating. However, there can be long-term disadvantages to the purchaser. The purchaser's organization can lose the engineering knowledge that might be important in optimizing trade-offs between engineering and operational aspects, and in planning future purchases. When the interface between engineering and operation is purely financial and legal, with separate companies working to different business objectives, conflicts of interests can arise, leading to sub-optimization and inadequate co-operation. Since the support contract, once awarded, cannot be practically changed or transferred, the supplier is in a monopoly situation. The case of Railtrack in the UK, which 'outsourced' all rail and other infrastructure maintenance to contractors, and in the process lost the knowledge necessary for effective management, resulting directly in a fatal crash and a network-wide crisis due to cracked rails, provides a stark warning of the potential dangers of the approach. It is interesting that, by contrast, the airlines continue to perform their own maintenance, and the commercial aircraft manufacturers concentrate on the business they know best.

17.8 Managing Lower-Level Suppliers

Lower-level suppliers can have a major influence on the reliability of systems. It is quite common that 80 % or more of failures of systems such as trains, aircraft, ships, factory and infrastructure systems, and so on can be 'bought' from the lower-level suppliers. In smaller systems, such as machines and electronic equipment, items such as engines, hydraulic pumps and valves, power supplies, displays, and so on are nearly always bought from specialist suppliers, and their failures contribute to system unreliability. Therefore it is essential that the project reliability effort is directed as much to these suppliers as to internal design, development and manufacturing. Continuous globalization and outsourcing also affect the work with lower-level suppliers. It is not uncommon these days to have suppliers located all over the globe including regions with little knowledge depth regarding design and manufacturing processes and with a less robust quality system in place.

To ensure that lower-level suppliers make the best contributions to system reliability, the following guidelines should be applied:

1 Rely on the existing commercial laws that govern trading relationships to provide assurance. In all cases this provides for redress if products or services fail to achieve the performance specified or implied. Therefore, if failures do occur, action to improve or other appropriate action can be demanded. In many cases warranty terms can also be exercised. However, if the contract stipulates a value of reliability to be achieved, such as MTBF or maximum proportion failing, then in effect failures are being invited. When failures do occur there is a tendency for discussion and argument to concentrate on statistical interpretations and other irrelevancies, rather than on engineering and management actions to prevent recurrence. We should specify success, not failure.

2 Do not rely solely on ISO 9000 or similar schemes to provide assurance. As explained above, these approaches provide no direct assurance of product or service reliability or quality.

3 Engineers should manage the selection and purchase of engineering products. It is a common practice for companies to assign this function to a specialized purchasing organization, and design engineers must submit specifications to them. This is based on the argument that engineers might not have knowledge of the business aspects that purchasing specialists do. However, only the engineers concerned can be expected to understand the engineering aspects, particularly the longer-term impact of reliability. Engineers can quickly be taught sufficient purchasing knowledge, or can be supported by purchasing experts, but

purchasing people cannot be taught the engineering knowledge and experience necessary for effective selection of engineering components and sub-systems.

4 Do not select suppliers on the basis of the price of the item alone. (This is point number 4 of Deming's famous '14 points for managers', Deming, 1987.) Suppliers and their products must be selected on the basis of total value to the system, over its expected life. This includes performance, reliability, support, and so on as well as price.

5 Create long-term partnerships with suppliers, rather than seek suppliers on a project-by-project basis or change suppliers for short-term advantage. In this way it becomes practicable to share information, rewards and risks, to the long-term benefit of both sides. Suppliers in such partnerships can teach application details to the system designers, and can respond more effectively to their requirements.

These points are discussed in more detail in O'Connor (2004).

17.9 The Reliability Manual

Just as most medium to large design and development organizations have internal manuals covering design practices, organizational structure, Quality Assurance (QA) procedures, and so on, so reliability management and engineering should be covered. Depending on the type of product and the organization adopted, reliability can be adequately covered by appropriate sections in the engineering and QA manuals, or a separate reliability manual may be necessary. In-house reliability procedures should not attempt to teach basic principles in detail, but rather should refer to appropriate standards and literature, which should of course then be made available. If reliability programme activities are described in military, national or industry standards, these should be referred to and followed when appropriate. The bibliographies at the end of each chapter of this book list the major references.

The in-house documents should cover, as a minimum the following subjects:

1 Corporate policy for reliability.
2 Organization for reliability.
3 Reliability procedures in design (e.g. design analysis, parts derating policy, parts, materials and process selection approval and review, critical items listing, design review).
4 Reliability test procedures.
5 Reliability data collection, analysis and action system, including data from test, warranty, etc (see Appendix 5).

The written procedures must state, in every case, who carries responsibility for action and who is responsible for providing the resources and capability. They must also state who provides supporting services. A section from the reliability manual may appear as shown in Table 17.1.

Table 17.1 *Reliability manual: responsibilities.*

		Department responsible		
Task	Reference	Prime	Resources	Support
Stress analysis (electronic)	Procedure XX	Project design	Reliability	Reliability
Reliability test	Procedure YY	Reliability	Environmental test	Project design
Reliability data	Procedure ZZ	Reliability	Reliability	QA,

17.10 The Project Reliability Plan

Every project should create and work to a written reliability plan. It is normal for customer-funded development projects to include a requirement for a reliability plan to be produced.

The reliability plan should include:

1 A brief statement of the reliability requirement.
2 The organization for reliability.
3 The reliability activities that will be performed (design analysis, test, reports).
4 The timing of all major activities, in relation to the project development milestones.
5 Reliability management of suppliers.
6 The standards, specifications and internal procedures (e.g. the reliability manual) which will be used, as well as cross-references to other plans such as for test, safety, maintainability and quality assurance.

When a reliability plan is submitted as part of a response to a customer request for proposals (RFP) in a competitive bid situation, it is important that the plan reflects complete awareness and understanding of the requirements and competence in compliance.

A reliability plan prepared as part of a project development, after a contract has been accepted, is more comprehensive than an RFP response, since it gives more detail of activities, time-scales and reporting. Since the project reliability plan usually forms part of the contract once accepted by the customer, it is important that every aspect is covered clearly and explicitly.

A well-prepared reliability plan is useful for instilling confidence in the supplier's competence to undertake the tasks, and for providing a sound reliability management plan for the project to follow.

Appendix 6 shows an example of a reliability and maintainability plan, including safety aspects.

17.10.1 Specification Tailoring

Specification tailoring is a term used to describe the process of suggesting alternatives to the customer's specification. 'Tailoring' is often invited in RFPs and in development contracts. A typical example occurs when a customer specifies a system and requires a formal reliability demonstration. If a potential supplier can supply a system for which adequate in-service reliability records exist, the specification could be tailored by proposing that these data be used in place of the reliability demonstration. This could save the customer the considerable expense of the demonstration. Also, it is not uncommon for a potential supplier to take an exception to certain requirements in the RFP if they do not appear feasible or possible for the supplier to implement. Other examples might arise out of trade-off studies, which might show, for instance, that a reduced performance parameter could lead to cost savings or reliability improvement.

17.11 Use of External Services

The retention of staff and facilities for analysis and test and the maintenance of procedures and training can only be cost-effective when the products involve fairly intensive and continuous development. In advanced product areas such as defence and aerospace, electronic instrumentation, control and communications, vehicles, and for large manufacturers of less advanced products such as domestic equipment and less complex industrial equipment, a dedicated reliability engineering organization is necessary, even if it is not a contractual requirement. Smaller companies with less involvement in risk-type development may have as great a need for reliability engineering expertise, but not on a continuous basis. External reliability engineering services can fulfil the requirements of smaller companies by providing the specialist support and facilities when needed. Reliability engineering consultants and specialist test establishments can often be useful to larger companies also, in support of internal staff and facilities. Since they are engaged full time across a number of different types of project they should be considered whenever new problems arise. However, they should be selected carefully and integrated in the project team.

Small companies should also be prepared to seek the help of their major customers when appropriate. This cooperative approach benefits both supplier and customer.

17.12 Customer Management of Reliability

When a product is being developed under a development contract, as is often the case with military and other public purchasing, the purchasing organization plays an important role in the reliability and quality programme. As has been shown, such organizations often produce standards for application to development contracts, covering topics such as reliability programme management, design analysis methods and test methods.

A reliability manager should be assigned to each project, reporting to the project manager. Project reliability management by a centralized reliability department, not responsible to the project manager, is likely to result in lower effectiveness. A central reliability department is necessary to provide general standards, training and advice, but should not be relied upon to manage reliability programmes across a range of projects. If there is a tendency for this to happen it is usually an indication that inadequate standards or training have been provided for project staff, and these problems should then be corrected.

The prime responsibilities of the purchaser in a development reliability programme are to:

1 Specify the reliability requirements (Sections 17.6 and 14.2).
2 Specify the standards and methods to be used.
3 Set up the financial and contractual framework (Section 17.7).
4 Specify the reporting requirements.
5 Monitor contract performance.

Proper attention to the first three items above should ensure that the supplier is effectively directed and motivated, so that the purchaser has visibility of activities and progress without having to become too deeply involved.

It is usually necessary to negotiate aspects of the specification and contract. During the specification and negotiation phases it is usual for a central reliability organization to be involved, since it is important that uniform approaches are applied. Specification tailoring (Section 17.10.1) is now a common feature of development contracting and this is an important aspect in the negotiation phase, requiring experience and knowledge of the situation of other contracts being operated or negotiated.

The supplier's reliability plan, prepared in response to the purchaser's requirement, should also be reviewed by the central organization, particularly for major contracts.

The contractor's reporting tasks are often specified in the statement of work (SOW). These usually include:

1 The reliability plan.
2 Design analysis reports and updates (prediction, FMECA, FTA, etc.).
3 Test reports.

Reporting should be limited to what would be useful for monitoring performance. For example, a 50-page FMECA report, tabulating every failure mode in a system, is unlikely to be useful to the purchaser. Therefore the statement of work should specify the content, format and size of reports. The detailed analyses leading to the reports should be available for specific queries or for audit.

The purchaser should observe the supplier's design reviews. Some large organizations assign staff to supplier's premises, to monitor development and to advise on problems such as interpretation of specifications. This can be very useful on major projects such as aircraft, ships and plant, particularly if the assigned staff are subsequently involved in operation and maintenance of the system.

There are many purchasers of equipment who do not specify complete systems or let total development contracts. Also, many such purchasers do not have their own reliability standards. Nevertheless, they can usually influence the reliability and availability of equipment they buy. We will use an example to illustrate how a typical purchaser might do this.

Example 17.2

A medium-sized food-processing plant is being planned by a small group of entrepreneurs. Amongst other things, the plant will consist of:

1 Two large continuous-feed ovens, which are catalogue items but have some modifications added by the supplier, to the purchaser's specification. These are the most expensive items in the plant. There is only one potential supplier.
2 A conveyor feed system.
3 Several standard machines (flakers, packaging machines, etc.).
4 A process control system, operated by a central computer, for which both the hardware and software will be provided by a specialist supplier to the purchaser's specification.

The major installations except item 4 will be designed and fitted by a specialist contractor; the process control system integration will be handled by the purchasers. The plant must comply with the statutory safety standards, and the group is keen that both safety and plant availability are maximized. What should they do to ensure this?

The first step is to ensure that every supplier has, as far as can be ascertained, a good reputation for reliability and service. The purchasers should survey the range of equipment available, and if possible obtain information on reliability and service from other users. Equipment and supplier selection should be based to a large extent on these factors.

For the standard machines, the warranties provided should be studied. Since plant availability is important, the purchasers should attempt to negotiate service agreements which will guarantee up-time, for example for guaranteed repair or replacement within 24 hours. If this is not practicable, they should consider, in conjunction with the supplier, what spares they should hold.

Since the ovens are critical items and are being modified, the purchasers should ensure that the supplier's normal warranty applies, and service support should be guaranteed as for the standard items. They should consider negotiating an extended warranty for these items.

The process control system, being a totally new development (except for the computer), should be very carefully specified, with particular attention given to reliability, safety and maintainability, as described below. Key features of the specification and contract should be:

1 Definition of safety-critical failure effects.
2 Definition of operational failure effects.
3 Validation of correct operation when installed.

4 Guaranteed support for hardware and software, covering all repairs and corrections found to be necessary.
5 Clear, comprehensive documentation (test, operating and maintenance instructions, program listings, program notes).

For this development work, the purchasers should consider invoking appropriate standards in the contract, such as BS 5760. For example, FMECA and FTA could be very valuable for this system, and the software development should be properly controlled and documented. The supplier should be required to show how those aspects of the specification and contract will be addressed, to ensure that the requirements are fully understood. A suitable consultant engineer might be employed to specify and manage this effort.

The installation contract should also cover reliability, safety and maintainability, and service.

During commissioning, all operating modes should be tested. Safety aspects should be particularly covered, by simulating as far as possible all safety-critical failure modes.

The purchasers should formulate a maintenance plan, based upon the guidelines given in Chapter 16. A consultant engineer might be employed for this work also.

Finally, the purchasers should insure themselves against the risks. They should use the record of careful risk control during development to negotiate favourable terms with their insurers.

17.13 Selecting and Training for Reliability

Within the reliability organization, staff are required who are familiar with the product (its design, manufacture and test) and with reliability engineering techniques. Therefore the same qualifications and experience as apply to the other engineering departments should be represented within the reliability organization. The objective should be to create a balanced organization, in which some of the staff are drawn from product engineering departments and given the necessary reliability training, and the others are specialists in the reliability engineering techniques who should receive training to familiarize them with the product. Reliability engineering should be included as part of the normal engineering staff rotation for career development purposes. By having a balanced department, and engineers in other departments with experience of reliability engineering, the reliability effort will have credibility and will make the most effective contribution.

Reliability engineers need not necessarily be specialists in particular disciplines, such as electronic circuit design or metallurgy. Rather, a more widely based experience and sufficient knowledge to understand the specialists' problems is appropriate. The reliability engineer's task is not to solve design or production problems but to help to prevent them, and to ascertain causes of failure. He or she must, therefore, be a communicator, competent to participate with the engineering specialists in the team and able to demonstrate the value and relevance of the reliability methods applied. Experience and knowledge of the product, including manufacturing, operation and maintenance, enables the reliability engineer to contribute effectively and with credibility. Therefore engineers with backgrounds in areas such as test, product support, and user maintenance should be short-listed for reliability engineering positions.

Since reliability engineering and quality control have much in common, quality control work often provides suitable experience from which to draw, provided that the quality control (QC) experience has been deeper than the traditional test and inspection approach, with no design or development involvement. For those in the reliability organization providing data analysis and statistical engineering support, specialist training is relatively more important than product familiarity.

The qualities required of the reliability engineering staff obviously are equally relevant for the head of the reliability function. Since reliability engineering should involve interfaces with several other functions, including such non-engineering areas as marketing and finance, this position should not be viewed as the end of the line for moderately competent engineers, but rather as one in which potential top management staff can develop general talents and further insight into the overall business, as well as providing further reliability awareness at higher levels in due course.

Specialists in statistics and lately six-sigma black belt masters (see Section 17.17.2) can make a significant contribution to the integrated reliability effort. Such skills are needed for design of experiments and analysis of data, and not many engineers are suitably trained and experienced. It is important that statistics experts working in engineering are made aware of the 'noisy' nature of the statistics generated, as described in earlier chapters. They should be taught the main engineering and scientific principles of the problems being addressed, and integrated into the engineering teams. They also have an important role to play in training engineers to understand and use the appropriate statistical methods.

Whilst selection and training of reliability people is important, it is also necessary to train and motivate all other members of the engineering team (design, test, production, etc.). Since product failures are nearly always due to human shortcomings, in terms of lack of knowledge, skill or effort, all involved with the product must be trained so that the chances of such failures are minimized. For example, if electronics designers understand electromagnetic interference as it affects their system they are less likely to provide inadequate protection, and test engineers who understand variation will conduct more searching tests. Therefore the reliability training effort must be related to the whole team, and not just to the reliability specialists.

Despite its importance, quality and reliability education is paradoxically lacking in today's engineering curricula. Few engineering schools offer degree programmes or even a sufficient variety of courses in quality or reliability methods. Therefore, the majority of the quality and reliability practitioners receive their professional training from their colleagues as 'on the job' training.

17.14 Organization for Reliability

Because several different activities contribute to the reliability of a product it is difficult to be categorical about the optimum organization to ensure effective management of reliability. Reliability is affected by design, development, production quality control, control of suppliers and subcontractors, and maintenance. These activities need to be coordinated, and the resources applied to them must be related to the requirements of the product. The requirements may be determined by a market assessment, by warranty cost considerations or by the customer. The amount of customer involvement in the reliability effort varies. The military and other public organizations often stipulate the activities required in detail, and demand access to design data, test records and other information, particularly when the procurement agency funds the development. At the other extreme, domestic customers are not involved in any way directly with the development and production programme. Different activities will have greater or lesser importance depending on whether the product involves innovative or complex design, or is simple and based upon considerable experience. The reliability effort also varies as the project moves through the development, production and in-use phases, so that the design department will be very much involved to begin with, but later the emphasis will shift to the production, quality control and maintenance functions. However, the design must take account of production, test and maintenance, so these downstream activities must be considered by the specification writers and designers.

Since the knowledge, skills and techniques required for the reliability engineering tasks are essentially the same as those required for safety analysis and for maintainability engineering, it is logical and effective to combine these responsibilities in the same department or project team.

Reliability management must be integrated with other project management functions, to ensure that reliability is given the appropriate attention and resources in relation to all the other project requirements and constraints.

Two main forms of reliability organization have evolved. These are described below.

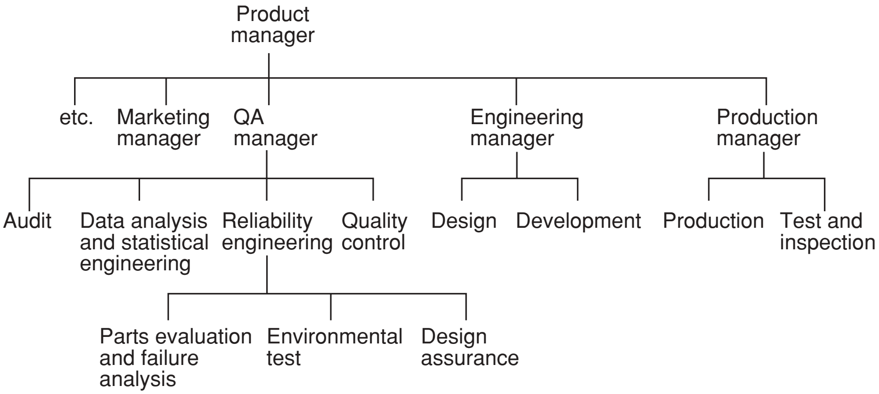


Figure 17.5 *QA based reliability organization.*

17.14.1 Quality Assurance Based Organization

The quality assurance (QA) based organization places responsibility for reliability with QA management, which then controls the 'quality' of design, maintenance, and so on, as well as of production. This organizational form is based upon the definition of *quality as the totality of features which bear on a product's ability to satisfy the requirement*. This is the formal European definition of quality. Consequently, in Europe the QA department or project QA manager is often responsible for all aspects of product reliability. Figure 17.5 shows a typical organization. The reliability engineering team interfaces mainly with the engineering departments, while quality control is mainly concerned with production. However, there is close coordination of reliability engineering and quality control, and shared functions: for example a common failure data collection and analysis system can be operated, covering development, production and in-use. The QA department then provides the feedback loop from in-use experience to future design and production. This form of organization is used by most manufacturers of commercial and domestic products.

17.14.2 Engineering Based Organization

In the engineering based organization, reliability is made the responsibility of the engineering manager. The QA (or quality control) manager is responsible only for controlling production quality and may report direct to the product manager or to the production manager. Figure 17.6 shows a typical organization. This type of organization is more common in the United States.

17.14.3 Comparison of Types of Organization

The QA based organization for reliability allows easier integration of some tasks that are common to design, development and production. The ability to operate a common failure data system has been mentioned. In addition, the statistical methods used to design experiments and to analyse development test and production failure data are the same, as is much of the test equipment and test methods. For example, the environmental test equipment used to perform reliability tests in development might be the same as that used for production reliability acceptance tests and screening. Engineers with experience or qualifications in QA are often familiar with reliability engineering methods, as their professional associations on both sides of the Atlantic include reliability in their areas of interest. However, for products or systems where a considerable amount

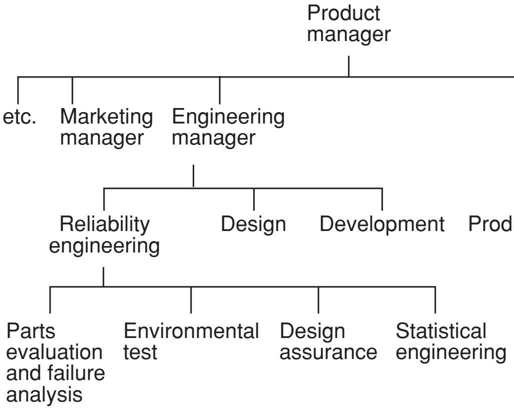


Figure 17.6 *Engineering based reliability organization.*

of innovative design is required, the engineering based organization has advantages, since more of the reliability effort will have to be directed towards design assurance, such as stress analysis, design review and development testing.

The main question to be addressed in deciding upon which type of organization should be adopted is whether the split of responsibility for reliability activities inherent in the engineering based organization can be justified in relation to the amount of reliability effort considered necessary during design and development. In fact the type of organization adopted is much less important than the need to ensure integrated management of the reliability programme. So long as the engineers performing reliability activities are properly absorbed into the project team, and report to the same manager as the engineers responsible for other performance aspects, functional departmental loyalties are of secondary importance. To ensure an integrated team approach reliability engineers should be attached to and work alongside the design and other staff directly involved with the project. These engineers should have access to the departmental supporting services, such as data analysis and component evaluation, but their prime responsibility should be to the project.

17.15 Reliability Capability and Maturity of an Organization

The ability to evaluate how well an organization can handle reliability aspects of the design, development and manufacturing processes (i.e. have the required tool sets, expertise, resources, and the reliability-focused priorities) requires objective criteria. However industry was lacking methods to quantify the capability of an organization to develop and build reliable products. This problem was addressed by developing standardized measurement criteria for assessing and quantifying the reliability capability of an organization. The evaluation methods for organizational reliability processes are *reliability capability* and *reliability maturity* assessments. Sometimes those terms are used interchangeably in regards to an organization and its reliability processes, however there are some subtle differences in the ways they are evaluated.

17.15.1 Reliability Capability

Reliability capability is a measure of the practices within an organization that contribute to the reliability of the final product and the effectiveness of these practices in meeting the reliability requirements of customers

(Tiku and Pecht, 2010). In order to produce industry-accepted criteria for assessing and quantifying the reliability capability IEEE has developed a standard IEEE Std 1624. Standardized and objective measurement criteria define the eight key reliability practices and their inputs, activities, and outputs.

These key reliability practices are the following:

- Reliability requirements and planning.
- Training and development.
- Reliability analysis.
- Reliability testing.
- Supply chain management.
- Failure data tracking and analysis.
- Verification and validation.
- Reliability improvements.

For each of these eight categories the standard defines the inputs, the required activities and the expected outputs. Each of the reliability practices are individually assessed (levels 1–5) with reference to the specified set of activities required to obtain a specific capability level. These five levels represent the metrics or measures of the organizational reliability capability and reflect stages in the evolutionary transition of that practice.

Five reliability capability levels of IEEE Std 1624 can be associated with five reliability maturity stages discussed in Silverman (2010). Those levels are: 1-Uncertainty; 2-Awakening, 3-Enlightenment, 4-Wisdom, 5-Certainty, which verbally characterize the levels of reliability maturity of an organization.

Although IEEE Std 1624 was developed for the electronics industry, it can be used for self-assessment by organizations or for supplier/customer relationship development in virtually any industry with no or minor adjustments.

17.15.2 Reliability Maturity

In 2004 the Automotive Industry Action Group (AIAG) published Reliability Methods Guideline (AIAG, 2004) containing 45 key reliability tools, some of which are covered in this chapter and the rest are in the other chapters. This activity was later expanded to develop an organizational capabilities maturity concept. The AIAG workgroup produced a reliability maturity assessment (RMA) manual that contains nine reliability categories:

A Reliability planning.
B Design for reliability.
C Reliability prediction and modelling.
D Reliability of mechanical components and systems.
E Statistical concepts.
F Failure reporting and analysis.
G Analysing reliability data.
H Reliability testing.
I Reliability in manufacturing.

Although the RMA category B is called ‘Design for reliability’, it contains only a subset of reliability tools discussed in Chapter 7. Each of the nine categories above included the appropriate set of reliability tools from ‘Reliability Methods Guideline’. For each reliability tool within the category the RMA suggested scoring criteria.

After assessment of the individual scores they are combined by categories resulting in a rating based on the percentage of the maximum available score for each category. The scores for each category can be combined based on the weighted averages to obtain the total score for an organization. Score above 60 % is classified as B-level and above 80 % is A-level. Scores below 60 % are considered as reliability deficiencies.

Both reliability maturity and capability assessments provide important tools to evaluate the organizational capability from a product reliability perspective. The assessments can also be used for supplier selection process and can be conducted as self-assessment and/or 2nd or 3rd party assessment. These activities also help to identify gaps and weaknesses in the reliability process and can be instrumental in developing an efficient product and process improvement plan.

17.16 Managing Production Quality

The production department should have ultimate responsibility for the manufacturing quality of the product. It is often said that quality cannot be 'inspected in' or 'tested in' to a product. The QA department is responsible for assessing the quality of production but not for the operations which determine quality. QA thus has the same relationship to production as reliability engineering has to design and development.

In much modern production, inspection and test operations have become integrated with production operations. For example, in operator control of computerized machine tools the machine operator might carry out workpiece gauging and machine calibration. Also, the costs of inspection and test can be considered to be production costs, particularly when it is not practicable to separate the functions or when, as in electronics production, the test policy can have a great impact on production costs. For these reasons, there is a trend towards routine inspection and test work being made the responsibility of production, with QA providing support services such as patrol inspection, and possibly final inspection, as well as training, calibration, and so on. Determination of inspection and test policy methods and staffing should then be primarily a production responsibility, with QA in a supporting role, providing advice and ensuring that quality standards will be met.

This approach to modern quality results in much smaller QA departments than under the older system whereby production produced and passed the products to QA for inspection and test at each stage. It also obviously reduces the total cost of production (production cost plus inspection and test costs). Motivation for quality is enhanced and QA staff are better placed to contribute positively, rather than acting primarily in a policing role. The quality circles movement has also heavily influenced this trend; quality circles could not operate effectively under the old approach.

The QA department should be responsible for:

1 Setting production quality criteria.
2 Monitoring production quality performance and costs.
3 QA training (SPC, motivational, etc.).
4 Specialist facilities and services.
5 Quality audit and registration.

These will be discussed in turn.

17.16.1 Setting Production Quality Criteria

The quality manager must decide the production quality criteria (such as tolerances, yields, etc.) to be met. These might have been set by the customer, as is often the case in commercial as well as in defence equipment manufacture, in which case the quality manager is the interface with the customer on production quality matters. Quality criteria apply to the finished product, to production processes and to bought-in materials and components. Therefore the quality manager should determine, or approve, the final inspection and test methods and criteria to ensure conformance. He or she should also determine such details as quality levels of components, quality control of suppliers and calibration requirements for test and measuring equipment.

17.16.2 Monitoring Production Quality Performance and Costs

The quality manager must be satisfied that the quality objectives are being attained or that action is being taken to ensure this. These include quality cost objectives, as described earlier. QA staff should therefore oversee and monitor functions such as failure reporting and final conformance inspection and test. The QA department should prepare or approve quality performance and cost reports, and should monitor and assist with problem-solving. The methods described in Chapters 11 and 15 are particularly appropriate for this task.

17.16.3 Quality Training

The quality manager is responsible for all quality control training. This is particularly important in training for operator control and quality circles, since all production people must understand and apply basic quality concepts such as simple SPC and data analysis.

17.16.4 Specialist Facilities and Services

The quality department provides facilities such as calibration services and records, vendor appraisal, component and material assessment, and data collection and analysis.

The assessment facilities used for testing components and materials so that their use can be approved are also the best services to use for failure investigation, since this makes the optimum use of expensive resources such as spectroscopic analysis equipment and scanning electron microscopes, and the associated specialist staff.

The joint use of these services in support of development, manufacturing and in-service is best achieved by operating an integrated approach to quality and reliability engineering.

17.16.5 Quality Audit and Registration

Quality audit is an independent appraisal of all of the operations, processes and management activities that can affect the quality of a product or service. The objective is to ensure that procedures are effective, that they are understood and that they are being followed.

Quality audit, like financial audit, requires both internal and external audit. Internal audit is a continuing function whereby QA staff review the operations and controls, and report on discrepancies. External audit is performed by the third party assessors on a regular schedule, typically annually, in order to obtain and retain registration to quality standards, in particular ISO9000.

An effective quality audit should include review of all design, development and production, test and inspection operations, as well as associated procedures and documentation. An important aspect is the assurance that personnel know and understand their role in the quality system, including relevant procedures and responsibilities.

The quality manager's responsibility for audit includes all internal audit, for ensuring that the company is successful in customer or third party audits and for quality auditing of suppliers. Preparation for external audit and being subjected to it can be a very important task, and much effort is involved. The quality department should be skilled in undertaking this responsibility with minimum disruption to normal design and production work. This demands thorough knowledge of the appropriate standards and methods and the ways in which they are applied. Training is an important feature to ensure that personnel will respond correctly during audit.

Reliability aspects are included in quality audit, since there is so much common ground, particularly in relation to failure reporting and corrective action. Ideally, internal audits should be carried out by the managers responsible for the procedures and tasks concerned, or by other knowledgeable staff on their behalf. The audits should have the objective of improving processes, not merely of determining compliance.

Rothery (1996) and Rabbitt and Bergh (1994) are amongst the many books on quality auditing to ISO9000.

17.17 Quality Management Approaches

17.17.1 Quality Systems

The quality systems approach is epitomized by ISO9000, described earlier. It is based on the premise that if the 'system' is described and followed, then the output (products, services) will be of high quality. A good quality system (effective procedures, training, etc.) is necessary, but it can provide only a baseline for achieving high quality and reliability. The quality systems approach based on ISO9000 or similar standards is the most widely-applied approach in most engineering business.

17.17.2 'Six Sigma'

The 'six sigma' approach was developed by the Motorola company in 1986. A six sigma process should operate within 6σ limits, which implies that 99.999 66 % of the products manufactured are statistically expected to be free of defects (3.4 defects per million). This in turn implies that every process follows the normal distribution. As explained in Chapter 2, this is not always the case. The six sigma process involves application of statistical and other methods to identify opportunities for process improvements. To this end specialists are trained in the methods, gaining credentials that culminate in 'black belt' and 'master black belt' status. They are given the task of finding opportunities, driving the improvements and training the organization in six sigma methods. Top management involvement and leadership is essential. The six sigma approach is credited with generating very large savings in several big companies, although its use is not without controversy.

The six sigma methods, including 'lean' six sigma and design for six sigma (DFSS) continue to be used to improve engineering practices, facilitate product improvement and generate cost saving. There is a multitude of references on the six sigma methods including books, periodic publications, various websites, forums, blogs, and so on.

17.17.3 Quality Circles

The quality circles approach to improving production quality was described in Chapter 15. It is an inherent part of the total quality management (TQM) approach described in Section 17.17.5.

17.17.4 Quality Awards

The first national quality awards to be introduced were the Deming Awards in Japan. These are presented every year to individuals, groups and companies which achieve notable quality levels or improvements. Later the idea was followed in the USA with the Baldrige Awards, named after the then Secretary for Trade. US companies submit themselves to the assessment process, which is conducted by independent assessors, to determine whether they achieve the highest scores in a range of company categories. The award winners gain good publicity in the years that they win, and they can also generate improvements as a result of their efforts to impress the assessors.

In Europe the European Foundation of Quality Management (EFQM) has produced a self-assessment guide, which companies can use to conduct their own evaluations. The EFQM Excellence Award is given to companies or other organizations that show outstanding results across a wide spectrum, not only quality and reliability.

17.17.5 Total Quality Management

The terms *total quality management* (TQM) (or *total quality control* (TQC)) are often used to describe a system whereby all the activities that contribute to product quality, not just production quality control, are appraised and controlled by one manager. In this context quality is defined as the totality of features which determine a product's acceptability, and as such includes appearance, performance, reliability, support, and so on.

Under this concept the quality manager has very wide authority for setting and monitoring quality standards, in this wide sense, throughout all functions of the organization. The quality manager then reports directly to the chief executive. It remains essential for line functions such as design, test and production to retain responsibility for their contributions to quality and reliability. However, the quality manager is responsible for ensuring that the total approach is coordinated, through the setting of standards, training and performance monitoring.

The TQM approach to reliability and quality can be very effective, particularly when applied to correct a situation in which quality is perceived as being lower than is required, but the reasons cannot be pinned down.

It is not easy to find people who can effectively fulfil the total quality management role. The task demands rather exceptional talents of persuasiveness and ability. It is easy for the quality manager and the organization to become dissociated from corporate realities, and the authority of the quality manager might be questioned by line departments and project managers.

The obvious solution to this is for the chief executive to personally 'drive' the TQM effort, by ensuring that all functions support it and by monitoring the results. This has the supreme advantage of showing that quality and reliability are of top level concern. Functions such as design reliability and production quality control can then be integrated with design and production, and coordination of standards, training, and so on, can be achieved through a chief executive's quality committee.

Only the chief executive can ensure total integration of the quality and reliability functions with the management of specifying, designing, producing and supporting the product. The increasing integration of design and production, and the pace and competitiveness of modern markets for technology-based products, demand that a fully integrated approach be used. This is to be found in many of the modern high technology companies that have grown up in the last 30 years, and in those older companies that have perceived quality and reliability as being matters too important to be left to chance or to lower levels of management. Their success has been largely due to this recognition, and to the commitment and involvement of the most senior executives.

Bergman and Klefsjo (1994) and O'Connor (2004) describe the management principles and methods of TQM.

17.18 Choosing the Methods: Strategy and Tactics

We must apply the methods that are known to be effective in creating reliable products. We should also try to avoid methods that are misleading or counter-productive. We can consider management methods to be either strategic or tactical. Strategic principles and methods prepare the organization for the work that must be done in the future. Tactical methods are applied when necessary.

The important strategic principles and methods are:

- Top management commitment, through all functions and product phases.
- Effective organization and people.
- Effective methods, capabilities and procedures.
- Supplier selection and development.
- Training, covering management, technology, the tactical methods and consistent with the organization and procedures.
- Research.

The important tactical methods are:

- Design for reliability (DfR) (QFD, FMECA, stress, variation, etc., and effective use of CAE tools) (Chapters 5, 7–10).
- Test, with the emphasis on HALT (Chapter 12) and design of experiments (Chapter 11).
- FRACAS (Chapters 12–16).
- Production quality control and improvement (Chapter 15).
- Maintenance planning and methods (when appropriate) (Chapter 16).

The strategic methods must be applied over the long term. The timescales for effective payback are typically one to five years, and the benefits continue to grow. However, without them the application of the tactical methods will not be as effective. Therefore they must be managed appropriately, and protected from short-term cost-cutting.

The tactical methods must be applied continuously on product design, development, production and in service.

17.19 Conclusions

In the years since the 1980s there has been a steady improvement in the reliability of engineering products and systems, despite increasingly high levels of complexity. This has been the result of spreading recognition that reliability is a key factor in competitive markets, and the realization that the costs of achieving high reliability are amply repaid by lower costs of failures and enhanced product reputation, even at reliability levels well in excess of previous achievement (as explained in Chapter 1). Companies across the spectrum of engineering have taken up the challenge and applied the tools described to improve the reliability of their products. As a result we have all benefitted.

Public safety consciousness has been a powerful spur to quality and reliability in some areas, such as nuclear power and air travel, where government agencies set up and monitor safety standards. Minimising product liability risks has become a major factor in design and development of a wide range of products. Reliability is obviously an essential contributor to safety, so the disciplines described in this book are an important part of the safety assurance process.

There has also been a continuing drive for high reliability in military products, and several of the formal methods for design reliability improvement started in the military in the United States, so that the US military standards on reliability analysis and methods are used in many non-military areas. Whilst the commercial world adopted many of the techniques developed by the military, there has been a swing the other way, with military buyers now applying commercial-type methods and warranties.

The great advances in computer-aided engineering have empowered designers to create and analyse better designs, so that errors are reduced and the transition to manufacturing is made easier and more trouble-free. Much of the improvement in reliability has also been due to improved quality of manufacturing operations. Automation has contributed to this, as well as other advances in manufacturing technology such as more precise machining and advances in measurement and test.

However, improvements in human performance as applied to engineering have probably contributed the most, as the principles taught by Drucker (Drucker, 1955), Deming (Deming, 1987) and other leaders have been more widely applied. Ultimately, the drive to high reliability can only come from management. Drucker did not address engineering specifically, but his philosophy is highly appropriate to engineering design, development, production and support. The application of the 'new management' to engineering is described in Clausing (1994) and O'Connor (2004).

Most successful engineering companies and organizations now accept and apply the philosophy, but there is sometimes a tendency to allow 'scientific' ideas to re-surface, inhibiting freedom and initiative. Also, since the main benefits of high reliability accrue downstream in the product life cycle, there can be a temptation to seek short-term savings by reducing effort on design analysis, test and other DfR methods. The methods we have described, and the ways in which they should be integrated within the whole product life cycle, represent the best practice of the most successful companies in today's engineering industry.

Questions

1. What are the main elements of an integrated reliability programme?
2. a Describe briefly the main cost headings associated with achieving high quality and reliability, and the main consequential costs of failure, in development, production and use.
   b Explain and discuss Deming's philosophy of overall quality and reliability cost minimization.
3. a What are the most important aspects to be considered in preparing reliability specifications?
   b Write an outline reliability specification for (i) a domestic TV set, (ii) a fighter aircraft, and (iii) a gearbox bearing.
4. Discuss the ways in which reliability can be covered in procurement contracts for complex systems.
5. What are the main elements of a project reliability plan? To which other project plans should it refer?
6. What is meant by 'total quality management'? How does the concept differ from the requirements of the international standard for quality systems (ISO9000), and how does it affect reliability?
7. Your firm designs, develops and manufactures a complex consumer product which sells into a highly competitive market. The firm has recently been losing its market share and this is thought to be due to an increasing reputation for unreliable products.

   You are currently developing a new product scheduled for volume production in about 18 months' time. This product includes several new technological features and is seen very much as a 'make-or-break' product as far as the firm's future survival is concerned. The design concept is 'frozen', but little development work has taken place.

   Outline the procedures you would adopt in development and subsequent volume production to ensure the retrieval of your firm's previous reputation for high reliability.
8. Review the nine reliability maturity categories in Section 17.15.2. In your opinion, are they equally important for an organization? How would you assign weights to those categories? Explain your choice.
9. What subject material you would include in the training of a reliability engineer? What university level courses would you consider for the Reliability Engineering program? Justify your choices?

Bibliography

Abbott, H. and Tyler, M. (1997) *Safer by Design: a Guide to the Management and Law of Designing for Product Safety*, Gower.

AIAG (2004) *Reliability Methods Guideline*. Developed by the Truck & Heavy Equipment Reliability Workgroup (THE-7). Available at http://www.aiag.org/staticcontent/committees/workgroup.cfm?workgroup=FTRM.

Barringer, P. (2011) *Reliability Standards*. Available at: http://www.barringer1.com/rs.htm.

Bergman, B. and Klefsjo, B. (1994) *Quality: from Customer Needs to Customer Satisfaction*, McGraw-Hill.

Breyfogle, F., Cupello, J. and Meadows, B. (2000) *Managing Six Sigma*, Wiley.

British Standard, BS 5760. *Reliability of Systems, Equipments and Components*. British Standards Institution, London.

Clausing, D. (1994) *Total Quality Development*, ASME Press.

Conti, T. (1993) *Building Total Quality*, Chapman and Hall.

Defence Standard 00–40: *The Management of Reliability and Maintainability*. HMSO UK.

Deming, W.E. (1987) *Out of the Crisis*, MIT Press.

Drucker, P. (1955) *The Practice of Management*, Heinemann.

Edosomwan, J. (1987) *Integrating Quality and Productivity Management*, Dekker.

GMW 3172 (2004) *'General Specification for Electrical/Electronic Component Analytical/Development/Validation (A/D/V) Procedures for Conformance to Vehicle Environmental, Reliability, and Performance Requirements'*, General Motors Worldwide Engineering standard. Available at www.global.ihs.com (Accessed 20 March 2011).

Hoyle, D. (2009) *ISO 9000 Quality Systems Handbook* (6th edition). Elsevier.

Hutchins, D. (1990) *In Pursuit of Quality*, Pitman.

IEC61508, *Functional Safety of Electrical/Electronic/Programmable Electronic Safety-related Systems*. International Standards Organisation.

IEEE (2008) Std 1624 IEEE Standard for Organizational Reliability Capability.

ISO/IEC60300, *Dependability Management*. International Standards Organization, Geneva.

ISO/IEC61508, *Functional Safety of Electrical/Electronic/Programmable Electronic Safety-related Systems*. International Standards Organization, Geneva.

ISO9000 (2000) *Quality Systems*. International Standards Organization, Geneva.

Kleyner, A. and Sandborn, P. (2008) *Minimizing Life Cycle Cost by Managing Product Reliability via Validation Plan and Warranty Return Cost*. *International Journal of Production Economics (IJPE)*, **112**, 796–807.

O'Connor, P.D.T. (2004) *The New Management of Engineering*. Available at www.lulu.com.

Pyzdek, T. (2001) *The Six Sigma Handbook*, McGraw-Hill.

Rabbitt, J. and Bergh, P. (1994) *The ISO 9000 Book: A Global Competitor's Guide to Compliance and Certification*, Amacom.

Rothery, B. (1996) *Standards and Certification in Europe*, Gower.

Silverman, M. (2010) *How Reliable is Your Product? 50 Ways to Improve Product Reliability*, Super Star Press, Silicon Valley, California.

Thomas, B. (1995) *The Human Dimension of Quality*, McGraw-Hill.

Tiku, S., Azarian, M. and Pecht, M. (2007) *Using a Reliability Capability Maturity Model to Benchmark Electronics Companies*, International Journal of Quality & Reliability Management, **24**(5), 547–563.

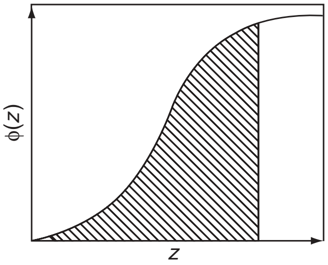
Tiku, S. and Pecht, M. (2010) *Validation of Reliability Capability Evaluation Model Using a Quantitative Assessment Process*. International Journal of Quality & Reliability Management, **27**(8), 938–952.

US MIL-STD-785. *Reliability Programs for Systems and Equipment—Development and Production*. Available from the National Technical Information Service, Springfield, Virginia.

Warranty Week (2011) The Newsletter for Warranty Management Professionals (online). Available at: http://www.warrantyweek.com/.

Appendix 1

The Standard Cumulative Normal Distribution Function



$$\Phi(z) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{z} \exp\left(\frac{-x^2}{2}\right) dx$$

for $0.00 \leq z \leq 4.00$

$$1 - \Phi(z) = \Phi(-z)$$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6985	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319

Practical Reliability Engineering, Fifth Edition. Patrick D. T. O'Connor and Andre Kleyner.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9430	0.9440
1.6	0.9452	0.9463	0.9474	0.9485	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9700	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9762	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	$0.9^{2}010$	$0.9^{2}061$	$0.9^{2}035$	$0.9^{2}086$	$0.9^{2}111$	$0.9^{2}134$	$0.9^{2}158$
2.4	$0.9^{2}180$	$0.9^{2}202$	$0.9^{2}224$	$0.9^{2}245$	$0.9^{2}266$	$0.9^{2}286$	$0.9^{2}305$	$0.9^{2}324$	$0.9^{2}343$	$0.9^{2}361$
2.5	$0.9^{2}379$	$0.9^{2}396$	$0.9^{2}413$	$0.9^{2}430$	$0.9^{2}446$	$0.9^{2}461$	$0.9^{2}477$	$0.9^{2}492$	$0.9^{2}506$	$0.9^{2}520$
2.6	$0.9^{2}534$	$0.9^{2}547$	$0.9^{2}560$	$0.9^{2}573$	$0.9^{2}586$	$0.9^{2}598$	$0.9^{2}609$	$0.9^{2}621$	$0.9^{2}632$	$0.9^{2}643$
2.7	$0.9^{2}653$	$0.9^{2}664$	$0.9^{2}674$	$0.9^{2}683$	$0.9^{2}693$	$0.9^{2}702$	$0.9^{2}711$	$0.9^{2}720$	$0.9^{2}728$	$0.9^{2}737$
2.8	$0.9^{2}745$	$0.9^{2}752$	$0.9^{2}760$	$0.9^{2}767$	$0.9^{2}774$	$0.9^{2}781$	$0.9^{2}788$	$0.9^{2}795$	$0.9^{2}801$	$0.9^{2}807$
2.9	$0.9^{2}813$	$0.9^{2}819$	$0.9^{2}825$	$0.9^{2}831$	$0.9^{2}836$	$0.9^{2}841$	$0.9^{2}846$	$0.9^{2}851$	$0.9^{2}856$	$0.9^{2}861$
3.0	$0.9^{2}865$	$0.9^{2}869$	$0.9^{2}874$	$0.9^{2}878$	$0.9^{2}882$	$0.9^{2}886$	$0.9^{2}889$	$0.9^{2}893$	$0.9^{2}897$	$0.9^{2}900$
3.1	$0.9^{3}032$	$0.9^{3}065$	$0.9^{3}096$	$0.9^{3}126$	$0.9^{3}155$	$0.9^{3}184$	$0.9^{3}211$	$0.9^{3}238$	$0.9^{3}264$	$0.9^{3}289$
3.2	$0.9^{3}313$	$0.9^{3}336$	$0.9^{3}359$	$0.9^{3}381$	$0.9^{3}402$	$0.9^{3}423$	$0.9^{3}443$	$0.9^{3}462$	$0.9^{3}481$	$0.9^{3}499$
3.3	$0.9^{3}517$	$0.9^{3}534$	$0.9^{3}550$	$0.9^{3}566$	$0.9^{3}581$	$0.9^{3}596$	$0.9^{3}610$	$0.9^{3}624$	$0.9^{3}638$	$0.9^{3}651$
3.4	$0.9^{3}663$	$0.9^{3}675$	$0.9^{3}687$	$0.9^{3}698$	$0.9^{3}709$	$0.9^{3}720$	$0.9^{3}730$	$0.9^{3}740$	$0.9^{3}749$	$0.9^{3}759$
3.5	$0.9^{3}767$	$0.9^{3}776$	$0.9^{3}784$	$0.9^{3}792$	$0.9^{3}800$	$0.9^{3}807$	$0.9^{3}815$	$0.9^{3}822$	$0.9^{3}822$	$0.9^{3}835$
3.6	$0.9^{3}841$	$0.9^{3}847$	$0.9^{3}853$	$0.9^{3}858$	$0.9^{3}864$	$0.9^{3}869$	$0.9^{3}874$	$0.9^{3}879$	$0.9^{3}883$	$0.9^{3}888$
3.7	$0.9^{3}892$	$0.9^{3}896$	$0.9^{4}004$	$0.9^{4}043$	$0.9^{4}116$	$0.9^{4}116$	$0.9^{4}150$	$0.9^{4}184$	$0.9^{4}216$	$0.9^{4}257$
3.8	$0.9^{4}277$	$0.9^{4}305$	$0.9^{4}333$	$0.9^{4}359$	$0.9^{4}385$	$0.9^{4}409$	$0.9^{4}433$	$0.9^{4}456$	$0.9^{4}478$	$0.9^{4}499$
3.9	$0.9^{4}519$	$0.9^{4}539$	$0.9^{4}557$	$0.9^{4}575$	$0.9^{4}593$	$0.9^{4}609$	$0.9^{4}625$	$0.9^{4}641$	$0.9^{4}655$	$0.9^{4}670$

Appendix 2

$\chi^2(\alpha, \nu)$ Distribution Values

Degrees of freedom	α(risk factor) = 1 − Confidence							
ν	0.995	0.990	0.975	0.95	0.90	0.80	0.70	0.60
1	0.0^4393	0.0^3157	0.03982	0.0^2393	0.0158	0.0642	0.148	0.275
2	0.0100	0.0201	0.0506	0.103	0.211	0.446	0.713	1.02
3	0.0717	0.115	0.216	0.352	0.584	1.00	1.42	1.87
4	0.207	0.297	0.484	0.711	1.06	1.65	2.19	2.75
5	0.412	0.554	0.831	1.15	1.61	2.34	3.00	3.66
6	0.676	0.872	1.24	1.64	2.20	3.07	3.83	4.57
7	0.989	1.24	1.69	2.17	2.83	3.82	4.67	5.49
8	1.34	1.65	2.18	2.73	3.49	4.59	5.53	6.42
9	1.73	2.09	2.70	3.33	4.17	5.38	6.39	7.36
10	2.16	2.56	3.25	3.94	4.87	6.18	7.27	8.30
11	2.60	3.05	3.82	4.57	5.58	6.99	8.15	9.24
12	3.07	3.57	4.40	5.23	6.30	7.81	9.03	10.2
13	3.57	4.11	5.01	5.89	7.04	8.63	9.93	11.1
14	4.07	4.66	5.63	6.57	7.79	9.47	10.8	12.1
15	4.60	5.23	6.26	7.26	8.55	10.3	11.7	13.0
16	5.14	5.81	6.91	7.96	9.31	11.2	12.6	14.0
17	5.70	6.41	7.56	8.67	10.1	12.0	13.5	14.9
18	6.26	7.01	8.23	9.39	10.9	12.9	14.4	15.9
19	6.84	7.63	8.91	10.1	11.7	13.7	15.4	16.9
20	7.43	8.26	9.59	10.9	12.4	14.6	16.3	17.8
21	8.03	8.90	10.3	11.6	13.2	15.4	17.2	18.8
22	8.64	9.54	11.0	12.3	14.0	16.3	18.1	19.7
23	9.26	10.2	11.7	13.1	14.8	17.2	19.0	20.7
24	9.89	10.9	12.4	13.8	15.7	18.1	19.9	21.7
25	10.5	11.5	13.1	14.6	16.5	18.9	20.9	22.6
26	11.2	12.2	13.8	15.4	17.3	19.8	21.8	23.6
27	11.8	12.9	14.6	16.2	18.1	20.7	22.7	24.5
28	12.5	13.6	15.3	16.9	18.9	21.6	23.6	25.5
29	13.1	14.3	16.0	17.7	19.8	22.5	24.6	26.5
30	13.8	15.0	16.8	18.5	20.6	23.4	25.5	27.4
35	17.2	18.5	20.6	22.5	24.8	27.8	30.2	32.3
40	20.7	22.2	24.4	26.5	29.1	32.3	34.9	37.1
45	24.3	25.9	28.4	30.6	33.4	36.9	39.6	42.0
50	28.0	29.7	32.4	34.8	37.7	41.4	44.3	46.9
75	47.2	49.5	52.9	56.1	59.8	64.5	68.1	71.3
100	67.3	70.1	74.2	77.9	82.4	87.9	92.1	95.8

Practical Reliability Engineering, Fifth Edition. Patrick D. T. O'Connor and Andre Kleyner.

								α	
0.50	0.40	0.30	0.20	0.10	0.05	0.025	0.010	0.005	ν
0.455	0.708	1.07	1.64	2.71	3.84	5.02	6.63	7.88	1
1.39	1.83	2.41	3.22	4.61	5.99	7.38	9.21	10.6	2
2.37	2.95	3.67	4.64	6.25	7.81	9.35	11.3	12.8	3
3.36	4.04	4.88	5.99	7.78	9.49	11.1	13.3	14.9	4
4.35	5.13	6.06	7.29	9.24	11.1	12.8	15.1	16.7	5
5.35	6.21	7.23	8.56	10.6	12.6	14.4	16.8	18.5	6
6.35	7.28	8.38	9.80	12.0	14.1	16.0	18.5	20.3	7
7.34	8.35	9.52	11.0	13.4	15.5	17.5	20.1	22.0	8
8.34	9.41	10.7	12.2	14.7	16.9	19.0	21.7	23.6	9
9.34	10.5	11.8	13.4	16.0	18.3	20.5	23.2	25.2	10
10.3	11.5	12.9	14.6	17.3	19.7	21.9	24.7	26.8	11
11.3	12.6	14.0	15.8	18.5	21.0	23.3	26.2	28.3	12
12.3	13.6	15.1	17.0	19.8	22.4	24.7	27.7	29.8	13
13.3	14.7	16.2	18.2	21.1	23.7	26.1	29.1	31.3	14
14.3	15.7	17.3	19.3	22.3	25.0	27.5	30.6	32.8	15
15.3	16.8	18.4	20.5	23.5	26.3	28.8	32.0	34.3	16
16.3	17.8	19.5	21.6	24.8	27.6	30.2	33.4	35.7	17
17.3	18.9	20.6	22.8	26.0	28.9	31.5	34.8	37.2	18
18.3	19.9	21.7	23.9	27.2	30.1	32.9	36.2	38.6	19
19.3	21.0	22.8	25.0	28.4	31.4	34.2	37.6	40.0	20
20.3	22.0	23.9	26.2	29.6	32.7	35.5	38.9	41.4	21
21.3	23.0	24.9	27.3	30.8	33.9	36.8	40.3	42.8	22
22.3	24.1	26.0	28.4	32.0	35.2	38.1	41.6	44.2	23
23.3	25.1	27.1	29.6	33.2	36.4	39.4	43.0	45.6	24
24.3	26.1	28.2	30.7	34.4	37.7	40.6	44.3	46.9	25
25.3	27.2	29.2	31.8	35.6	38.9	41.9	45.6	48.3	26
26.3	28.2	30.3	32.9	36.7	40.1	43.2	47.0	49.6	27
27.3	29.2	31.4	34.0	37.9	41.3	44.5	48.3	51.0	28
28.3	30.3	32.5	35.1	39.1	42.6	45.7	49.6	52.3	29
29.3	31.3	33.5	36.3	40.3	43.8	47.0	50.9	53.7	30
34.3	36.5	38.9	41.8	46.1	49.8	53.2	57.3	60.3	35
39.3	41.6	44.2	47.3	51.8	55.8	59.3	63.7	66.8	40
44.3	46.8	49.5	52.7	57.5	61.7	65.4	70.0	73.2	45
49.3	51.9	54.7	58.2	63.2	67.5	71.4	76.2	79.5	50
74.3	77.5	80.9	85.1	91.1	96.2	100.8	106.4	110.3	75
99.3	102.9	106.9	111.7	118.5	124.3	129.6	135.6	140.2	100

Appendix 3

Kolmogorov–Smirnov Tables

Critical values, $d_{alpha};(n)^{a}$, of the maximum absolute difference between sample $F_n(x)$ and population $F(x)$ cumulative distribution.

Number of trials, *n*	Level of significance, *α*			
	0.10	0.05	0.02	0.01
1	0.95000	0.97500	0.99000	0.99500
2	0.77639	0.84189	0.90000	0.92929
3	0.63604	0.70760	0.78456	0.82900
4	0.56522	0.62394	0.68887	0.73424
5	0.50945	0.56328	0.62718	0.66853
6	0.46799	0.51926	0.57741	0.61661
7	0.43607	0.48342	0.53844	0.57581
8	0.40962	0.45427	0.50654	0.54179
9	0.38746	0.43001	0.47960	0.51332
10	0.36866	0.40925	0.45662	0.48893
11	0.35242	0.39122	0.43670	0.46770
12	0.33815	0.37543	0.41918	0.44905
13	0.32549	0.36143	0.40362	0.43247
14	0.31417	0.34890	0.38970	0.41762
15	0.30397	0.33760	0.37713	0.40420
16	0.29472	0.32733	0.36571	0.39201
17	0.28627	0.31796	0.35528	0.38086
18	0.27851	0.30936	0.34569	0.37062
19	0.27136	0.30143	0.33685	0.36117
20	0.26473	0.29408	0.32866	0.35241
21	0.25858	0.28724	0.32104	0.34427
22	0.25283	0.28087	0.31394	0.33666
23	0.24746	0.27490	0.30728	0.32954
24	0.24242	0.26931	0.30104	0.32286

Practical Reliability Engineering, Fifth Edition. Patrick D. T. O'Connor and Andre Kleyner.

Critical values, d_{alpha};(n)[a], of the maximum absolute difference between sample $F_n(x)$ and population $F(x)$ cumulative distribution.

Number of trials, n	Level of significance, α			
	0.10	0.05	0.02	0.01
25	0.23768	0.26404	0.29516	0.31657
26	0.23320	0.25907	0.28962	0.31064
27	0.22898	0.25438	0.28438	0.30502
28	0.22497	0.24993	0.27942	0.29971
29	0.22117	0.24571	0.27471	0.29466
30	0.21756	0.24170	0.27023	0.28987
31	0.21412	0.23788	0.26596	0.28530
32	0.21085	0.23424	0.26189	0.28094
33	0.20771	0.23076	0.25801	0.27677
34	0.20472	0.22743	0.25429	0.27279
35	0.20185	0.22425	0.26073	0.26897
36	0.19910	0.22119	0.24732	0.26532
37	0.19646	0.21826	0.24404	0.26180
38	0.19392	0.21544	0.24089	0.25843
39	0.19148	0.21273	0.23786	0.25518
40[b]	0.18913	0.21012	0.23494	0.25205

[a] Values of $d_\alpha(n)$ such that $p(\max)|F^n(x) - F(x)|d^\alpha(n) = \alpha$.

[b] $N > 40 \approx \frac{1.22}{N^{1/2}}, \frac{1.36}{N^{1/2}}, \frac{1.51}{N^{1/2}}$ and $\frac{1.63}{N^{1/2}}$ for the four levels of significance.

Appendix 4

Rank Tables (5 %, 95 %)

5 % RANKS

	SAMPLE SIZE									
j\n	1	2	3	4	5	6	7	8	9	10
1	5.000	2.532	1.695	1.274	1.021	0.851	0.730	0.639	0.568	0.512
2		22.361	13.535	9.761	7.644	6.285	5.337	4.639	4.102	3.677
3			36.840	24.860	18.925	15.316	12.876	11.111	9.775	8.726
4				47.237	34.259	27.134	22.532	19.290	16.875	15.003
5					54.928	41.820	34.126	28.924	25.137	22.244
6						60.696	47.930	40.031	34.494	30.354
7							65.184	52.932	45.036	39.338
8								68.766	57.086	49.310
9									71.687	60.584
10										74.113

5 % RANKS

	SAMPLE SIZE									
j\n	11	12	13	14	15	16	17	18	19	20
1	0.465	0.426	0.394	0.366	0.341	0.320	0.301	0.285	0.270	0.256
2	3.332	3.046	2.805	2.600	2.423	2.268	2.132	2.011	1.903	1.806
3	7.882	7.187	6.605	6.110	5.685	5.315	4.990	4.702	4.446	4.217
4	13.507	12.285	11.267	10.405	9.666	9.025	8.464	7.969	7.529	7.135
5	19.958	18.102	16.566	15.272	14.166	13.211	12.377	11.643	10.991	10.408
6	27.125	24.530	22.395	20.607	19.086	17.777	16.636	15.634	14.747	13.955
7	34.981	31.524	28.705	26.358	24.373	22.669	21.191	19.895	18.750	17.731
8	43.563	39.086	35.480	32.503	29.999	27.860	26.011	24.396	22.972	21.707
9	52.991	47.267	42.738	39.041	35.956	33.337	31.083	29.120	27.395	25.865
10	63.564	56.189	50.535	45.999	42.256	39.101	36.401	34.060	32.009	30.195
11	76.160	66.132	58.990	53.434	48.925	45.165	41.970	39.215	36.811	34.693

Practical Reliability Engineering, Fifth Edition. Patrick D. T. O'Connor and Andre Kleyner.

5 % RANKS (Continued)

	SAMPLE SIZE									
j\n	11	12	13	14	15	16	17	18	19	20
12		77.908	68.366	61.461	56.022	51.560	47.808	44.595	41.806	39.358
13			79.418	70.327	63.656	58.343	53.945	50.217	47.003	44.197
14				80.736	72.060	65.617	60.436	56.112	52.420	49.218
15					81.896	73.604	67.381	62.332	58.088	54.442
16						82.925	74.988	68.974	64.057	59.897
17							83.843	76.234	70.420	65.634
18								84.668	77.363	71.738
19									85.413	78.389
20										86.089

5 % RANKS

	SAMPLE SIZE									
j\n	21	22	23	24	25	26	27	28	29	30
1	0.244	0.233	0.223	0.213	0.205	0.197	0.190	0.183	0.177	0.171
2	1.719	1.640	1.567	1.501	1.440	1.384	1.332	1.284	1.239	1.198
3	4.010	3.822	3.651	3.495	3.352	3.220	3.098	2.985	2.879	2.781
4	6.781	6.460	6.167	5.901	5.656	5.431	5.223	5.031	4.852	4.685
5	9.884	9.411	8.981	8.588	8.229	7.899	7.594	7.311	7.049	6.806
6	13.245	12.603	12.021	11.491	11.006	10.560	10.148	9.768	9.415	9.087
7	16.818	15.994	15.248	14.569	13.947	13.377	12.852	12.367	11.917	11.499
8	20.575	19.556	18.634	17.796	17.030	16.328	15.682	15.085	14.532	14.018
9	24.499	23.272	22.164	21.157	20.238	19.396	18.622	17.908	17.246	16.633
10	28.580	27.131	25.824	24.639	23.559	22.570	21.662	20.824	20.050	19.331
11	32.811	31.126	29.609	28.236	26.985	25.842	24.793	23.827	22.934	22.106
12	37.190	35.254	33.515	31.942	30.513	29.508	28.012	26.911	25.894	24.953
13	41.720	39.516	37.539	35.756	34.139	32.664	31.314	30.072	28.927	27.867
14	46.406	43.913	41.684	39.678	37.862	36.209	34.697	33.309	32.030	30.846
15	51.261	48.454	45.954	43.711	41.684	39.842	38.161	36.620	35.200	33.889
16	56.302	53.151	50.356	47.858	45.607	43.566	41.707	40.004	38.439	36.995
17	61.559	58.020	54.902	52.127	49.636	47.384	45.336	43.464	41.746	40.163
18	67.079	63.091	59.610	56.531	53.779	51.300	49.052	47.002	45.123	43.394
19	72.945	68.409	64.507	61.086	58.048	55.323	52.861	50.621	48.573	46.691
20	79.327	74.053	69.636	65.819	62.459	59.465	56.770	54.327	52.099	50.056
21	86.705	80.188	75.075	70.773	67.039	63.740	60.790	58.127	55.706	53.493
22		87.269	80.980	76.020	71.828	68.176	64.936	62.033	59.403	57.007
23			87.788	81.711	76.896	72.810	69.237	66.060	63.200	60.605
24				88.265	82.388	77.711	73.726	70.231	67.113	64.299
25					88.707	83.017	78.470	74.583	71.168	68.103
26						89.117	83.603	79.179	75.386	72.038
27							89.498	84.149	79.844	76.140
28								89.853	84.661	80.467
29									90.185	85.140
30										90.497

5 % RANKS

	SAMPLE SIZE									
j\n	31	32	33	34	35	36	37	38	39	40
1	0.165	0.160	0.155	0.151	0.146	0.142	0.138	0.135	0.131	0.128
2	1.158	1.122	1.086	1.055	1.025	0.996	0.969	0.943	0.919	0.896
3	2.690	2.604	2.524	2.448	2.377	2.310	2.246	2.186	2.129	2.075
4	4.530	4.384	4.246	4.120	3.999	3.885	3.778	3.676	3.580	3.488
5	6.578	6.365	6.166	5.978	5.802	5.636	5.479	5.331	5.190	5.057
6	8.781	8.495	8.227	7.976	7.739	7.516	7.306	7.107	6.919	6.740
7	11.109	10.745	10.404	10.084	9.783	9.499	9.232	8.979	8.740	8.513
8	13.540	13.093	12.675	12.283	11.914	11.567	11.240	10.931	10.638	10.361
9	16.061	15.528	15.029	14.561	14.122	13.708	13.318	12.950	12.601	12.271
10	18.662	18.038	17.455	16.909	16.396	15.913	15.458	15.028	14.622	14.237
11	21.336	20.618	19.948	19.319	18.730	18.175	17.653	17.160	16.694	16.252
12	24.077	23.262	22.501	21.788	21.119	20.491	19.898	19.340	18.812	18.312
13	26.883	25.966	25.111	24.310	23.560	22.855	22.191	21.565	20.973	20.413
14	29.749	28.727	27.775	26.884	26.049	25.265	24.527	23.832	23.175	22.553
15	32.674	31.544	30.491	29.507	28.585	27.719	26.905	26.138	25.414	24.729
16	35.657	34.415	33.258	32.177	31.165	30.216	29.324	28.483	27.690	26.940
17	38.698	37.339	36.074	34.894	33.789	32.754	31.781	30.865	30.001	29.185
18	41.797	40.317	38.940	37.657	36.457	35.332	34.276	33.283	32.346	31.461
19	44.956	43.349	41.656	40.466	39.167	37.951	36.809	35.736	34.725	33.770
20	48.175	46.436	44.823	43.321	41.920	40.609	39.380	38.224	37.136	36.109
21	51.458	49.581	47.841	46.225	44.717	43.309	41.988	40.748	39.581	38.480
22	54.810	52.786	50.914	49.177	47.560	46.049	44.634	43.307	42.058	40.881
23	58.234	56.055	54.344	52.181	50.448	48.832	47.320	45.902	44.569	43.314
24	64.739	59.314	57.235	55.239	53.385	51.658	50.045	48.534	47.114	45.778
25	65.336	62.810	60.493	58.355	56.374	54.532	52.812	51.204	49.694	48.275
26	69.036	66.313	63.824	61.534	59.416	57.454	55.624	53.914	52.311	50.805
27	72.563	69.916	67.237	64.754	62.523	60.429	58.483	56.666	54.966	53.370
28	76.650	73.640	70.748	68.113	65.695	63.483	61.392	59.463	57.661	55.972
29	81.054	77.518	74.375	71.535	68.944	66.561	64.357	62.309	60.399	58.612
30	85.591	81.606	76.150	75.069	72.282	69.732	67.384	65.209	63.185	61.294
31	90.789	86.015	82.127	78.747	75.728	72.990	70.482	68.168	66.021	64.021
32		91.063	86.415	82.619	79.312	76.352	73.663	71.196	68.916	66.797
33			91.322	86.793	83.085	79.848	76.946	74.304	71.876	69.629
34				91.566	87.150	83.526	80.357	77.510	74.915	72.525
35					91.797	87.488	83.946	80.841	78.048	75.497
36						92.015	87.809	84.344	81.302	78.560
37							92.222	88.115	84.723	81.741
38								92.419	88.405	85.085
39									92.606	88.681
40										92.784

5 % RANKS

	SAMPLE SIZE									
j\n	41	42	43	44	45	46	47	48	49	50
1	0.125	0.122	0.119	0.116	0.114	0.111	0.109	0.107	0.105	0.102
2	0.874	0.853	0.833	0.814	0.795	0.778	0.761	0.745	0.730	0.715
3	2.024	1.975	1.928	1.884	1.842	1.801	1.762	1.725	1.689	1.655
4	3.402	3.319	3.240	3.165	3.093	3.025	2.959	2.897	2.836	2.779
5	4.930	4.810	4.695	4.586	4.481	4.382	4.286	4.195	4.108	4.024
6	6.570	6.409	6.256	6.109	5.969	5.836	5.708	5.586	5.469	5.357
7	8.298	8.093	7.898	7.713	7.536	7.366	7.205	7.050	6.902	6.760
8	10.097	9.847	9.609	9.382	9.166	8.959	8.762	8.573	8.392	8.218
9	11.958	11.660	11.377	11.107	10.850	10.605	10.370	10.146	9.931	9.725
10	13.872	13.525	13.195	12.881	12.582	12.296	12.023	11.762	11.512	11.272
11	15.833	15.436	15.058	14.698	14.355	14.028	13.715	13.416	13.130	12.856
12	17.838	17.389	16.961	16.554	16.166	15.796	15.443	15.105	14.782	14.472
13	19.883	19.379	18.901	18.445	18.012	17.598	17.203	16.825	16.464	16.117
14	21.964	21.406	20.875	20.370	19.889	19.430	18.993	18.574	18.174	17.790
15	24.081	23.466	22.881	22.326	21.796	21.292	20.810	20.350	19.910	19.488
16	26.230	25.557	24.918	24.311	23.732	23.180	22.654	22.151	21.671	21.210
17	28.412	27.679	26.984	26.323	25.694	25.095	24.523	23.977	23.455	22.955
18	30.624	29.831	29.078	28.363	27.683	27.034	26.416	25.825	25.261	24.721
19	32.867	32.011	31.200	30.429	29.696	28.997	28.331	27.696	27.088	26.507
20	35.138	34.219	33.348	32.520	31.733	30.984	30.269	29.588	28.936	28.313
21	37.440	36.455	35.522	34.636	33.794	32.993	32.229	31.500	30.804	30.138
22	39.770	38.719	37.722	36.777	35.879	35.025	34.210	33.434	32.692	31.980
23	42.129	41.009	39.949	38.943	37.987	37.078	36.212	35.387	34.599	33.845
24	44.518	43.328	42.201	41.133	40.118	39.154	38.235	37.360	36.524	35.726
25	46.937	45.674	44.480	43.347	42.273	41.251	40.279	39.353	38.469	37.625
26	49.388	48.050	46.785	45.587	44.451	43.371	42.344	41.366	40.432	39.541
27	51.869	50.454	49.117	47.852	46.652	45.513	44.430	43.398	42.415	41.476
28	54.385	52.889	51.478	50.143	48.878	47.678	46.537	45.451	44.416	43.428
29	56.935	55.356	53.868	52.461	51.129	49.866	48.666	47.524	46.436	45.399
30	59.522	57.857	56.288	54.807	53.406	52.078	50.817	49.618	48.477	47.388
31	62.149	60.393	58.741	57.183	55.710	54.315	52.991	51.734	50.537	49.396
32	64.820	62.968	61.228	59.590	58.042	56.578	55.190	53.871	52.617	51.423
33	67.539	65.585	63.753	62.029	60.404	58.868	57.413	56.032	54.720	53.470
34	70.311	68.248	66.318	64.505	62.798	61.187	59.662	58.217	56.844	55.538
35	73.146	70.963	68.927	67.020	65.227	63.537	61.940	60.427	58.991	57.627
36	76.053	73.738	71.587	69.578	67.694	65.921	64.247	62.664	61.164	59.738
37	79.049	76.584	74.306	72.185	70.203	68.341	66.587	64.931	63.362	61.874
38	82.160	79.517	77.093	74.849	72.759	70.805	68.963	67.228	65.589	64.034
39	85.429	82.561	79.964	77.580	75.370	73.309	71.378	69.561	67.846	66.222
40	88.945	85.759	82.944	80.392	78.046	75.870	73.838	71.932	70.136	68.440
41	92.954	89.196	86.073	83.310	80.802	78.494	76.350	74.347	72.465	70.691
42		93.116	89.437	86.374	83.661	81.196	78.924	76.812	74.836	72.978
43			93.270	89.666	86.662	83.998	81.573	79.337	77.256	75.306
44				93.418	89.887	86.939	84.321	81.936	79.734	77.683
45					93.560	90.098	87.204	84.631	82.285	80.117
46						93.695	90.300	87.459	84.929	82.621
47							93.825	90.494	87.703	85.216
48								93.950	90.681	87.939
49									94.069	90.860
50										94.184

95 % RANKS

	SAMPLE SIZE									
$j\backslash n$	1	2	3	4	5	6	7	8	9	10
1	95.000	77.639	63.160	52.713	45.072	39.304	34.816	31.234	28.313	25.887
2		97.468	86.465	75.139	65.741	58.180	52.070	47.068	42.914	39.416
3			98.305	90.239	81.075	72.866	65.874	59.969	54.964	50.690
4				98.726	92.356	84.684	77.468	71.076	65.506	60.662
5					98.979	93.715	87.124	80.710	74.863	69.646
6						99.149	94.662	88.889	83.125	77.756
7							99.270	95.361	90.225	84.997
8								99.361	95.898	91.274
9									99.432	96.323
10										99.488

95 % RANKS

	SAMPLE SIZE									
$j\backslash n$	11	12	13	14	15	16	17	18	19	20
1	23.840	22.092	20.582	19.264	18.104	17.075	16.157	15.332	14.587	13.911
2	36.436	33.868	31.634	29.673	27.940	26.396	25.012	23.766	22.637	21.611
3	47.009	43.811	41.010	38.539	36.344	34.383	32.619	31.026	29.580	28.262
4	56.437	52.733	49.465	46.566	43.978	41.657	39.564	37.668	35.943	34.366
5	65.019	60.914	57.262	54.000	51.075	48.440	46.055	43.888	41.912	40.103
6	72.875	68.476	64.520	60.928	57.744	54.835	52.192	49.783	47.580	45.558
7	80.042	75.470	71.295	67.497	64.043	60.899	58.029	55.404	52.997	50.782
8	86.492	81.898	77.604	73.641	70.001	66.663	63.599	60.784	58.194	55.803
9	92.118	87.715	83.434	79.393	75.627	72.140	68.917	65.940	63.188	60.641
10	96.668	92.813	88.733	84.728	80.913	77.331	73.989	70.880	67.991	65.307
11	99.535	96.954	93.395	89.595	85.834	82.223	78.809	75.604	72.605	69.805
12		99.573	97.195	93.890	90.334	86.789	83.364	80.105	77.028	74.135
13			99.606	97.400	94.315	90.975	87.623	84.366	81.250	78.293
14				99.634	97.577	94.685	91.535	88.357	85.253	82.269
15					99.659	97.732	95.010	92.030	89.009	86.045
16						99.680	97.868	95.297	92.471	89.592
17							99.699	97.989	95.553	92.865
18								99.715	98.097	95.783
19									99.730	98.193
20										99.744

95 % RANKS

	SAMPLE SIZE									
j\n	21	22	23	24	25	26	27	28	29	30
1	13.295	12.731	12.212	11.735	11.293	10.883	10.502	10.147	9.814	9.503
2	20.673	19.812	19.020	18.289	17.612	16.983	16.397	15.851	15.339	14.860
3	27.055	25.947	24.925	23.980	23.104	22.289	21.530	20.821	20.156	19.533
4	32.921	31.591	30.364	29.227	28.172	27.190	26.274	25.417	24.614	23.860
5	38.441	36.909	35.193	34.181	32.961	31.824	30.763	29.769	28.837	27.962
6	43.698	41.980	40.390	38.914	37.541	36.260	35.062	33.940	32.887	31.897
7	48.739	46.849	45.097	43.469	41.952	40.535	39.210	37.967	36.800	35.701
8	53.954	51.546	49.643	47.873	46.221	44.677	43.230	41.873	40.597	39.395
9	58.280	56.087	54.046	52.142	50.364	48.700	47.139	45.673	44.294	42.993
10	62.810	60.484	58.315	56.289	54.393	52.616	50.948	49.379	47.901	46.507
11	67.189	64.746	62.461	60.321	58.316	56.434	54.664	52.998	51.427	49.944
12	71.420	68.874	66.485	64.244	62.138	60.158	58.293	56.536	54.877	53.309
13	75.501	72.869	70.391	68.058	65.861	63.791	61.839	59.996	58.254	56.605
14	79.425	76.728	74.176	71.764	69.487	67.336	65.303	63.380	61.561	59.837
15	83.182	80.444	77.836	75.361	73.015	70.792	68.686	66.691	64.799	63.005
16	86.755	84.006	81.366	78.843	76.441	74.158	71.988	69.927	67.970	66.111
17	90.116	87.397	84.752	82.204	79.762	77.430	75.207	73.089	71.073	69.154
18	93.219	90.589	87.978	85.431	82.970	80.604	78.338	76.173	74.106	72.133
19	95.990	93.540	91.019	88.509	86.052	83.672	81.378	79.176	77.066	75.047
20	98.281	96.178	93.832	91.411	88.994	86.623	84.318	82.092	79.950	77.894
21	99.756	98.360	96.348	94.099	91.771	89.440	87.148	84.915	82.753	80.669
22		99.767	98.433	96.505	94.344	92.101	89.851	87.633	85.468	83.367
23			99.777	98.499	96.648	94.569	92.406	90.232	88.083	85.981
24				99.786	98.560	96.780	94.777	92.689	90.584	88.501
25					99.795	98.616	96.902	94.969	92.951	90.913
26						99.803	98.668	97.015	95.148	93.194
27							99.810	98.716	97.120	95.314
28								99.817	98.761	97.218
29									99.823	98.802
30										99.829

95 % RANKS

	SAMPLE SIZE									
j\n	31	32	33	34	35	36	37	38	39	40
1	9.211	8.937	8.678	8.434	8.203	7.985	7.778	7.581	7.394	7.216
2	14.409	13.985	13.585	13.207	12.850	12.512	12.191	11.885	11.595	11.319
3	18.946	18.394	17.873	17.381	16.915	16.474	16.054	15.656	15.277	14.915
4	23.150	22.482	21.850	21.253	20.688	20.152	19.643	19.159	18.698	18.259
5	27.137	26.360	25.625	24.931	24.272	23.648	23.054	22.490	21.952	21.440
6	30.964	30.084	29.252	28.465	27.718	27.010	26.337	25.696	25.085	24.503
7	34.665	33.687	32.763	31.887	31.056	30.268	29.518	28.804	28.124	27.475
8	38.261	37.190	36.176	35.216	34.305	33.439	32.616	31.832	31.084	30.371
9	41.766	40.606	39.507	38.466	37.477	36.537	35.643	34.791	33.979	33.203
10	45.190	43.945	42.765	41.645	40.582	39.571	38.608	37.691	36.815	35.979
11	48.542	47.214	45.956	44.761	43.626	42.546	41.517	40.537	39.601	38.706
12	51.825	50.419	49.086	47.819	46.615	45.468	44.376	43.334	42.339	41.388
13	55.044	53.564	52.159	50.823	49.552	48.341	47.187	46.086	45.034	44.028
14	58.203	56.651	55.177	53.775	52.440	51.168	49.955	48.796	47.689	46.630
15	61.302	59.683	58.144	56.678	55.282	53.951	52.680	51.466	50.305	49.195
16	64.343	62.661	61.060	59.534	58.080	56.691	55.366	54.098	52.886	51.725
17	67.326	65.585	63.926	62.343	60.833	59.391	58.012	56.693	55.431	54.222
18	70.251	68.456	66.742	65.106	63.543	62.049	60.620	59.252	57.942	56.686
19	73.117	71.272	69.509	67.823	66.210	64.668	63.190	61.776	60.419	59.119
20	75.922	74.034	72.225	70.493	68.835	67.246	65.723	64.264	62.864	61.520
21	78.664	76.738	74.889	73.116	71.415	69.784	68.219	66.717	65.275	63.891
22	81.338	79.382	77.499	75.689	73.951	72.280	70.676	69.135	67.654	66.230
23	83.939	81.961	80.052	78.212	76.440	74.735	73.094	71.517	69.999	68.539
24	86.460	84.472	82.545	80.680	78.881	77.145	75.473	73.862	72.310	70.815
25	88.891	86.907	84.971	83.091	81.270	79.509	77.809	76.168	74.586	73.060
26	91.219	89.255	87.325	85.439	83.604	81.825	80.101	78.435	76.825	75.270
27	93.422	91.505	89.596	87.717	85.878	84.087	82.347	80.660	79.027	77.447
28	95.470	93.635	91.772	89.916	88.086	86.292	84.542	82.840	81.188	79.587
29	97.310	95.615	93.834	92.024	90.217	88.433	86.682	84.972	83.306	81.688
30	98.841	97.396	95.752	94.021	92.261	90.501	88.760	87.050	85.378	83.746
31	99.835	98.878	97.476	95.880	94.198	92.483	90.768	89.069	87.399	85.763
32		99.840	98.912	97.552	96.001	94.364	92.694	91.021	89.362	87.729
33			99.845	98.945	97.623	96.114	94.521	92.893	91.260	89.639
34				99.849	98.975	97.690	96.222	94.669	93.081	91.487
35					99.854	99.004	97.754	96.324	94.810	93.260
36						99.858	99.031	97.814	96.420	94.943
37							99.861	99.057	97.871	96.511
38								99.865	99.081	97.925
39									99.869	99.104
40										99.872

95 % RANKS

	SAMPLE SIZE									
j\n	41	42	43	44	45	46	47	48	49	50
1	7.046	6.884	6.730	6.582	6.440	6.305	6.175	6.050	5.931	5.816
2	11.055	10.804	10.563	10.334	10.113	9.902	9.700	9.506	9.319	9.140
3	14.571	14.241	13.927	13.626	13.338	13.061	12.796	12.541	12.297	12.061
4	17.840	17.439	17.056	16.690	16.339	16.002	15.679	15.369	15.071	14.784
5	20.951	20.483	20.036	19.608	19.198	18.804	18.427	18.064	17.715	17.379
6	23.947	23.416	22.907	22.420	21.954	21.506	21.076	20.663	20.266	19.883
7	26.854	26.262	25.694	25.151	24.630	24.130	23.650	23.188	22.744	22.317
8	29.689	29.037	28.413	27.814	27.241	26.691	26.162	25.623	25.164	24.694
9	32.461	31.752	31.073	30.422	29.797	29.198	28.622	28.068	27.535	27.022
10	35.180	34.415	33.682	32.980	32.306	31.659	31.037	30.439	29.864	29.309
11	37.851	37.032	36.247	35.495	34.773	34.079	33.413	32.772	32.154	31.560
12	40.478	39.607	38.772	37.971	37.202	36.463	35.753	35.069	34.411	33.778
13	43.065	42.143	41.259	40.410	39.596	38.813	38.060	37.336	36.698	35.933
14	45.615	44.644	43.712	42.817	41.958	41.132	40.338	39.573	38.836	38.126
15	48.131	47.110	46.132	45.193	44.290	43.422	42.587	41.783	41.008	40.262
16	50.612	49.546	48.522	47.539	46.594	45.665	44.810	43.968	43.156	42.373
17	53.062	51.950	50.883	49.857	48.871	47.922	47.009	46.129	45.280	44.462
18	55.482	54.326	53.215	52.148	51.122	50.134	49.183	48.266	47.382	46.530
19	57.871	56.672	55.520	54.413	53.348	52.322	51.334	50.382	49.463	48.577
20	60.230	58.991	57.799	56.653	55.549	54.487	53.463	52.476	51.523	50.604
21	62.560	61.281	60.051	58.867	57.727	56.629	55.570	54.549	53.563	52.612
22	64.861	63.545	62.273	61.057	59.882	58.749	57.556	56.602	55.584	54.801
23	67.133	65.781	64.478	63.223	62.013	60.846	59.721	58.634	57.585	56.572
24	69.376	67.989	66.652	65.363	64.121	62.922	61.765	60.647	59.568	58.524
25	71.588	70.169	68.800	67.480	66.205	64.975	63.787	62.640	61.531	60.459
26	73.769	72.320	70.922	69.571	68.267	67.007	65.790	64.613	63.476	62.375
27	75.919	74.443	73.016	71.637	70.304	69.016	67.771	66.566	65.401	64.274
28	78.035	76.534	75.082	73.677	72.317	71.002	69.730	68.500	67.308	66.155
29	80.117	78.594	77.119	75.689	74.306	72.966	71.668	70.412	69.196	68.017
30	82.162	80.621	79.125	77.674	76.268	74.905	73.584	72.304	71.064	69.862
31	84.166	82.611	81.099	79.630	78.203	76.819	75.477	74.175	72.912	71.687
32	86.128	84.564	83.039	81.554	80.111	78.708	77.346	76.023	74.739	73.493
33	88.042	86.475	84.942	83.446	81.988	80.569	79.190	77.848	76.545	75.279
34	89.903	88.340	86.805	85.302	83.834	82.402	81.007	79.650	78.329	77.045
35	91.702	90.153	88.623	87.119	85.645	84.204	82.797	81.426	80.090	78.790
36	93.430	91.907	90.391	88.892	87.418	85.972	84.557	83.175	81.826	80.511
37	95.070	93.591	92.102	90.618	89.150	87.704	86.285	84.895	83.536	82.210
38	96.598	95.190	93.744	92.287	90.834	89.395	87.977	86.584	85.218	83.882
39	97.976	96.681	95.305	93.891	92.464	91.041	89.630	88.238	86.870	85.528
40	99.126	98.025	96.760	95.414	94.030	92.633	91.238	89.854	88.488	87.144
41	99.875	99.147	98.071	96.835	95.518	94.164	92.795	91.427	90.069	88.728
42		99.878	99.167	98.116	96.907	95.618	94.291	92.950	91.608	90.275
43			99.881	99.186	98.158	96.975	95.714	94.414	93.098	91.781
44				99.883	99.205	98.199	97.041	95.805	94.531	93.240
45					99.886	99.222	98.238	97.103	95.892	94.643
46						99.889	99.239	98.275	97.163	95.976
47							99.891	99.255	98.311	97.221
48								99.893	99.270	98.345
49									99.895	99.285
50										99.897

Appendix 5

Failure Reporting, Analysis and Corrective Action System (FRACAS)

ANALYSES

1 Listing of failures:

Failure Report No.	System	Sub-system	Assembly	Part

Output selectable by system, sub-system, assembly, part. Also by time period and other appropriate feature, for example location, modification status.

2 Pareto analysis of top 10–20 failure modes, selectable as above (Chapter 13). See Note 1.
3 MTBF for system, sub-system, assembly. Show number of failures for each MTBF value. Selectable as above.
4 Trend analyses, selectable as above. See Chapter 13.
5 Probability/hazard plots and derivation of distribution parameters. Selected as above. See Chapter 3.

Notes

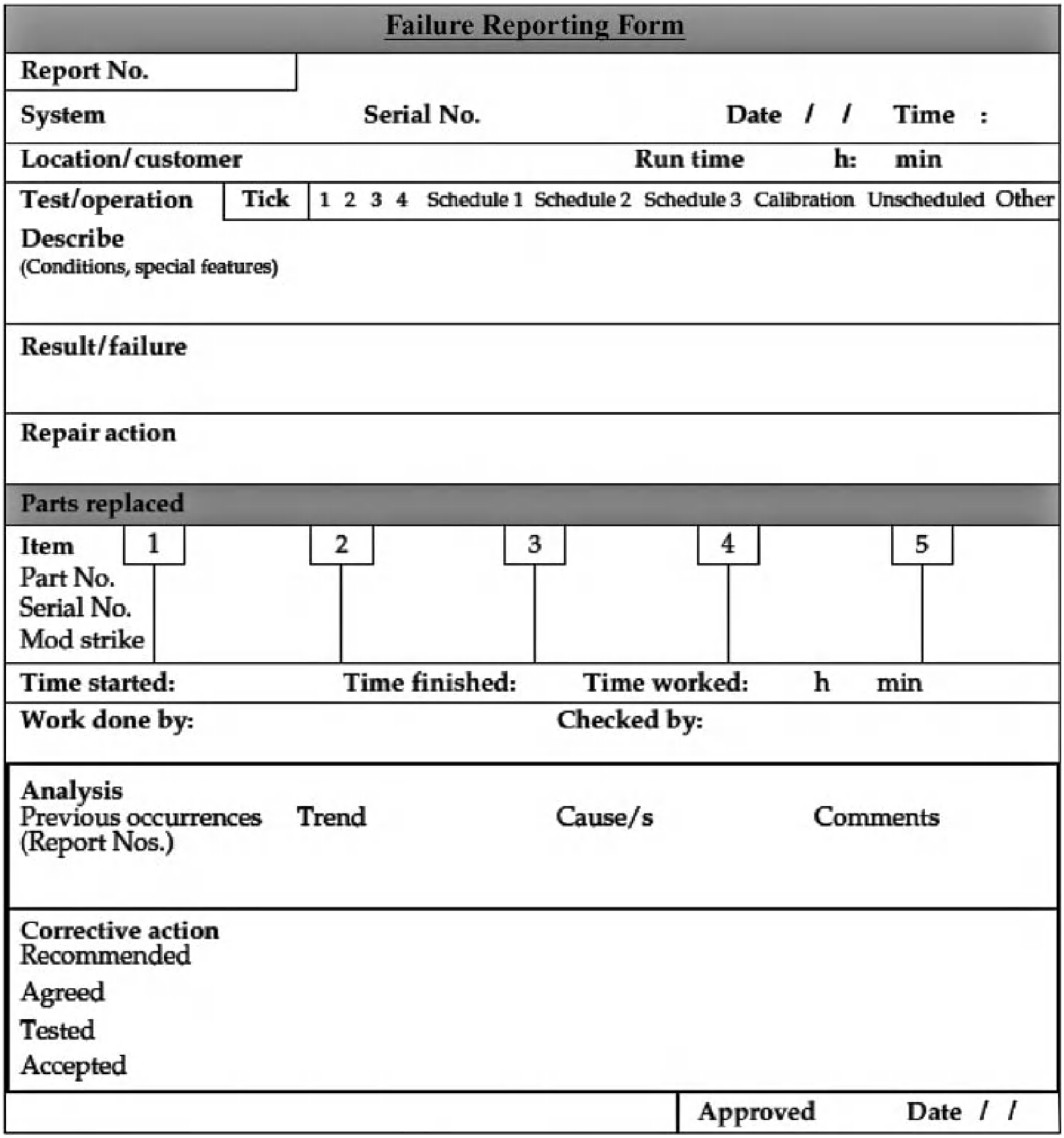
1 The Pareto analysis should form the basis for the critical items list (Section 7.4.7). Failure reporting for critical items should be amplified as necessary to aid investigation, and special reporting, for example by phone, fax or e-mail, could be used.
2 Calendar time could be used instead of run time when appropriate. Then trend analysis, probability plots, and MTBF would be calculated on a calendar time basis. This is acceptable if run time is quite closely correlated with calendar time and if run time data are not easily obtainable, for example if the equipment does not have run time indicators. Using calendar time can be easier and cheaper, since it is then not necessary to obtain the run time at each failure and total (fleet) run time. Only the startup date for each unit and the total number of units in use need be ascertained.

Practical Reliability Engineering, Fifth Edition. Patrick D. T. O'Connor and Andre Kleyner.

3 The FRACAS should be made simple to operate and to understand.

4 Input data should be sufficiently detailed to be meaningful to users. Coding of data such as causes of failure, either by the person filling in the job report or by people at the data centre, can lead to ambiguity and errors.

5 Data input directly into portable computers, rather than onto paper, can greatly improve the quality and speed of data collection and analysis.



Failure Reporting Form

Report No.

System Serial No. Date / / Time :

Location/customer Run time h: min

Test/operation | Tick | 1 2 3 4 Schedule 1 Schedule 2 Schedule 3 Calibration Unscheduled Other

Describe
(Conditions, special features)

Result/failure

Repair action

Parts replaced

Item 1 2 3 4 5
Part No.
Serial No.
Mod strike

Time started: Time finished: Time worked: h min

Work done by: Checked by:

Analysis
Previous occurrences (Report Nos.) Trend Cause/s Comments

Corrective action
Recommended
Agreed
Tested
Accepted

Approved Date / /

Appendix 6

Reliability, Maintainability (and Safety) Plan Example

RELIABILITY AND MAINTAINABILITY (AND SAFETY) PLAN SUPER SYSTEM

CONTENTS

PART 1 RELIABILITY, MAINTAINABILITY (AND SAFETY) PLAN OVERVIEW

PART 2 RELIABILITY AND MAINTAINABILITY ENGINEERING TASKS

PART 3 SAFETY ENGINEERING TASKS

Practical Reliability Engineering, Fifth Edition. Patrick D. T. O'Connor and Andre Kleyner.

1. RELIABILITY, MAINTAINABILITY (AND SAFETY) PLAN OVERVIEW

1.1 Introduction

This Plan describes the organization and responsibilities for the reliability, maintainability (and safety) (RAMS) engineering tasks that will be integrated into the design, development, production and in-service support activities for the Super system project. It also describes the RAMS tasks that will be undertaken.

The RAMS requirements form part of the overall performance requirements for the system, as described in Reference 1.

The RAMS tasks will managed and performed in compliance with the requirements and guidelines in Reference 2.

During the design and development stage, the Company will ensure that their system and equipment suppliers and subcontractors fully understand and comply with the RAMS requirements and with the RAMS engineering tasks specified to them.

In order to achieve these requirements the design, development test, and production philosophy will be for the creation of intrinsically robust, failure-free designs, including the design of all production processes, and followed by stringent production quality assurance and improvement. The failure-free design (FFD) philosophy of hardware and of processes will ensure that all stresses, variations, and other potential or actual causes of failure will be identified and corrected, by the adoption of an integrated, concurrent approach to design, development, and production control. The primary objective of the reliability programme will be to ensure that designs are inherently robust in relation to manufacturing processes and to the environmental conditions of storage, maintenance and operation, throughout the life of the system. To this end, all design analyses and tests will be directed towards identifying and eliminating causes of failure. Particular features of the RAMS programme in this respect will be:

—Prediction and measurement of reliability will be performed as described, but these activities will be treated as secondary to the primary objective of creating an inherently failure-free design.

—The effects of variation of environmental conditions, parameter values and manufacturing processes will be assessed by analysis and by the use of statistically designed tests, including Taguchi methods, to ensure that all designs are robust in relation to all sources of variation over the life of the system.

—The methods of Highly Accelerated Life Testing (HALT) and Highly Accelerated Stress Screening (HASS) will be applied to development and production testing. The objective of these tests will be to force failures by applying high stresses, so that designs and processes can be optimized.

—The reliability test programme will be fully integrated with the overall development test programme. A common failure reporting and corrective action system (FRACAS) will be applied. All failures will be fully investigated and corrective action taken to prevent recurrence. The test programme will include sub-system and system level tests, statistical experiments to assess variation, environmental tests, trials, as well as tests specifically designed to stimulate reliability growth. Reliability growth in development will be monitored in terms of problems discovered and corrected. All test and failure data and corrective action information will be reported, with assessments of reliability achieved and expected in relation to development programme objectives and the in-service requirements.

—By the end of the development programme the Company will have demonstrated that the system and subsystem designs are inherently capable of being produced and tested, and of withstanding the storage, operating and maintenance environments without failure during the in-service life. All relevant failures which occur during development testing will have been corrected by changes to designs or processes, and the effectiveness of the corrective action will have been proved.

—All subcontractors will be required to undertake reliability programmes based upon the same philosophy and methods. The results of their analyses and tests will be closely monitored to ensure a common approach, and to ensure that design improvements and corrective action is implemented promptly and effectively. Subcontractors will be selected on the basis of their proven excellence in the technologies involved, and they will be motivated to adopt the same philosophy for RAMS.

1.2 Reliability, availability, maintainability and safety (RAMS) requirements

The RAMS requirements for the project, related to a service life of are:

—System reliability requirement: not more than *N* failures per, causing.
—System maintainability requirement: repairs to be performed in not more than minutes for % of failures.
—(Safety)

Where responsibility for design and development of a sub- system or component is let by the company to a subcontractor the RAMS requirements for that system, sub-system or equipment will be fully specified by the Project RAMS Engineer in accordance with the relevant Company Procedure.

1.3 RAMS tasks

To ensure achievement and assurance of the RAMS requirements, comprehensive reliability, maintainability (and safety) engineering tasks will be applied. Parts 2 and 3 detail the RAMS engineering tasks that will be undertaken to satisfy the particular requirements in the specification. The tasks will be applied where relevant, from the initial design phase through to, and including, the in-service warranty period.

The RAMS tasks that will be applied at appropriate stages of design and development are:

—Quality function deployment (QFD) to identify and prioritize key design and process requirements.
—Reliability modelling of the system and sub-systems using reliability block diagrams.
—Reliability predictions and apportionment for sub-systems and components, as appropriate.
—Failure analysis by failure mode, effect and criticality analysis (FMECA) and fault tree analysis (FTA).
—Testing, with the emphasis on HALT.
—Failure reporting, analysis and corrective action (FRACAS).
—Maintainability analysis and demonstration.
—System safety analysis by the application of hazard identification and hazard analysis techniques.
—RAMS reviews, in which compliance with RAMS tasks will be audited.
—Production quality assurance and improvement methods.

The Work Plan for these tasks is shown in Appendix 1, and the list of RAMS deliverables is in Appendix 2.

2. RELIABILITY AND MAINTAINABILITY ENGINEERING TASKS

2.1 Reliability modelling

RBDs will be constructed following the guidelines given in. . . . , and will be updated to reflect the state of the design. RBD interfaces will be agreed with and controlled by the Project RAMS Engineer.

2.2 Reliability prediction and apportionment

Reliability prediction will be carried out, covering all areas of design following the guidelines set out in. . . . The methods and data sources will be declared to the Customer.

The reliability prediction process will be started at the commencement of the project, and will be updated to take account of design changes. Predictions will be used to identify high risk components and sub-systems, and for updating reliability apportionments.

2.3 Failure modes, effects and criticality analysis

FMECA will be carried out following the guidelines given in. . . . FMECA software will be used to create and record the analysis. Wherever appropriate computer-aided design (CAD) models and data will be used as inputs and for analysis.

The relevant Design, Quality Assurance and Safety staff will be responsible for formal response to the analyses regarding preventive actions, compensating factors or the effects on safety.

2.4 Fault tree analysis

FTA will be undertaken on equipment designs where safety-related Top Events have been identified. To enable FTA to be undertaken Top Events will be defined and described.

FTA software will be used to perform and record the analysis.

2.5 Reliability testing

An integrated test programme will be conducted, as described in the Project Test Plan (Reference 3). The main features of the test approach will be:

(HALT)
(Taguchi)

2.6 Failure reporting and RAMS monitoring

During the design and development phase of the project the Company will operate a failure reporting, analysis and corrective action system (FRACAS) as described in the Company Quality Manual.

Failure reporting action will be taken on all failures that occur on hardware and software used on the Company and subcontractor tests and trials undertaken during the design and development phase of the project. Failures during production testing will be reported and managed in accordance with the Company Quality Manual.

Reliability and maintainability achievement will be monitored during all development testing.

2.7 *Production reliability assurance*

The Quality Assurance Plan describes the methods that will be applied prior to and during production to ensure that production systems will achieve the reliability requirements. The QA activities will be integrated with the reliability activities wherever appropriate.

Particular features of the QA programme to ensure reliability will be:

—FMECA will be used for the derivation of production functional test and inspection methods. Production FMECA will also be performed.

—Statistical Process Control (SPC) will be applied to all manufacturing processes in which variation can affect yield and reliability. SPC limits will be based where relevant on the results of analysis and test of development hardware, particularly when statistical experiments have been conducted to optimize product and process designs.

—Production stress screening methods (HASS) will be developed as part of the development test programme. HASS will be applied to all production hardware, at sub-system and system level, and will be tailored to provide the optimum screens for the items concerned. HASS profiles and durations will be continuously monitored and modified during production, to ensure the most cost-effective approach. The HASS will provide assurance that all production hardware is function and capable of withstanding the storage and use environments.

—All failures occurring at any test or inspection stage will be investigated, with the objective of preventing recurrence. The objective of the failure reporting and corrective action system (FRACAS), in conjunction with monitoring of SPC, will be to generate continuous improvement of all processes.

—All subcontractors will be required to work to the same philosophy of continuous improvement. Their performance will be closely monitored, and they will be assisted where necessary.

2.8 *Maintainability analysis and demonstration*

Maintainability will be analysed during design to ensure that the requirements are achieved, and will be measured during development. Preventive maintenance tasks will be optimized using the reliability centred maintenance (RCM) method.

2.9 *In-service RAMS monitoring (ISRM)*

The Company will maintain a FRACAS for all systems in service, to monitor RAMS achievement, to ensure that all reliability requirements are achieved, and that any shortcomings or failures are promptly investigated and corrected.

3. SAFETY ENGINEERING TASKS

3.1 *Preliminary hazard analysis*

The Company and subcontractors will apply preliminary hazard analysis (PHA) to all areas of design responsibility in accordance with the Company RAMS Manual and Reference. . . .

The output from the PHA will consist of documented hazards associated with the system.

3.2 System and sub-system hazard analysis

System and sub-system hazard analysis (SHA) will be carried out to identify hazards associated with the system and sub-system design, which may not have been identified in the PHA, including component fault modes, critical human error inputs, and hazards resulting from interfaces within the equipment. The techniques that will be applied will be HAZOPS, FMECA, FTA, and event tree analysis (ETA), as appropriate. They will be applied to all areas of design responsibility in accordance with the Company RAMS Manual.

The output from the SHA will detail system and sub-system hazards, their severity and probability values, together with recommendations for actions necessary to eliminate them, or to control the risk to a level that is agreed to be as low as practicable (ALARP).

3.3 Hazard tracking (Hazard log)

Hazard Tracking will be applied to all conditions which could possibly produce a Catastrophic (Severity Level 1) effect or Hazardous (Severity Level 2) effect, as defined in Reference. . . .

Hazards will be tracked from the point of identification until the hazard is eliminated or the associated risk is reduced to a level agreed with the Customer as being acceptable. The Hazard Tracking system will be maintained after design work is complete and throughout the warranty period.

A Hazard Log will be maintained, containing as a minimum:

—Description of each hazard, by nature, cause, and effect.
—Severity rating.
—Status of actions to resolve.
—Traceability of resolution to the point of risk acceptance.

The Company will ensure that the hazard tracking system is correctly maintained, and will make the records available to the Customer for audit and review.

4. PROJECT RAMS ENGINEERING MANAGEMENT AND REPORTING

4.1 Responsibilities

For the effective management of RAMS engineering formal management procedures and guidelines on analysis techniques will be applied for all tasks outlined in the RAMS Plan The management procedures are contained within the relevant Company Procedures The application of the RAMS Plan will be the responsibility of the Project Reliability and Safety Engineer on behalf of the Project Manager Subcontractors will be required to prepare RAMS Plans that comply with the Project RAMS Plan as appropriate to the sub systems.

4.2 RAMS reviews

The RAMS Review is the formal audit of the RAMS engineering tasks undertaken by the Company and implemented as detailed in the RAMS Work Plan.

RAMS Reviews will be conducted in accordance with the relevant Company Procedure, on all areas of design for which the Company and subcontractors are responsible. The responsibility for conducting RAMS Reviews is assigned to the Company Project Reliability (and Safety) Engineer.

On completion of a RAMS Review, a RAMS Progress Report will be raised by the Company, to highlight areas of non-conformance or risk, and to advise on the extent to which the RAMS requirements are expected to be achieved.

References

1. Super System Specification.
2. (Detailed method guidelines/descriptions: relevant company procedures, standards, etc.).
3. Super System Test Plan.

Appendices

1. RAMS Work Plans.
2. RAMS Deliverables.

Appendix 7

Matrix Algebra Revision

The solution of the second-order matrix

$$\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}$$

is $a_1\, b_2 - a_2\, b_1$.

The solution of the third-order matrix

$$\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix}$$

is

$$a_1 \begin{vmatrix} b_2 & c_2 \\ b_3 & c_3 \end{vmatrix} - b_1 \begin{vmatrix} a_2 & c_2 \\ a_3 & c_3 \end{vmatrix} + c_1 \begin{vmatrix} a_2 & b_2 \\ a_3 & b_3 \end{vmatrix}$$

$a_1, a_2, \ldots, c_2, c_3$ are called the *elements* of the matrix.

The lower order matrix associated with a_1, that is

$$\begin{vmatrix} b_2 & c_2 \\ b_3 & c_3 \end{vmatrix}$$

is called the *cofactor* of a_1, denoted A_1. Similarly,

$$-\begin{vmatrix} a_2 & c_2 \\ a_3 & c_3 \end{vmatrix}$$

is the cofactor of b_1, denoted B_1, and so on. Thus the solution of the third order matrix is

$$a_1 A_1 + b_1 B_1 + c_1 C_1$$

and so on for higher orders.

Practical Reliability Engineering, Fifth Edition. Patrick D. T. O'Connor and Andre Kleyner.

To get the cofactor signs correct, think of the element positions in the matrix as having positive and negative signs associated with them, as follows:

$$\begin{vmatrix} + & - & + & - & . & . & . \\ - & + & - & + & . & . & . \\ + & - & + & - & . & . & . \\ . & . & . & . & & & \\ . & . & . & . & & & \\ . & . & . & . & & & \end{vmatrix}$$

Matrix multiplication

$$\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} \times \begin{vmatrix} A_1 & B_1 \\ A_2 & B_2 \end{vmatrix}$$
$$= \begin{vmatrix} (a_1A_1 + b_1A_2) & (a_1B_1 + b_1B_2) \\ (a_2A_1 + b_2A_2) & (a_2B_1 + b_2B_2) \end{vmatrix}$$

Index

Practical Reliability Engineering, Fifth Edition. Patrick D. T. O'Connor and Andre Kleyner.