Milan Holický

Introduction to Probability and Statistics for Engineers

Springer

Introduction to Probability and Statistics for Engineers

Milan Holický

Introduction to Probability and Statistics for Engineers

Springer

Milan Holický
Klokner Institute, Department of Structural Reliability
Czech Technical University in Prague
Prague
Czech Republic

ISBN 978-3-642-38299-4 ISBN 978-3-642-38300-7 (eBook)
DOI 10.1007/978-3-642-38300-7
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013945183

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The theory of probability and mathematical statistics is becoming indispensable knowledge for an increasing number of engineers and scientists. This is caused by the enlarging significance of the economic and societal consequences of technological systems due to uncertainties affecting performance. That is why the fundamental concepts and procedures used to analyse, design, execute and utilize these systems are, at present, based on the theory of probability and mathematical statistics.

However, these probabilistic and statistical concepts are still not properly understood or interpreted by experts, including engineers and scientists involved in the various stages of system developments. The present book is an attempt to improve this undesirable situation by providing easily understandable descriptions of seemingly intricate probabilistic concepts. The objective of the book is to provide a concise and transparent explanation of the most important concepts commonly used in engineering and science.

The book is based on the lecture notes of the author, used in undergraduate and graduate courses at technical universities. The text is written in simple language using limited mathematical tools and emphasising the didactic aspects. All the theoretical concepts and procedures have been systematically illustrated by numerical examples. Selected, commercially available software products (in particular Excel and MATHCAD) are utilized in the practical numerical examples given.

Other than a basic knowledge of undergraduate mathematics, no special prerequisites are needed. This book is in the form of a text book, but can also be used as a reference handbook by undergraduate and graduate students, engineers and scientists, and by all specialists in the field of statistical evaluation of data, reliability analysis and risk assessment.

Due to the limited scope of the book, some concepts and procedures have been introduced without due theoretical development. In such cases, a reference to specialised literature is provided. However, in order to make the text understandable, the theoretical procedures are often illustrated by explanatory examples, which extend the main text and also indicate further possible applications of the general theoretical concepts.

The author expresses his gratitude to Mr. Phil Shoenfelt and Ms. Jana Pallierová from the Klokner Institute of the Czech Technical University in Prague, and to the editors of Springer for their kind help in the preparation of the manuscript.

Prague Prof. Dr. Milan Holický
October 2012

Contents

Chapter 1
Introduction

The theory of probability and mathematical statistics is becoming an indispensable tool in the analysis of many contemporary tasks in engineering and science. Wide-ranging experience confirms that probability is one of the most important concepts in modern science. In spite of that, comprehensive understanding of the fundamental principles of the theory of probability and statistics still seems to be inadequate. Consequently, the results of numerous applications are often understood and interpreted loosely and imprecisely.

Nevertheless, an increasing number of procedures applied in various technical fields are nowadays based on the theory of probability and mathematical statistics [1, 2]. The majority of new professional rules and provisions, codes of practice and standards are based on principles of the theory of probability and statistics [3–5]. Consequently, many engineers, scientists, experts, public officers and other experts participating in any process of decision making are directly or indirectly confronted with new knowledge, procedures and terminology [3, 6–8]. The present book is an attempt to provide a concise introductory text aimed at all classes of involved specialists, experts and the general public.

It is well recognised that many technological systems including engineering works suffer from a number of significant uncertainties which may appear at all stages of design, execution and use [9, 10]. Some uncertainties can never be avoided or eliminated entirely and must therefore be taken into account when analysing, designing, executing and using the system. The following types of uncertainties can be generally recognised [11]:

- Natural randomness of actions, material properties and geometric data;
- Statistical uncertainties due to limited available data;
- Uncertainties of theoretical models due to simplifications;
- Vagueness due to inaccurate definitions of performance requirements;
- Gross errors in design, execution and operation of the system;
- Lack of knowledge of the behaviour of elements in real conditions.

Note that the order of the listed uncertainties corresponds approximately to the decreasing scope of current knowledge, as well as to the lack of theoretical tools,

M. Holický, *Introduction to Probability and Statistics for Engineers*,
DOI 10.1007/978-3-642-38300-7_1, © Springer-Verlag Berlin Heidelberg 2013

making it problematic to take these uncertainties into account in any analysis. Depending on the nature of the system, environmental conditions and applied influences, some types of uncertainties may become critical, while some may turn out to be insignificant.

The above uncertainties are sometimes classified into two main groups:

– Aleatoric (random) uncertainties;
– Epistemic (notional) uncertainties.

Some of the above mentioned uncertainties are mainly aleatoric (for example natural randomness), the other can be classified as epistemic (for example lack of knowledge). Nevertheless theoretical tools to analyse uncertainties are confined primarily to the theory of probability and mathematical statistics.

The natural randomness and statistical uncertainties (mainly aleatoric) may be relatively well described by the available methods of the theory of mathematical statistics. In fact the International Standards [3–5] provide some guidance on how to proceed. However, lack of reliable experimental data, i.e. statistical uncertainty, particularly in the case of new elements, actions, environmental influences, and also specific geometrical data, causes significant problems.

Moreover, the available sample data are often not homogeneous and are obtained under differing conditions (for example in the case of material properties, imposed loads, environmental influences and hidden internal dimensions). Then, it becomes difficult – if not impossible – to use these data for developing general theoretical models.

Uncertainties regarding applied theoretical models used to analyse the system may be, to a certain extent, assessed on the basis of theoretical and experimental investigation. The vagueness caused by inaccurate definitions (in particular of some performance requirements) may be partially described by the theory of fuzzy sets. Up to now, however, these methods have been of little practical significance, as suitable experimental data are very difficult to obtain. Knowledge about the behaviour of new materials and structural systems is gradually improving thanks to newly developing theoretical tools and experimental techniques.

The lack of available theoretical tools is obvious in the instances of gross error and lack of knowledge (epistemic uncertainties), which are, nevertheless, often the decisive causes of system failures. In order to limit the gross errors caused by human activity, quality management systems, including the methods of statistical inspection and control (see for example [3–5]), can be effectively applied.

A number of theoretical methods and operational techniques have been developed and used to control the unfavourable effects of various uncertainties during a specified working life. Simultaneously, the theory of reliability has been developed to describe and analyse the uncertainties in a rational way, and to take them into account in analysis and verification of system performance. In fact, the development of the whole reliability theory was initiated by observed insufficiencies and failures caused by various uncertainties. At present the theory of reliability is extensively used in many technical areas to develop and calibrate operational procedures for assessing reliability.

That is why the theory of probability and mathematical statistics is becoming an indispensable discipline in many branches of engineering and science. This is also caused by the increasing economic and social consequences of various failures affecting the performance of complex technological systems. Fundamental concepts and procedures used in analysis of these systems are inevitably based on the theory of probability and mathematical statistics.

The objective of this book is to provide a concise description of basic theoretical concepts and practical tools of probability theory and mathematical statistics that are commonly used in engineering and science. The book starts with the fundamental principles of probability theory that are supplemented by evaluation of experimental data, theoretical models of random variables, sampling theory, distribution updating and tests of statistical hypotheses. In addition two-dimensional random samples and a short description of random functions are provided. Examples concerning various technical problems are used throughout the book to illustrate basic theoretical concepts and to demonstrate their potential when applied to the analysis of engineering and scientific systems. The text is structured into 12 chapters and 7 annexes as follows.

The basic concepts of probability theory, including the Bayesian approach, are introduced in Chap. 2. Chapter 3 deals with the necessary techniques for evaluation of statistical data. General concepts used to describe various discrete and continuous random variables are generally introduced in Chap. 4. Selected theoretical models of discrete and continuous random variables commonly used in engineering and science are presented in Chaps. 5 and 6. Chapter 7 is devoted to fundamental linear and nonlinear functions of random variables describing engineering and scientific problems. The estimation of population parameters, together with sampling distributions, is described in Chap. 8. One of the keywords of engineering and scientific applications is the term fractile (sometimes called quantile); relevant statistical techniques for determining the fractiles of theoretical models and for estimating them from limited sample data are described in Chap. 9. The testing of statistical hypotheses is covered in Chap. 10. Two-dimensional samples and populations, coefficients of correlation and regression (including their estimation from a sample), and tests of the population coefficients of correlation and regression are described in Chap. 11. Finally, random functions, more and more frequently applied in engineering and science, are briefly introduced in Chap. 12. Annexes summarize characteristic of random samples, parameters of population, theoretical models of random variables, parameters of functions of random variables, techniques for estimating fractiles, conventional models of random variables, and include a brief table of standardized normal distribution.

The book is based on limited mathematical tools and all the theoretical procedures are systematically illustrated by numerical examples. The text is written in simple language with an emphasis on didactic aspects. Except for a basic knowledge of undergraduate mathematics, no special prerequisites are required. The aim of the book is to provide a concise introductory textbook and handbook that will provide quick answers to practical questions.

The size of the book has been kept deliberately limited and, consequently, some procedures are introduced without detailed theoretical development. In such cases a reference to specialised literature is provided (particularly to the book [1] and other publications [12–19]). In order to make the text understandable, the theoretical procedures are often explained with the help of illustrative examples that supplement the main text and indicate further possible applications of the theoretical concepts in engineering and science. Due to the limited extend of the book numerical values of random variables required in examples are mostly taken from easily accessible tables provided by software products like EXCEL, MATHCAD, STATISTICA and numerous other products, or from Statistical tables available on the internet. A brief numerical table for the distribution function of a normal distribution is also provided in Appendix 7.

The book has the character of a text book, but can be used also as a concise handbook. It is primarily intended for undergraduate and graduate students of engineering and science, for engineers, scientific workers, and other specialists participating in the field of evaluation of statistical data, reliability analysis, risk assessment and decision making.

References

1. Ang, A.H.-S., Tang, W.H.: Probabilistic Concepts in Engineering. Emphasis on Applications to Civil Environmental Engineering. Wiley, New York (2007)
2. Devore, J., Farnum, N.: Applied Statistics for Engineers and Scientists. Thomson, London (2005)
3. Dunin-Barkovskij, I.V., Smirnov, N.V.: The Theory of Probability and Mathematical Statistics in Engineering. Technical and Theoretical Literature, Moscow (1955) (in Russian)
4. Gurskij, E.I.: The Theory Probability with Elements of Mathematical Statistics. Higher School, Moscow (1971) (in Russian)
5. Holický, M.: Reliability Analysis for Structural Design. SUNN MeDIA, Stellenbosch (2009)
6. Blockley, D.I.: The Nature of Structural Design and Safety. Ellis Hotwood Limited, Chichester (1980)
7. EN 1990: Eurocode – Basis of Structural Design. CEN, Brussels (2002)
8. ISO 2394: General Principles on Reliability for Structures. ISO, Geneve (1998)
9. ISO 12491: Statistical Methods for Quality Control of Building Materials and Components. ISO, Geneve (1997)
10. ISO 13822: Assessment of Existing Structures. ISO, Geneve (2002)
11. ISO 3534-1: Statistics – Vocabulary and Symbols – Part 1: Probability and General Statistical Terms. ISO, Geneve (1993)
12. ISO 3534-2: Statistics – Vocabulary and Symbols – Part 2: Statistical Quality Control. ISO, Geneve (1993)
13. Stewart, M.G., Melchers, R.E.: Probabilistic Risk Assessment of Engineering Systems. Chapman & Hall, London (1997)
14. Melchers, R.E.: Structural Reliability: Analysis and Prediction. Wiley, New York (1999)
15. Christensen, P.T., Baker, M.: Structural Reliability Theory and Its Applications. Springer, Berlin (1982)
16. Madsen, H.O., Krenk, S., Lind, N.C.: Methods of Structural Safety. Prentice-Hall Inc., Englewood Cliffs (1986)
17. Nowak, A.S., Collins, K.R.: Reliability of Structures. McGraw Hill, Boston (2000)
18. Schneider, J.: Introduction to Safety and Reliability of Structures. IABSE, Zürich (1997)
19. Krishnana, V.: Probability and Random Processes. Wiley, New York (2005)

Chapter 2
Basic Concepts of Probability

The basic concepts of the theory of probability, frequently applied in engineering and scientific tasks, are best illustrated by practical examples. Fundamental terms like "experiment", "random event" and "sample space" are supplemented by descriptions of the common relations between random events. The key notion of probability is defined, taking into account historical approaches and practical interpretations related to engineering and scientific applications. The basic rules for the calculation of probability are illustrated by numerical examples. The essential concept of conditional probability is clarified in detail and was used to develop the Bayes' theorem. Various applications of the theorem are demonstrated by examples taken from engineering. An extension of the Bayes' theorem is used to develop operational procedures for probability updating.

2.1 Experiment, Random Event, Sample Space

This chapter gives a concise overview of the most important concepts and terms of the theory of probability, especially those which are most often used in reliability analyses of civil engineering works and systems. The presentation of some concepts and laws is rather intuitive without rigorous mathematical proofs. More detailed explanations of all the relevant concepts, theorems and rules may be found in specialised literature [1, 2].

The most significant fundamental concepts of the theory of probability applied in structural reliability include:

- Experiment;
- Random event; and
- Sample space (space of events).

These terms are used in classical probability theory, but are also applicable in contemporary probability theory based on the theory of sets.

M. Holický, *Introduction to Probability and Statistics for Engineers*,
DOI 10.1007/978-3-642-38300-7_2, © Springer-Verlag Berlin Heidelberg 2013

2.1.1 Experiment

An experiment in the theory of probability is understood as the realization of a certain set of conditions π. In the classical approach to the theory of probability it is often assumed that the experiments can be repeated arbitrarily (e.g. tossing a dice, testing a concrete cube) and the result of each experiment can be unambiguously used to declare whether a certain event occurred or did not occur (e.g., when tossing a dice, obtaining or not obtaining a predetermined number, or when a concrete cube exceeds or does not exceed a specified value).

However, in practical applications of the theory of probability the assumption of arbitrary repeatable experiments, each of which leads to an unequivocal result (even though not known beforehand), is not always realistic (e.g. in the construction industry, where usually only a very limited number of experiments can be performed). Contemporary usage of the theory of probability allows for more general concepts, wherein terms such as experiment, event and sequence of events are related to the general theory of sets.

The concept of an experiment remains applicable in general. However, specification of the conditions π is of the utmost importance, irrespective of whether the experiment can be realistically carried out. In some cases the experiment can only be carried out hypothetically. In any case, the specification of the conditions π needs to be as accurate and as complete as possible. The results and interpretation of any experiment should always be related to these conditions. The comparison of experiments carried out under different conditions may lead to serious mistakes and misunderstandings. Description of the appropriate set of conditions and their verification should therefore become an indispensable part of every probabilistic analysis.

2.1.2 Random Event

Probability theory deals with the results of experiments that are not unequivocally determined in advance by the appropriate set of conditions π, or with experiments for which a set of conditions that would lead to an unequivocal result either cannot be provided during an experiment, or is not known at all (or partly unknown). An experiment of this kind is called a random experiment. The result of a random experiment is characterised by events that could, but will not necessarily, occur when the conditions π are realized. Such events are called random events and are usually denoted by a capital letter from the beginning of the alphabet, e.g. A or B (possibly with an index). An event that will necessarily occur every time the conditions π are realized is called a certain event – denoted here by the symbol U; an event that can never occur is called an impossible event – usually denoted as V.

2.1.3 Sample Space

The sample space Λ of a certain random experiment denotes all events which can occur by the realization of a specified set of conditions π, i.e. those which can be outcomes of the experiment under consideration. The sample space can be finite (tossing a dice), or infinite (testing a concrete cube in a testing machine). In some cases a system of elementary events can be found, i.e. a system of events that cannot be further divided (e.g. tossing a dice numbers 1–6 represents such elementary events). In other cases the system of elementary events is not obvious, or does not exist (testing a cube in a testing machine).

All the fundamental terms that have been introduced – experiment, set of conditions π, event and sample space Λ – are clarified by the following three simple examples, which constitute an integral part of this summary. Besides a complementary explanation of the relevant terms, the following examples provide further information on the general principles and mathematical tools used to describe real conditions and accepted simplifying assumptions.

Example 2.1. Tossing a dice is a traditional (and from an educational viewpoint a very useful) example of a random experiment. In this case the set of conditions π is trivial. The dice is balanced and symmetrical and cast in a correct way that will not affect the number obtained when the dice is tossed.

The certain event U denotes the event where any of the numbers 1, 2, 3, 4, 5 or 6 occur.

The impossible event V denotes the event when other numbers appear (e.g. 0, 7, 9 etc.).

Elementary events E_i, $i = 1, 2$ to 6, which cannot be further divided, denote the number $i = 1, 2$ to 6. If the given set of conditions π is satisfied, the occurrence of every elementary event is equally possible. In this case we can say that it is a system of equally possible elementary events.

Random events A_i, for example the appearance of the number 1, can be denoted as $A_1 = E_1$; the appearance of the even numbers as $A_2 = E_2 \cup E_4 \cup E_6$; the appearance of the numbers divisible by three as $A_3 = E_3 \cup E_6$; the appearance of the numbers divisible by two or three as $A_4 = E_2 \cup E_3 \cup E_4 \cup E_6$, etc. The sample space Λ (i.e. the total of all possible events which may occur at a toss) is, in this case, obviously finite.

Example 2.2. The cylinder strength of a specific concrete is examined. The random experiment is the loading of a test cylinder into a testing machine. The set of conditions π includes the composition, treatment and age of concrete, the dimensions of the cube, the speed of the loading, etc. The investigated random event is the failure of the concrete cylinder at a certain level of loading. If the loading is sufficiently high, the cylinder always fails; at sufficiently low levels of loading the failure will never occur. At the loading level corresponding to the characteristic cylinder strength of concrete the failure may occur (e.g. 5 % of all cases on average) or may not (e.g. 95 % of all cases).

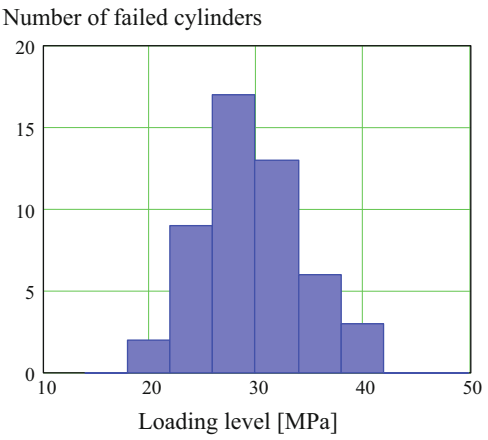


Fig. 2.1 The number of failed cylinders versus the loading level

Elementary events can be defined in many ways, e.g. by a system of intervals of equal width within a certain loading range. Consider concrete of the grade C 20 having the characteristic cylinder strength 20 MPa. The possible range of loading from 10 to 50 MPa is split into the intervals of 4 MPa and elementary events are defined as the failure of a cylinder at the loading level within each interval. Figure 2.1 shows the results of 50 experiments under specified conditions π. Solid bars in each interval indicate the number of failed cylinders. It follows from Fig. 2.1 that two cylinders failed at the loading level of 18–22 MPa, nine cylinders failed in the next interval of 22–26 MPa, 17 cylinders failed in the interval of 24–30 MPa, etc. Without doubt, it is not a system of equally possible events (see Fig. 2.1). The sample space Λ consists of any one-sided or two-sided interval and is obviously infinite. Figure 2.1 shows a frequently used graphical representation of experimental results, called a histogram, which is commonly used for the development of probabilistic models describing basic variables.

Example 2.3. Consider throwing a dart onto a board indicated in Fig. 2.2. Each throw represents one realization of a random experiment. The set of conditions π includes the distance of the board from the throwing point, the size of the board, the type of dart and other conditions for throwing.

It is assumed that every point of the board can be hit with equal likelihood, and that the board is always hit (these are, undoubtedly, questionable assumptions).

The hitting of the whole board is therefore a certain event U.

An impossible event V is a throw that misses the board.

A random event, though, may be the hitting of any smaller area, A or B, drawn on the board (Fig. 2.2), or of any combination of such areas. The system of all possible areas on the board represents an infinite sample space Λ.

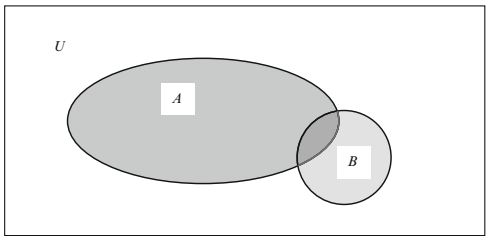


Fig. 2.2 An example of throwing a dart onto a board – Venn diagram

2.2 Relations Between Random Events

Example 2.3 indicates a common representation of random events (see Fig. 2.2) using oval figures illustrating the mutual relations between the random events A, B, C,... (such a representation is called a Venn diagram). In Fig. 2.2 the certain event U is represented by the whole rectangle, two random events A and B are symbolized by the ovals. Let us consider some basic relations between events A and B, which lead to the definition of other important terms and to the derivation of some general relationships between the random events. Other diagrams similar to Fig. 2.2 may illustrate all the following relationships and combinations of random events. A detailed explanation including formal mathematical proofs of all rules may be found in specialised literature [3, 4].

If an event B occurs every time the conditions π are realized, as a result of which an event A occurs, we say that event A implies event B, which is usually symbolically denoted as $A \subset B$. If the events A and B occur simultaneously every time the conditions π are realized, we say that an intersection of the two events occurs. An intersection of the events A and B is denoted as $A \cap B$. If at least one of the events A and B occurs at every realization of the conditions π, a union of the two events occurs. This is denoted by $A \cup B$. If event A occurs but event B does not, then the difference A–B of the two events occurs. Events A and $\bar{A}$ are complementary events (we also say that event $\bar{A}$ is the opposite of event A) if it holds simultaneously that $A \cup \bar{A} = U$ and $A \cap \bar{A} = V$. It can be shown that the following simple rules (the commutative, associative and distributive laws) hold for the intersection and union of random events:

$$A \cap B = B \cap A, \quad A \cup B = B \cup A \tag{2.1}$$

$$(A \cap B) \cap C = A \cap (B \cap C), \quad (A \cup B) \cup C = A \cup (B \cup C) \tag{2.2}$$

$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C), \quad A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \tag{2.3}$$

These basic rules lead to the definition of more complicated relations of the intersection and the union of a system of events A_i:

$$\bigcap_i A_\mathrm{i} = A_1 \cap A_2 \cap A_3 \cap \ldots \cap A_\mathrm{n}$$
$$\bigcup_i A_\mathrm{i} = A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_\mathrm{n} \tag{2.4}$$

The following rules (the so-called de Morgan rules), the validity of which follows from the above relations, are sometimes effectively applied in practical computations of probabilities of complex events

$$\overline{\bigcap_i A_i} = \overline{A_i} \cup \overline{A_2} \cup \ldots \cup \overline{A_n}$$
$$\overline{\bigcup_i A_i} = \overline{A_1} \cap \overline{A_2} \cap \ldots \cap \overline{A_n} \tag{2.5}$$

The use of these rules is evident from the two following examples.

Example 2.4. A simple serial system loaded by forces P consists of two elements as shown in Fig. 2.3. Failure F of the system may occur due to failure F_1 of the element 1 or due to failure F_2 of the element 2:

$$F = F_1 \cup F_2$$

The complementary event $\overline{F}$ (no failure) is, according to relation (2.5), described by an event for which it holds

$$\overline{F} = \overline{F_1 \cup F_2} = \overline{F_1} \cap \overline{F_2}$$

Example 2.5. A town C is supplied with water from two sources, A and B, by a pipeline, which consists of three independent branches 1, 2 and 3 (see the scheme in Fig. 2.4). Let us denote F_1 as the failure of branch 1, F_2 the failure of branch 2 and F_3 the failure of branch 3. In a case where the sources A and B have sufficient capacity to supply the town C, the lack of water in that town is described by the event $(F_1 \cap F_2) \cup F_3$; here, either the branch 3 fails or the branches 1 and 2 fail. For the analysis of this event, however, it may be expedient to study a complementary event describing the sufficiency of water in the town C.

According to de Morgan's rules the complementary event of the sufficiency of water in the town C is described by the event

$$\overline{(F_1 \cap F_2) \cup F_3} = (\overline{F_1} \cup \overline{F_2}) \cap \overline{F_3}$$

where the event $(\overline{F_1} \cup \overline{F_2})$ represents sufficient water in the junction of branches 1 and 2, which is at the same time the beginning of branch 3.

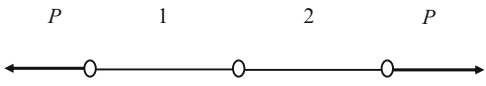


Fig. 2.3 A serial system

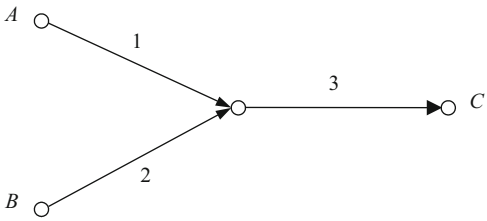


Fig. 2.4 Water supply system of a town C from sources A and B

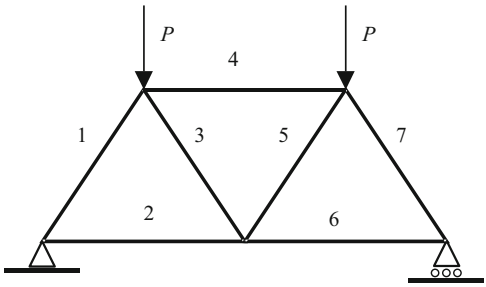


Fig. 2.5 Statically determinate truss structure

Example 2.6. Let us consider the statically determinate truss structure shown in Fig. 2.5, consisting of seven members and loaded by forces P. The aim is to describe an event F as that in which a structural failure occurred. Let F_i denote the event in which a failure of the element $i = 1, 2, \ldots, 7$ occurred.

The failure of the whole structure (event F) occurs if a failure of at least one of the members occurs. Therefore it holds that

$$F = \bigcup_{i=1}^{7} F_i$$

With regard to the conditions of manufacture of the individual members the events F_i may be mutually dependent and thus are not exclusive. In the computation of the probability of failure it may then be expedient to consider the complementary event $\overline{F}$ for which it holds, according to de Morgan's rules (2.5)

$$\overline{F} = \overline{\bigcup_{i=1}^{7} F_i} = \bigcap_{i=1}^{7} \overline{F_i}$$

Similar relationships may be effectively used when analysing complex technical systems.

The following additional terms are often used. We say that a system of events A_i forms a complete system of events if the union of these events is a certain event U. In that case at least one event, A_i, always occurs. A complete system of mutually exclusive events is another term that is sometimes used when analysing complex events. In that case only one event A_i always occurs.

2.3 Definitions of Probability

2.3.1 Classical Definition

Probability describes the occurrence of random events. The definition of probability is, however, a mathematically intricate problem. Historically it has experienced an interesting evolution, reflecting the remarkable development of the theory of probability and its practical applications. The classical definition of probability is based on a complete system of elementary events. Let an event A consist of m out of n, equally likely elementary events where the total number n is formed by a complete system of mutually exclusive events. The probability of event A is then given by the quotient

$$\mathrm{P}(A) = m/n \tag{2.6}$$

For probability defined in this way it obviously holds that

$$0 \leq \mathrm{P}(A) = m/n \leq 1 \tag{2.7}$$

$$\mathrm{P}(U) = n/n = 1, \quad \mathrm{P}(V) = 0/n = 0 \tag{2.8}$$

It can also be shown for a system of mutually exclusive events A_i that the probability of the union of these events is given by the relation

$$\mathrm{P}\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} \mathrm{P}[A_i] \tag{2.9}$$

The classical definition of probability is fully acceptable for many elementary cases, such as the tossing of a dice in Example 2.1. However, if the dice is not symmetrical, the classical definition obviously fails. Examples 2.2 and 2.3 further indicate that a finite system of elementary events is not sufficient for the fundamental problems of civil engineering. In the attempt to deal with these insufficiencies other definitions of probability gradually emerged.

2.3.2 Geometrical Definition

The geometrical definition of probability is related to the throwing of a dart in Example 2.3. According to this definition, the probability of an event A is equal to the quotient of the surface area of event A, the denoted area(A), and of the surface area of the certain event U, the denoted area(U), i.e. by the quotient

$$\mathrm{P}(A) = \mathrm{area}(A)/\mathrm{area}(U) \tag{2.10}$$

Thus, the geometric definition attempts to eliminate one insufficiency of the classical definition, which lies in the finite number of elementary events. However, this definition still does not take into account the reality that not all the points on the board (event U) have the same possibility of occurrence. Obviously, the "surface area" is not an appropriate measure of occurrence; this difficulty is still to be solved.

2.3.3 Statistical Definition

The statistical definition of probability is based on the results of an experiment repeated many times. Let us consider a sequence of n realizations of an experiment. Assume that a certain event A comes up $m(A)$ times out of these n experiments. It appears that the relative frequency of the occurrence of the event A, i.e. the fraction $m(A)/n$, attains an almost constant value with an increasing number of realizations n. This phenomenon is called the statistical stability of relative frequencies, i.e. of the fraction $m(A)/n$. The value to which the relative frequency $m(A)/n$ approaches as n increases ($n \to \infty$) is accepted as an objective measure of the occurrence of the event A and is called the probability $\mathrm{P}(A)$ of the event A:

$$\mathrm{P}(A) = \lim_{n\to\infty} \frac{m(A)}{n} \tag{2.11}$$

However, the assumption of statistical stability and convergence indicated in Eq. (2.11) (i.e. the limit of the quantity derived from the results of experiments) causes some mathematical difficulties.

2.3.4 Axiomatic Definition

The classical, geometrical as well as statistical definitions of probability attempt to define not only the probability, but also to propose a rule for its computation – something that is extremely difficult and perhaps impossible, to achieve.

The long-term effort to define the basic terms of the theory of probability seems to reach fruition with the so-called axiomatic system, which is accepted all over the world. The axiomatic system defines only the term of probability and its fundamental properties without providing any practical instructions for its determination.

Note that Eqs. (2.7, 2.8, and 2.9) characterize the common properties of the classical, geometrical as well as statistical definition of probability:

1. The probability of a certain event is equal to 1;
2. The probability of an impossible event is equal to 0; and
3. If an event A is a union of partial and mutually exclusive events $A_1, A_2, \ldots, A_n$, then the probability of event A is equal to the sum of probabilities of the partial events.

The axiomatic definition of probability introduces these general properties as axioms. Probability P is a real function, defined in a sample space Λ above the certain event U with these properties:

1. If $A \in \Lambda$, then

$$P(A) = \geq 0 \tag{2.12}$$

2. For the certain event U, it holds that

$$\mathrm{P}(U) = 1 \tag{2.13}$$

3. If $A_i \in \Lambda$, $i = 1, 2, \ldots$ and if for arbitrary i and j $A_i \cap A_j = V$, then

$$\mathrm{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathrm{P}(A_i) \tag{2.14}$$

It can be seen that the above-mentioned three axioms are confirmed by the classical, geometrical and statistical definitions. Moreover, the axiomatic definition of probability also fits the new concept of probability as a measure of the fulfilment of a statement or requirement, often assessed only by intuition and a subjective view (an expert judgement). This subjective definition is sometimes called Bayesian probability. However, it should be noted that in this approach the concept of reproducible (repeatable) random events, which forms the basis for the probability determination of an event, is completely impossible.

Note that by using the above axioms, the modern theory of probability transfers into the general theory of sets. Probability is then defined as a non-negative additive function of sets, which can be interpreted as a generalization of the term "surface area" in the geometrical definition of probability.

2.4 Basic Rules for the Computation of Probabilities

Using Eqs. (2.6, 2.7, 2.8, and 2.9) or axioms Eqs. (2.12, 2.13, and 2.14), other rules, which can be useful in computations of probabilities, can be derived. If A_i, $i = 1, 2, \ldots n$, form a complete system of events, then it evidently holds that

$$\mathrm{P}\left(\bigcup_{i=1}^{n} A_i\right) = \mathrm{P}(U) = 1 \tag{2.15}$$

If an event A is a union of partial and mutually exclusive events, A_i, $i = 1, 2, \ldots, n$, we can write

$$\mathrm{P}(A) = \mathrm{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathrm{P}(A_i) \tag{2.16}$$

For the probability of the union of two arbitrary events A and B (which do not have to be exclusive) the principle of the summation of probabilities holds

$$\mathrm{P}(A \cup \mathrm{B}) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B) \tag{2.17}$$

which follows from Eq. (2.16) for mutually exclusive events A and $B - (A \cap B)$, of which the union is the studied event $A \cup B$.

If A_i, $i = 1, 2, \ldots, n$, is a complete system of mutually exclusive events, then we obtain from Eq. (2.15)

$$\mathrm{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathrm{P}(A_i) = \mathrm{P}(U) = 1 \tag{2.18}$$

From Eq. (2.18) for complementary events A and $\bar{A}$ it follows that

$$\mathrm{P}(\bar{A}) = 1 - \mathrm{P}(A) \tag{2.19}$$

Example 2.7. Let us determine the probability that a serial system described in Example 2.4 will fail. Let $\mathrm{P}(F_1) = 0.05$, $\mathrm{P}(F_2) = 0.05$ and $\mathrm{P}(F_1 \cap F_2) = 0.02$. Then, considering the relation (2.17), we find that

$$\mathrm{P}(F_1 \cup F_2) = \mathrm{P}(F_1) + \mathrm{P}(F_2) - \mathrm{P}(F_1 \cap F_2) = 0.05 + 0.05 - 0.02 = 0.08$$

Note that the events F_1 and F_2 are not exclusive (failures of both elements can occur simultaneously). If they were exclusive, the probability of failure would be 0.10. Other details concerning this example will be provided in the next section by the principle of multiplication of probabilities.

2.5 Conditional Probability

Conditional probability $P(A|B)$ of the event A under a complementary condition that another event B has occurred simultaneously (or before), and has a non-zero probability, is an important concept in the contemporary theory of probability which is often used in the theory of structural reliability. The conditional probability $P(A|B)$ is defined as the fraction

$$P(A|B) = P(A \cap B)/P(B) \tag{2.20}$$

This relation is the basis of the so-called Bayes concept of the theory of probability (Thomas Bayes (1702–1761)). In two special cases important simplifications of relation (2.20) are valid. If events A and B are exclusive, i.e. $A \cap B = V$, then $P(A|B) = 0$; if an event A implies an event B, i.e. it holds that $A \subset B$, then $P(A|B) = P(A)/P(B)$. If $B \subset A$, then $P(A|B) = 1$. These rules follow directly from the basic properties of probability described in Sects. 2.2 and 2.3.

A general rule for the multiplication of probabilities follows from Eq. (2.20)

$$P(A \cap B) = P(B)\ P(A|B) \tag{2.21}$$

Consider again the special cases. If the events A and B are exclusive, i.e. $A \cap B = V$, then $P(A|B) = 0$ and also $P(A \cap B) = 0$; if $A \subset B$, then $P(A|B) = P(A)/P(B)$ and $P(A \cap B) = P(B)$; if, conversely, $B \subset A$, then $P(A|B) = 1$ and $P(A \cap B) = P(B)$.

We say that events A and B are independent (the occurrence of event B does not influence the probability of the occurrence of event A) if it holds that $P(A|B) = P(A)$. Consider the special cases introduced above. If events A and B are exclusive, then they are dependent because $P(A|B) = 0 \neq P(A)$ (if A is not an impossible event). If $A \subset B$, then A and B are dependent events, because $P(A|B) = P(A)/P(B) \neq P(A)$, if conversely $B \subset A$, then the events A and B are dependent, because $P(A|B) = 1 \neq P(A)$. Therefore independent events A and B must not be exclusive, i.e. $A \cap B \neq V$, and satisfy the trivial relations $A \not\subset B$ and $B \not\subset A$.

If two events A and B are independent (and therefore it holds that $A \cap B \neq V$, $A \not\subset B$ and $B \not\subset A$), then it follows from Eq. (2.21)

$$P(A \cap B) = P(A)\ P(B) \tag{2.22}$$

Relation (2.22) is the principle of the multiplication of probabilities, according to which the probability of intersection (a simultaneous occurrence of two independent random events) is given by the product of their probabilities. This fundamental rule is needed for probability integration in the theory of reliability.

Example 2.8. Taking into account relation (2.21), the following relation can be written for the probability of failure of a serial system, as described in Example 2.7

$$\mathrm{P}(F) = \mathrm{P}(F_1 \cup F_2) = \mathrm{P}(F_1) + \mathrm{P}(F_2) - \mathrm{P}(F_1 \cap F_2)$$
$$= \mathrm{P}(F_1) + \mathrm{P}(F_2) - \mathrm{P}(F_1)\mathrm{P}(F_2|F_1) = 0.10 - 0.05\mathrm{P}(F_2|F_1)$$

If the events F_1 and F_2 are independent, then $\mathrm{P}(F_2|F_1) = \mathrm{P}(F_2)$ and the failure probability is given as

$$\mathrm{P}(F) = \mathrm{P}(F_1 \cup F_2) = \mathrm{P}(F_1) + \mathrm{P}(F_2) - \mathrm{P}(F_1)\ \mathrm{P}(F_2) = 0.10 - 0.0025 = 0.0975$$

If the events F_1 and F_2 are perfectly dependent ($F_1 \subset F_2$), i.e. $\mathrm{P}(F_2|F_1) = 1$, then

$$\mathrm{P}(F) = \mathrm{P}(F_1 \cup F_2) = \mathrm{P}(F_1) + \mathrm{P}(F_2) - \mathrm{P}(F_1) = 0.10 - 0.05 = 0.05$$

The serial system acts in this case as a single element. Thus, in general, the probability of failure of the serial system under consideration fluctuates from 0.05 to 0.0975 depending on the degree of dependence of the events F_1 and F_2.

Assume that an event A can occur only by the realization of one of the mutually exclusive events B_i, $i = 1, 2, \ldots, n$ ($n = 5$ in Fig. 2.5), for which the probabilities $\mathrm{P}(B_i)$ are known. If the conditional probabilities $\mathrm{P}(A|B_i)$ are also known (obviously $\mathrm{P}(A|B_5) = 0$), then the probability of the event A can be assessed as

$$\mathrm{P}(A) = \sum_{i=1}^{n} \mathrm{P}(B_i)\ \mathrm{P}(A|B_i) \tag{2.23}$$

which is called the theorem of total probability.

2.6 Bayes' Theorem

When an event A occurs, it is natural to investigate which of the events B_i caused A, i.e. what is the probability of the individual hypotheses B_i assuming that A occurred (see Fig. 2.6), which is denoted as the probability $\mathrm{P}(B_i|A)$. A very important relation follows from relations (2.20, 2.21 and 2.23)

$$\mathrm{P}(B_i|A) = \frac{\mathrm{P}(B_i)\mathrm{P}(A|B_i)}{\sum_{i=1}^{n} \mathrm{P}(B_j)\mathrm{P}(A|B_j)} \tag{2.24}$$

which is often referred to as the Bayes rule or theorem.

The following important usage of the theory of structural reliability illustrates the common procedure for the practical application of the Bayes rule. If the failure of a structure, denoted as event A, can be caused by one of the hypotheses B_i whose

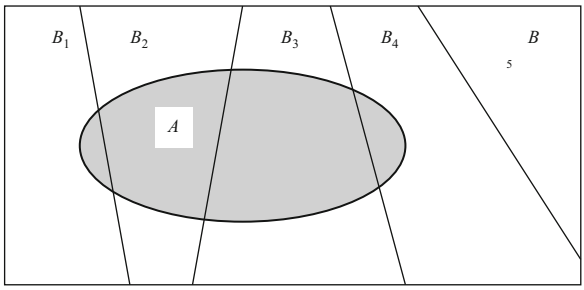


Fig. 2.6 An event A and mutually exclusive events B_i

probabilities of occurrence $P(B_i)$ are known from previous experience, and if the conditional probabilities $P(A|B_i)$ that the failure A occurred in consequence of a hypothesis B_i are also known, then the probability of failure $P(A)$ can be determined from the principle of total probability described by Eq. (2.23).

If, however, the failure A did occur, i.e. the event A occurred, and then the probabilities of the individual hypotheses, which could have caused the failure, are of importance. We are therefore interested in the conditional probabilities $P(B_i|A)$, which can be determined by use of the Bayes rule (2.24).

The practical use of relations (2.23) and (2.24) is illustrated by the following examples.

Example 2.9. For the assessment of an existing reinforced concrete structure, control tests are available which indicate that the actual strength is lower than the characteristic value 20 MPa (event B_1) with the probability $p_1' = P(B_1) = 0.05$, and greater than 20 MPa (event B_2) with the probability $p_2' = P(B_2) = 0.95$. For the subsequent verification of the strength of the concrete, an inaccurate non-destructive method is used. Let A denote the possibility that the concrete strength assessed by the non-destructive method is greater than 20 MPa. Assume that errors of the non-destructive method can be expressed by conditional probabilities

$$P(A|B_1) = 0.30,\ P(A|B_2) = 0.90$$

Thus, due to the inaccuracy of the non-destructive method, concrete with a strength lower than 20 MPa can be considered as concrete with a strength greater than 20 MPa with a probability of 0.30; at the same time, concrete of a strength greater than 20 MPa can be considered with a probability of 0.90.

The probability of the occurrence of event A (non-destructive strength is greater than 20 MPa) follows from the principle of complete probability (2.23)

$$P(A) = \sum_{i=1}^{2} P(B_i)P(A|B_i) = 0.05 * 0.30 + 0.95 * 0.90 = 0.87$$

This means that when using the inaccurate non-destructive method, concrete strength greater than 20 MPa will be predicted with a probability 0.87. Note that if the non-destructive tests were absolutely accurate, e.g. if the conditional probabilities were

$$\mathrm{P}(A|B_1) = 0,\ \mathrm{P}(A|B_2) = 1$$

it would follow from Eq. (2.23)

$$\mathrm{P}(A) = \sum_{i=1}^{2} \mathrm{P}(B_i)\mathrm{P}(A|B_i) = 0.05 * 0 + 0.95 * 1 = 0.95$$

However, from a practical point of view another question is more important: what is the probability $\mathrm{P}(B_2|A)$ of hypothesis B_2, that concrete for which the non-destructive test indicates a strength greater than 20 MPa (meaning that event A occurred) really does have a strength greater than 20 MPa (meaning that event B_2 occurred)? This probability can be assessed directly by using the Bayes rule (2.24) for the probability of hypotheses

$$\mathrm{P}(B_2|A) = \frac{\mathrm{P}(B_2)\mathrm{P}(A|B_2)}{\sum_{i=1}^{2} \mathrm{P}(B_j)\mathrm{P}(A|B_j)} = \frac{0.95 * 0.90}{0.05 * 0.30 + 0.95 * 0.90} = 0.98$$

Thus, if the strength is greater than 20 MPa according to the non-destructive test, then the probability that the concrete really does have a strength greater than 20 MPa increases from the original value of 0.95–0.98.

Bayes' rule is widely applied in many other practical situations in engineering practice, e.g. in situations where previous information about the distribution of probabilities is updated with regard to newly acquired knowledge. This important procedure of probability updating is described in the following section.

2.7 Probability Updating

Bayes' rule (2.24) is often applied to the so-called updating of the distribution of probabilities, which is based on random experiments (often repeated) isolated in time. Similarly, as in Sect. 2.5, it is assumed that these probabilities $\mathrm{P}(B_i)$ are known from previous experience (sometimes remote, vague or merely subjective). That is why they are called original (*a priori*) probabilities and they are denoted simply as $p_i' = \mathrm{P}(B_i)$.

Experiments are then carried out to determine the conditional probabilities $\mathrm{P}(A|B_i)$ of the studied event A, under the assumption that event B_i occurred, the outcomes of which can be considered as the measures of likelihood that the cause of

event A was the very event B_i. These conditional probabilities, or values proportional to them, are therefore called likelihoods $l_i \propto \mathrm{P}(A|B_i)$; the symbol $\propto$ means "proportional to" (likelihoods l_i thus need not necessarily be normalized to the conventional interval $<0, 1>$). We are inquiring about updated (*a posteriori*) probabilities $p_i'' = \mathrm{P}(B_i|A)$ of an event (i.e. hypothesis) B_i updated in accordance with the result of a new experiment (concerning event A). An important relation for p_i'' follows directly from the Bayes rule (2.24):

$$p_i'' = \frac{p_i' l_i}{\sum_j p_j' l_j} \tag{2.25}$$

Formula (2.25) obviously holds generally for likelihoods which, unlike probabilities, are not normalized to the interval $<0, 1>$ and only express the relative contribution of the events (hypotheses) B_i on the observed event A.

Relation (2.25) is a basis for the updating of probabilities, which is often applied in many engineering procedures, particularly in the assessment of existing structures. It is in these cases that present information is combined with previous (often subjective) information, i.e. with information about a structure at various points in time, usually quite remote. This is the reason why it is necessary to verify the conditions under which the previous information was obtained and to resist the temptation to apply non-homogeneous data, which may be misleading and could lead to serious mistakes and misunderstandings.

Example 2.10. Consider again the reinforced concrete structure described in Example 2.9. We observe that from previous control tests original (*a priori*) probabilities are known: $p_1' = \mathrm{P}(B_1) = 0.05$ (the probability that the real strength is lower than the characteristic value of 20 MPa, which is event B_1) and $p_2' = \mathrm{P}(B_2) = 0.95$ (the probability that the real strength is greater than 20 MPa, event B_2).

In the subsequent assessment of the structure supplementary tests of concrete strength are carried out using core samples, which are sufficiently accurate (unlike the non-destructive tests from previous Example 2.9). Thus in analysing the results it is not necessary to consider the inaccuracies. These tests suggest that the likelihood of event B_1 is $l_1 \propto \mathrm{P}(A|B_1) = 0.2$ and the likelihood of event B_2 is $l_2 \propto \mathrm{P}(A|B_2) = 0.8$ (the likelihoods introduced being already normalized). Updated (*a posteriori*) probabilities follow from relation (2.25)

$$p_1'' = \frac{p_1' l_1}{\sum_{j=1}^{2} p_j' l_j} = \frac{0.05 * 0.20}{0.05 * 0.20 + 0.95 * 0.80} = 0.01$$

$$p_2'' = \frac{p_2' l_2}{\sum_{j=1}^{2} p_j' l_j} = \frac{0.95 * 0.80}{0.05 * 0.20 + 0.95 * 0.80} = 0.99$$

Thus the updated (*a posteriori*) distribution of probabilities p_i'' is more favourable than the original (*a priori*) distribution of probabilities p_i'.

Note that when the supplementary tests suggest that the likelihoods of both events B_1 and B_2 are equal, e.g. $l_1 = \mathrm{P}(A|B_1) = l_2 = \mathrm{P}(A|B_2) = 0.5$, the updated probabilities equal the original ones ($p_i' = p_i''$). If, however, the analysis of event A showed that the likelihood of event B_1 is greater than the likelihood of event B_2, e.g. $l_1 \propto \mathrm{P}(A|B_1) = 0.7$ and $l_2 \propto \mathrm{P}(A|B_2) = 0.3$, the *a posteriori* probabilities change significantly:

$$p_1'' = \frac{p_1' l_1}{\sum_{j=1}^{2} p_j' l_j} = \frac{0.05 * 0.70}{0.05 * 0.70 + 0.95 * 0.30} = 0.11$$

$$p_2'' = \frac{p_2' l_2}{\sum_{j=1}^{2} p_j' l_j} = \frac{0.95 * 0.30}{0.05 * 0.70 + 0.95 * 0.30} = 0.89$$

The updated (*a posteriori*) distribution of probabilities p_i'' is then less favourable than the original (*a priori*) distribution p_i'. However, the influence of the *a priori* distribution still seems to prevail; it disappears only in the case of an extreme distribution of likelihoods, e.g. when l_1 approaches one ($l_1 \propto \mathrm{P}(A|B_1) \to 1$) and at the same time l_2 approaches zero ($l_2 \propto \mathrm{P}(A|B_2) \to 0$). But, in practice, the distribution of likelihoods is usually similar to the distribution of *a priori* probabilities.

Example 2.11. Tensile components of an existing structure have been designed for a load of 2 kN. After reconstruction, the load on each of these components is increased to 2.5 kN. Prior experience shows that the elements are able to resist a load of 2.5 kN (event B) with a probability $p_1' = \mathrm{P}(B) = 0.8$ and they fail with a probability $p'_2 = \mathrm{P}(\overline{B}) = 0.2$. Furthermore, it is known from prior experience that half of these components cannot withstand a load of 2.5 N but are able to bear a lower load of 2.3 kN (event A). Knowing this, the probability $p_1' = \mathrm{P}(B) = 0.8$ can be updated by testing one of these components up to 2.3 kN.

Let us suppose that the test is successful, i.e. the element does not fail with the 2.3 kN load. The likelihood of event B, i.e. $l_1 \propto \mathrm{P}(A|B) = 1$, and of event $\overline{B}$, i.e. $l_2 \propto \mathrm{P}(A|\bar{B}) = 0.5$, are estimated from the result of this test. Then an *a posteriori* probability follows from relation (2.25),

$$p_1'' = \frac{p_1' l_1}{\sum_{j=1}^{2} p_j' l_j} = \frac{0.80 * 0.10}{0.80 * 1.0 + 0.20 * 0.5} = 0.89$$

Thus the *a priori* probability $p_1' = 0.8$ is updated to the value $p_1'' = 0.89$. The updating of probabilities can now be repeated by another test where the *a posteriori*

probability obtained in the previous step will be considered as *a priori* information. If the other test is also successful, then the new *a posteriori* probability will be

$$p_1'' = \frac{p_1' l_1}{\sum_{j=1}^{2} p_j' l_j} = \frac{0.89 * 1.0}{0.89 * 1.0 + 0.11 * 0.5} = 0.94$$

This repetitive procedure of updating probabilities is quite characteristic in practical applications.

However, what happens when the first test is not successful? If the likelihoods l_1 and l_2 are estimated in this case as $l_1 \propto \mathrm{P}(A|B) = 0.5$ and $l_2 \propto \mathrm{P}(A|\bar{B}) = 1.0$, it follows for the *a posteriori* probability p_1''

$$p_1'' = \frac{p_1' l_1}{\sum_{j=1}^{2} p_j' l_j} = \frac{0.80 * 0.5}{0.80 * 0.5 + 0.20 * 1.0} = 0.67$$

which is an unfavourable reduction of the original (*a priori*) value $p_1' = 0.8$. In such a case it may be useful to carry out additional tests and repeat the updating.

2.8 Bayesian Networks

The Bayesian (causal) networks represent important extensions of Bayes' theorem more and are increasingly used in a number of different fields. Several software tolls have recently been developed to analyse The Bayes' (for example Hugin, GeNie, [3] both of which are available on the internet).

The term "Bayesian networks" was coined by Judea Pearl in 1985 [5, 6] to emphasize three aspects:

- The often subjective nature of the input information.
- The reliance on Bayes' conditioning as the basis for updating information.
- The distinction between causal and evidential modes of reasoning, which underscores Thomas Bayes' famous paper of 1763.

Acyclic directed graphs in which nodes represent random variables and arcs (arrows) show the direct probabilistic dependencies among them (causal links). In addition to the chance nodes, the Bayesian network may include deterministic nodes, decision nodes and utility (value) nodes.

- Chance nodes, drawn usually as ovals, denote random variables that may be described by discrete (in some cases also by continuous) distribution.
- Deterministic nodes, drawn usually as double ovals, denote deterministic variables.

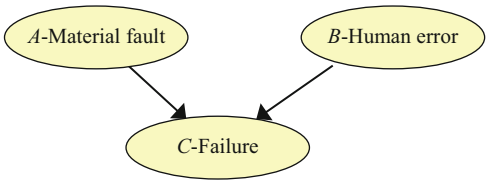


Fig. 2.7 Bayesian network describing failure of a system

– Decision nodes, drawn as rectangles, denote variables that are under the decision-maker's control and are used to model the decision-maker's options.
– Utility nodes (also called Value nodes), drawn usually as hexagons, denote variables that contain information about the decision-maker's goals and objectives. They express the decision-maker's preference for a particular outcome, rather than those outcomes which may have directly preceded it.

The theory of Bayesian network evaluation is quite an extensive topic that is outside the scope of this introductory text. Further information may be found in publications [5–7]. In general, the calculation of probabilities is based on the concept of conditional probabilities and on the theorem of total probability (Sect. 2.5). Consider the simple network of three chance nodes shown in Fig. 2.7. It describes the failure of a system (chance node C) assumed to be caused by material faults (chance node A) and human error (chance node B).

The probability distribution of chance node C follows from the generalised theorem of total probability (Sect. 2.5) as

$$\mathrm{P}(C_k) = \sum\nolimits_{i,j} \mathrm{P}(C_k|A_iB_j)P(A_i)P(B_j) \tag{2.26}$$

Here the subscripts k, i and j denote the states of the chance nodes (two or more). Other procedures of Bayesian network analysis (probability updating, introducing evidence) are provided in special publications [5–7] and software products [3]. The following example illustrates only the main procedure of network evaluation indicated by Eq. (2.26).

Example 2.12. Consider the example indicated in Fig. 2.7. The input data of the network consists of the initial probabilities of the parent's nodes A and B, and the conditional probabilities of the child node C. The three nodes have two alternative states only: negative (fault, error, failure) and positive (no fault, no error, safe). The following two tables show the initial probabilities of nodes A and B, and the conditional probabilities of the child node C.

Initial probabilities of chance nodes A and B

Node A		Node B	
A_1 – Fault	0.05	B_1 – Fault	0.10
A_2 – No fault	0.95	B_2 – No fault	0.90

Conditional probabilities of chance node C (C_1 – Failure, C_2 – Safe)

Node A	A_1 – Fault		A_2 – No fault	
Node B	B_1 – Error	B_2 – No error	B_1 – Error	B_2 – No error
$P(C_1\|A,B)$	0.5	0.1	0.1	0.01
$P(C_2\|A,B)$	0.50	0.9	0.9	0.99

The resulting probabilities of node C follows from Eq. (2.26) as

$$\begin{aligned} P(C_1) &= P(C_1|A_1,B_1)P(A_1)P(B_1) + P(C_1|A_1,B_2)P(A_1)P(B_2) \\ &+ P(C_1|A_2,B_1)P(A_2)P(B_1) + P(C_1|A_2,B_2)P(A_2)\ P(B_2) = 0.025 \end{aligned}$$

$$\begin{aligned} P(C_2) &= P(C_2|A_1,B_1)P(A_1)P(B_1) + P(C_2|A_1,B_2)P(A_1)P(B_2) \\ &+ P(C_2|A_2,B_1)P(A_2)P(B_1) + P(C_2|A_2,B_2)P(A_2)\ P(B_2) = 0.975 \end{aligned}$$

Thus the failure probability of the system is $P(C_1) = 0.025$; the complementary safe state has a probability of $P(C_2) = 0.975$.

Example 2.13. Figure 2.8 shows an example of an influential diagram describing a structure under persistent and fire design situations.

The influential diagram in Fig. 2.8 contains seven chance nodes of oval shape (nodes number 1,2,3,4,5,12,14), four decision nodes of rectangular shape (6,5,15,16) and six utility nodes of diamond shape (8,9,10,11,13,17).

Note that each decision affects the state of the utility nodes and at the same time may generate some costs. For example the decision concerning the sprinklers (decision node 6) affects the state of chance node 2 (extent of the sprinklers) and at the same time generates some additional costs (utility node 8) caused by the installation of sprinklers (which may represent a considerable investment). The sprinklers (chance node 2) may further affect the amount of smoke (chance node 12), the development of any fire (chance node 3) having at least two states: fire stop and fire flashover.

Utility nodes 8, 10 and 17 may be described by the amount of money needed; the other utility nodes (nodes 9, 11 and 13) may also include social consequences, such as injury and loss of human life. Then, however, there is the problem of finding a common unit for economic and social consequences. In some cases these two aspects of consequence are combined, and include some compensation costs for loss of human life; in other cases they are systematically separated for ethical reasons.

Thus the influential diagram offers both the probabilistic analysis and the utility aspect of the system. In such a way the influential diagram is a powerful tool for investigating any system, and for providing rational background information for decision-makers.

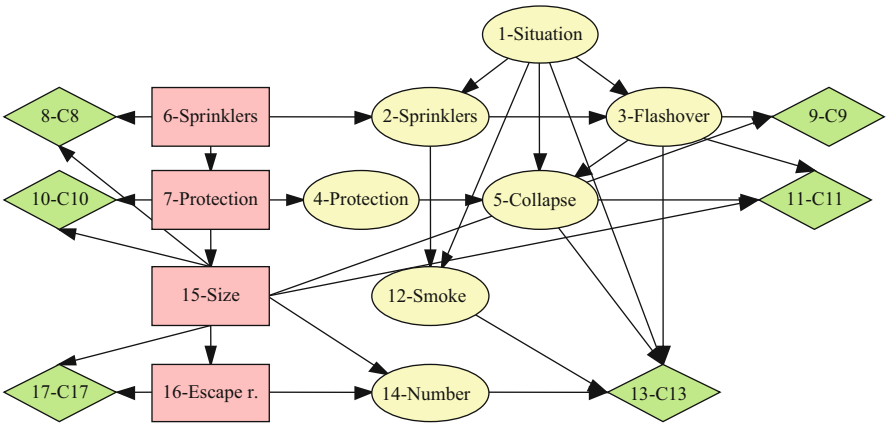


Fig. 2.8 Bayesian network describing a structure designed to withstand fire

2.9 Notes on the Fuzzy Concept

Present methods for analysing uncertainties in engineering and science also include the application of fuzzy sets and fuzzy logic. The concept of fuzzy sets is an extension of the classical crisp set theory. It was introduced by L.A. Zadeh in 1965 [4]. Its application in engineering was indicated in later publications [8, 9]. Fuzzy concepts are often regarded as concepts which, in their application, are neither completely true nor completely false, or which are partly true and partly false. In mathematics and statistics, a fuzzy variable (such as "the temperature", "performance" or "serviceability") is a value which could lie within a certain range defined by quantitative limits, and can be usefully described verbally with imprecise categories (such as "high", "medium" or "low").

The difference between the classical "crisp" set and a fuzzy set can be described as follows. In classical crisp set theory any element in the universe is either a member of a given set or not. An indicator function attains just two values 1 or 0. In a fuzzy set the indicator may attain any value in the interval $<0,1>$ and is termed by Zadeh [4] the membership function. If the membership is 1, then the element is definitely a member of the set; if the membership is 0, then the element is definitely not a member.

The basic properties of membership function are indicated in Fig. 2.9. Point x is a full member of fuzzy set A, point x' is partly a member of A, and $x^{¨}$ is not a member of A. In a mathematical way, fuzzy set A is symbolically represented by a pair (A,ν) where A is the fuzzy set and ν is the mapping $A \rightarrow <0,1>$ of a set A to the interval $<0,1>$. The mapping ν is called membership function.

The following example illustrates a possible application of the fuzzy concept in engineering. The example indicates an analysis of vagueness or imprecision in the definition of structural performance, as described in detail in paper [10].

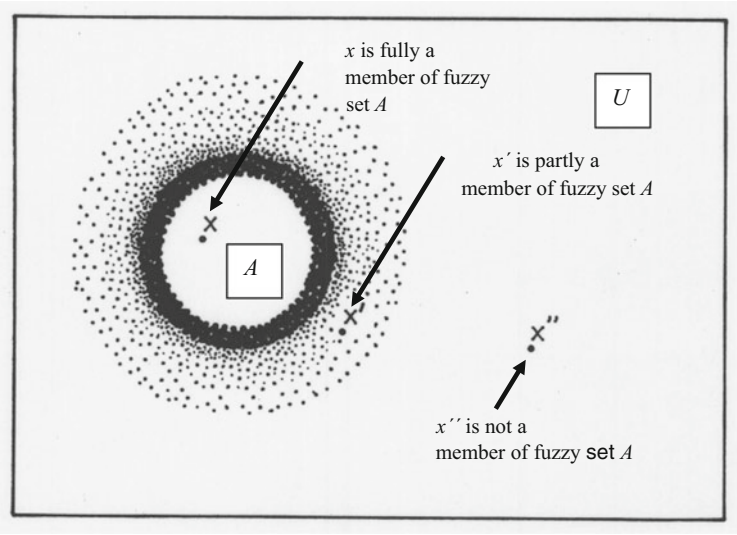


Fig. 2.9 A fuzzy set A

Example 2.14. Fuzziness due to vagueness and imprecision in the definition of performance requirement R is described by the membership function $\nu_R(x)$ indicating the degree of a structure's membership in a fuzzy set of damaged (unserviceable) structures; here x denotes a generic point of a relevant performance indicator (a deflection or a root mean square of acceleration) considered when assessing structural performance. Common experience indicates that a structure is losing its ability to comply with specified requirements gradually and within a certain transition interval $<r_1, r_2>$.

The membership function $\nu_R(x)$ describes the degree of structural damage (lack of functionality). If the rate (the first derivative) $\mathrm{d}\nu_R(x)/\mathrm{d}x$ of the "performance damage" in the interval $<r_1, r_2>$ is constant (a conceivable assumption), then the membership function $\nu_R(x)$ has a piecewise linear form as shown in Fig. 2.10. It should be emphasized that $\nu_R(x)$ describes the non-random (deterministic) part of uncertainty in the performance requirement R related to economic and other consequences of inadequate performance. In addition, the randomness of requirement R at any damage level $\nu = \nu_R(x)$ may be described by the probability density function $\varphi_R(x|\nu)$ (see Fig. 2.10), for which a normal distribution, having the constant coefficient of variation $V_R = 0.10$, is commonly assumed.

The fuzzy probabilistic measure of structural performance is defined [10] by the damage function $\Phi_R(x)$ given as the weighted average of damage probabilities reduced by the corresponding damage level (some theoretical aspects of the example are clarified later in this book, details of the theoretical development are given in [10]). Applied terms and additional details may be found in documents [11–14].

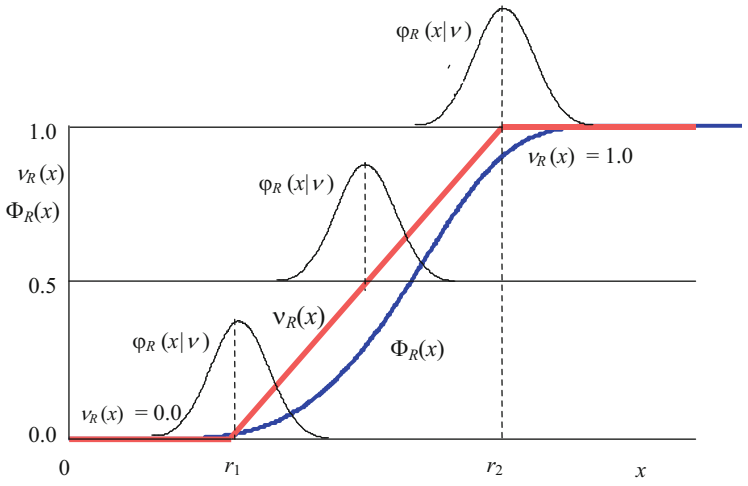


Fig. 2.10 A membership function $\nu_R(x)$ of fuzzy structures R

$$\Phi_R(x) = \frac{1}{N} \int_0^1 \nu \left(\int_{-\infty}^{x} \varphi_R(x'|\nu) \mathrm{d}x' \right) \mathrm{d}\nu$$

Here N denotes a factor normalizing the damage function $\Phi_R(x)$ to the conventional interval $<0, 1>$ (see Fig. 2.10) and x' is a generic point of x. The density of the damage $\varphi_R(x)$ follows from (1) as

$$\phi_R(x) = \frac{1}{N} \int_0^1 \nu \ \phi_R(x|\nu) \ \mathrm{d}\nu$$

The damage function $\Phi_R(x)$ and density function $\varphi_R(x)$ may be considered as a generalized fuzzy-probability distribution of the performance requirement, R, one that is derived from the fuzzy concept and the membership function $\nu_R(x)$ and can be used similarly as classical probability distributions.

References

1. Ang, A.H.-S., Tang, W.H.: Probabilistic Concepts in Engineering. Emphasis on Applications to Civil Environmental Engineering. Wiley, New York (2007)
2. Devore, J., Farnum, N.: Applied Statistics for Engineers and Scientists. Thomson, London (2005)
3. Hugin System: Version 5.7, professional. Hugin Expert A/S, Aalborg, Denmark. (see also software product GeNie, http//:genie.sis.pitt.edu, (2010))

4. Zadeh, L.A.: Fuzzy sets. Info. Control **8**(3), 338–353 (1965)
5. Pearl, J.: Fusion, propagation, and structuring in belief networks. Artif. Intell. (Elsevier) **29**(3), 241–288 (1986)
6. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Representation and Reasoning Series, 2nd printing edn. Morgan Kaufmann, San Francisco (1988)
7. Jensen, F.V.: Introduction to Bayesian Networks. Aalborg University, Aalborg, Denmark (1996)
8. Brown, C.B., Yao, J.T.P.: Fuzzy sets and structural engineering. J. Struct. Eng. **109**(5), 1211–1225 (1983)
9. Blockley, D.I.: The Natire of Structural Design. Ellis Horwood Limited, Chichester (1980)
10. Holický, M.: Fuzzy probabilistic optimisation of building performance. Autom. Constr. **8**(4), 437–443 (1999). ISSN 0926/5805
11. ISO 12491: Statistical Methods for Quality Control of Building Materials and Components. ISO, Geneve (1997)
12. ISO 3534-1: Statistics – Vocabulary and Symbols – Part 1: Probability and General Statistical Terms. ISO, Geneve (1993)
13. ISO 3534-2: Statistics – Vocabulary and Symbols – Part 2: Statistical Quality Control. ISO, Geneve (1993)
14. Holický, M.: Reliability Analysis for Structural Design. SUNN MeDIA, Stellenbosch (2009)

Chapter 3
Evaluation of Statistical Data

The evaluation of statistical data representing a random sample taken from a particular population is frequently moment the first step in many engineering and scientific tasks. The concept of a general population and the random samples taken from it is introduced and supplemented by the definition of commonly used sample characteristics. Emphasis is put on the characteristics, summarized in Appendix 1, that usually provide the initial background information for the specification of a theoretical model of population. Sample characteristics regularly used in engineering and science describe the location, dispersion, asymmetry and kurtosis of statistical data. The general rules and computational techniques used for determining sample characteristics of a single random sample, and also for the combination of two random samples, are illustrated by examples.

3.1 Population and Random Sample

The concepts of population and random sample are extremely important for the appropriate interpretation of statistical data and their analysis. Population, or "the universe", is the totality of items under consideration. A population may be finite (N sampling units) or infinite. Rather than examining the entire group of N units a small part of the population, that is a sample of n units, may be examined instead. A precise definition regarding a population is often difficult to come by, but must be provided in order to interpret outcomes of statistical investigation correctly [1, 2]. An excellent description of the basic technique is given in [3, 4] and a short review is provided in [5]. The correct terminology and procedures are available in International Standards [6–8].

A sample is one or more units taken from a population and is intended to provide information on that population. It may serve as a basis for decision-making about the population, or about the process which produced it. The term "random sample" refers to the samples that are taken from a population in such a way that all possible units have the same probability of being taken. The number of sampling units,

M. Holický, *Introduction to Probability and Statistics for Engineers*,
DOI 10.1007/978-3-642-38300-7_3, © Springer-Verlag Berlin Heidelberg 2013

called sample size n, may be considerably different. Commonly, samples are considered to be very small ($n < 10$), small ($n < 30$), large ($n > 30$) or very large ($n > 100$). Obviously, with increasing size the samples become more representative. However, the sampling procedure is equally important.

If a sample is representative of a population, important conclusions about it can often be inferred from an analysis of the sample. This phase of statistics is called inductive statistics, or statistical inference, and is covered in subsequent chapters. The phase of statistics that seeks only to describe and analyse a given sample is called descriptive, or deductive, statistics to which is devoted this Chapter.

Example 3.1. A factory produces 70 units of the same type. A random sample of 10 units can be taken from the population of 70 units using a table, or a generator of random numbers within a range of 1–70. A sample can then be created by taking the units whose serial numbers are equal to ten generated random numbers.

3.2 Characteristics of Location

The basic characteristic of sample location (or its main tendency) is the sample mean m_X given as

$$m_X = \frac{1}{n} \sum_1^n x_i \tag{3.1}$$

Here x_i denotes sample units. If the sample units are ordered from the smallest to greatest unit then the subscripts i are generally changed to (i), and the units are then denoted $x_{(i)}$.

Another characteristic of location is median defined the point separating ordered sequence of data into two parts such that half of the data is less than the median and half of the data greater than the median.

Example 3.2. A random sample of measurements of concrete strength contains ten measurements $x_i = \{27; 30; 33; 29; 30; 31; 26; 38; 35; 32\}$ in MPa. The measured data, in order of scale, is $x_{(i)} = \{26; 27; 29; 30; 30; 31; 32; 33; 35; 38\}$ in MPa:

The sample mean m_X and the median $\tilde{m}_X$ are given as

$$m_X = \frac{1}{10}\left(\sum x_i\right) = 31.1\,\text{MPa}, \quad \tilde{m}_X = \frac{1}{2}\left(x_{(5)} + x_{(6)}\right) = 30.5\,\text{MPa}$$

3.3 Characteristics of Dispersion

The basic characteristic of dispersion is called the variance

$$s_X^2 = \frac{1}{n}\sum_1^n (x_i - m_X)^2 \tag{3.2}$$

In practical applications the standard deviation s_X is commonly used instead of "variance".

Another measure of dispersion that is frequently applied in engineering and science is called the coefficient of variation

$$v_X = \frac{s_X}{m_X} \tag{3.3}$$

This is, in fact, a measure of relative dispersion normalised by the sample mean m_X. It is frequently used in engineering when the sample mean m_X is not very small. If the sample mean m_X is relatively small then the standard deviation should be used instead.

In the case of very small samples ($n \leq 10$) additional measure of dispersion, called sample range, is sometimes used; it is defined simply as the difference between of the greatest and smallest sample unit, $x_{(n)} - x_{(1)}$.

In same specific cases also the mean deviation MD, or average deviation, defined as the mean of differences $|x_i - m_X|$ is also used

$$MD_X = \frac{1}{n}\sum_{i=1}^n |x_i - m_X| \tag{3.4}$$

Example 3.3. The variance of the sample given in Example 3.1 $x_i = \{27; 30; 33; 29; 30; 31; 26; 38; 35; 32\}$in MPa is given as

$$s_X^2 = \frac{1}{n}\sum_1^n (x_i - m_X)^2 = 11.69(\text{MPa})^2$$

The standard deviation is thus

$$s_X = \sqrt{s_X^2} = \sqrt{11.69} = 3.42\,\text{MPa}$$

Example 3.4. The coefficient of variation of the data in the random sample given in Example 3.2 $x_i = \{27; 30; 33; 29; 30; 31; 26; 38; 35; 32\}$ in MPa, is given as

$$v_X = \frac{3.42}{31.1} = 0.11 = 11\%$$

Example 3.5. Considering ordered measurements from Example 3.2 $x_{(i)} = \{26; 27; 29; 30; 30; 31; 32; 33; 35; 38\}$in MPa, the variation range and the mean deviations are:

$$x_{(n)} - x_{(1)} = 38 - 26 = 12 \text{ MPa}$$

$$MD_X = \frac{1}{n} \sum_{i=1}^{n} |x_i - m_X| = 2.72 \text{ MPa}$$

3.4 Characteristics of Asymmetry and Kurtosis

The characteristics of asymmetry and peakedness (kurtosis) are used less frequently than the characteristics of location (the mean m_X) and the characteristic of dispersion (the variance s_X^2). However, the characteristics of asymmetry and peakedness provide valuable information about the nature of the sample, in particular the distribution of observation to the left and right of the mean and the concentration of observation about the mean. This information, concerning in particular the skewness, may be extremely useful for determining the appropriate theoretical model (probability distribution) of population.

The following moment characteristics are most often used. The coefficient of asymmetry is defined on the basis of the central moment of the third order as

$$a_X = \frac{1}{n s_X^3} \sum_{i=1}^{n} (x_i - m_X)^3 \tag{3.5}$$

Similarly the coefficient of kurtosis is related to the central moment of the fourth order as

$$e_X = \frac{1}{n s_X^4} \sum_{i=1}^{n} (x_i - m_X)^4 - 3 \tag{3.6}$$

Note that the above defined coefficients of asymmetry and kurtosis should be close to zero for samples taken from population having normal distribution.

The coefficient of asymmetry is positive when more sample data is on the left of the mean, positive when more data is on the right of the mean. The coefficient of kurtosis is positive when the sample data is located mostly in the vicinity of the mean, negative when the data is distributed more uniformly. Both these characteristics (skewness a_X and kurtosis e_X) are strongly dependent on abnormal deviations of some sample units (outliers), or errors, particularly in the case of small samples ($n < 30$). Then their evaluation may be highly uncertain (and may suffer from so-called statistical uncertainty due to limited data).

Example 3.6. Considering again data from Example 3.2 given as $x_i = \{27; 30; 33; 29; 30; 31; 26; 38; 35; 32\}$ in MPa, the coefficients of asymmetry and kurtosis are:

$$a_X = \frac{1}{ns_X^3} \sum_{i=1}^{n} (x_i - m_X)^3 = 0.46$$

$$e_X = \frac{1}{ns_X^4} \sum_{i=1}^{n} (x_i - m_X)^4 - 3 = -0.44$$

The positive coefficient of asymmetry indicates that more observations are on the left of the mean (in fact 6 of 10 values are on the left of the mean). A slightly negative coefficient of kurtosis indicates low peakedness (observed values seem to be distributed slightly more uniformly than those of normal distribution). Note that the investigated sample is very small (10 values only), and the coefficients obtained, a_X and e_X may be inaccurate.

It is interesting to note that there is an empirical relationship between the skewness a_X the mean m_X, the median $\tilde{m}_X$ and the standard deviation s_X (called sometimes as Pearson coefficient of skewness) in the form

$$a_X \approx 3(m_X - \tilde{m}_X)/s_X{}^{\circ}$$

Considering the results of previous Examples 3.2 and 3.3 $m_X = 31.1\,\text{MPa}$, $\tilde{m}_X = 30.5\,\text{MPa}$, $s_X = 3.42\,\text{MPa}$ and it follows that

$$a_X \approx \frac{3(31.1 - 30.5)}{3.42} = 0.53$$

This seems to be a good approximation of the above obtained moment skewness $a_X = 0.46$. It also demonstrates the intuitively expected result that if the median $\tilde{m}_X$ is less than the mean m_X, then the skewness a_X is positive. Consequently more data is located left of the mean than right of the mean.

3.5 General and Central Moments

Most of the samples characteristics described above belong to so called moment characteristics that are based on general or central moments of the data. The general moment (about the origin) of the order l ($l = 1, 2, 3, \ldots$) is defined as the arithmetic mean of the sum of l-powers

$$m_l^* = \frac{1}{n} \sum_{i=1}^{n} x_i^l \tag{3.7}$$

The central moment (about the mean) of the order l is similarly given as

$$m_l = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_X)^l \tag{3.8}$$

The moment characteristics can be then defined as follows.

$$m_X = m_1^* \tag{3.9}$$

$$s_X = \sqrt{m_2} \tag{3.10}$$

$$a_X = \frac{m_3}{m_2^{3/2}} \tag{3.11}$$

$$e_X = \frac{m_4}{m_2^2} - 3 \tag{3.12}$$

In numerical calculation it is sometime useful to apply the following relations between the general and central moments

$$m_2 = m_2^* - m_X^2 \tag{3.13}$$

$$m_3 = m_3^* - 3m_X m_2^* + 2m_X^3 \tag{3.14}$$

$$m_4 = m_4^* - 4m_X m_3^* + 4m_X^2 m_2^* - 3m_X^4 \tag{3.15}$$

When computers are used to evaluate statistical samples Eqs.. (3.13, 3.14, and 3.15) are not directly used.

3.6 Combination of Two Random Samples

Sometimes it is necessary to combine two random samples taken from one population, assuming that the characteristics of both the samples are known, but the original observations x_i are not available. It must be emphasised that only homogeneous samples of the same origin (taken from one population under the same conditions) should be combined. Violation of this important assumption could lead to incorrect results.

Assume that a first sample of the size n_1 has the characteristics m_1, s_1, a_1, while a second sample of the size n_2 has the characteristics m_2, s_2, a_2. Only three basic characteristics are considered here (the coefficients of kurtosis are rarely available for combined samples). The resulting characteristics of a combined sample of the size n can be determined from the following expressions:

$$n = n_1 + n_2 \tag{3.16}$$

$$m = \frac{n_1 m_1 + n_2 m_2}{n} \tag{3.17}$$

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n} + \frac{n_1 n_2}{n^2}(m_1 - m_2)^2 \tag{3.18}$$

$$a = \frac{1}{s^3} \times \left[\frac{n_1 s_1^3 a_1 + n_2 s_2^3 a_2}{n} + \frac{3 n_1 n_2 (m_1 - m_2)\left(s_1^2 - s_2^2\right)}{n^2} - \frac{n_1 n_2 (n_1 - n_2)(m_1 - m_2)^3}{n^2} \right] \tag{3.19}$$

It is interesting to note that the standard deviation s is dependent not only on the standard deviations of two initial samples s_1 and s_2, but also on the means of both the samples. Similarly, the skewness a also depends on the characteristics of the lower order (means and standard deviations). The relationship for the kurtosis is not included as it is not commonly used.

It should be noted that if the original data is available then it can be analysed as one sample; relationships (3.16, 3.17, 3.18, and 3.19) can then be used for checking newly obtained results. The most important thing is the verification of the hypothesis that both samples are taken from one population.

Example 3.7. An example of the practical application of Eqs. (3.16, 3.17, 3.18, and 3.19) is shown underneath.

Samples	n	m	s	a	v
Sample 1	10	30.1	4.4	0.5	0.15
Sample 2	15	29.2	4.1	0.5	0.14
Combined	25	29.56	4.25	0.53	0.14

Note that a different number of sample units may affect the characteristics of the resulting combined sample. An EXCEL sheet has been developed for calculation if this is the case.

Sometimes it may occur that the size of one sample, say n_1, is not known, and only the first two characteristics m_1, s_1 are available. This is a typical situation when updating previous data with the characteristics m_1, s_1, using newly observed data of the size n_2 with the characteristics m_2, s_2. Then the Bayesian approach may be used for assessing the unknown value n_1 and a corresponding degree of freedom ν_1. The following text is presented here as a guide on how to proceed in that case, just for information and without the appropriate mathematical clarification.

In accordance with the Bayesian concept [1, 3], the unknown value n_1 and a corresponding degree of freedom ν_1 may be assessed using the relations for the coefficients of variation of the mean and standard deviation $V(\mu)$ and $V(\sigma)$, (the

parameters μ and σ are considered as random variables in Bayes' concept) for which it holds

$$n_1 = [s_1/(m_1 V(\mu))]^2, \quad \nu_1 = 1/\left(2V(\sigma)^2\right) \tag{3.20}$$

Both unknown variables n_1 and ν_1 may be assessed independently (generally $\nu_1 \neq n_1 - 1$), depending on previous experience with a degree of uncertainty of the estimator of the mean μ and the standard deviation σ of the population. Note that for a new sample it holds that $\nu_2 = n_2 - 1$.

When the sample size n_1 and the degree of freedom ν_1 are estimated, the degree of freedom ν is given as [3]

$$\nu = \nu_1 + \nu_2 - 1 \quad \text{if} \quad n_1 \geq 1, \quad \nu = \nu_1 + \nu_2 \quad \text{if} \quad n_1 = 0 \tag{3.21}$$

Then the resulting size of the combined sample n and the mean m is given by Eqs. (3.16) and (3.17); the standard deviation s is determined from a modified Eq. (3.18) as

$$s^2 = \left[\nu_1 s_1^2 + \nu_2 s_2^2 + \frac{n_1 n_2}{n}(m_1 - m_2)^2\right]/\nu \tag{3.22}$$

The above relationship may be easily applied using the EXCEL sheet or other software tools.

Example 3.8. Suppose that from the prior production of a given type of concrete the following information is available regarding its strength

$$m_1 = 30.1\,\text{MPa}, \quad V(\mu) = 0.50, \quad s_1 = 4.4\ \text{MPa}, \quad V(\sigma) = 0.28.$$

For the unknown characteristics n_1 and ν_1 it follows from Eq. (3.20) that

$$n_1 = \left(\frac{4.4}{30.1}\frac{1}{0.50}\right)^2 \approx 0, \quad \nu_1 = \frac{1}{2 \times 0.28^2} \approx 6$$

Thus, the following characteristics are subsequently considered: $n_1 = 0$ and $\nu_1 = 6$.

To verify the quality of the concrete, new measurements have been carried out using specimens from the same type of concrete. The following strength characteristics have been obtained:

$$n_2 = 5, \quad \nu_2 = n_2 - 1 = 4, \quad m_2 = 29.2\ \text{MPa}, \quad s_2 = 4.6\ \text{MPa}.$$

Using Eqs. (3.16, 3.17, 3.18 and 3.19), the updated characteristics are as follows:

$$n = 0 + 5 = 5$$

$$\nu = 6 + 4 = 10$$

$$m = \frac{0 \times 30.1 + 5 \times 29.2}{5} = 29.2 \text{ MPa}$$

$$s^2 = \left[6 \times 4.4^2 + 4 \times 5.6^2 + \frac{0 \times 5}{5}(30.1 - 29.2)^2\right]/10 = 4.5^2 \text{MPa}^2$$

Thus, using the previous information, the standard deviation of the new measurements could be decreased from s = 5.6 MPa to s = 4.5 MPa.

However, it should be noted that the combination of the previous information with the current measurements might not always lead to favourable results. For example, if the coefficients of variation are $w(\mu) = 0.2$ and $w(\sigma) = 0.6$, then the unknown characteristics n_1 and ν_1 follow from Eq. (3.20) as

$$n_1 = \left(\frac{4.4}{30.1}\frac{1}{0.2}\right)^2 \approx 1; \quad \nu_1 = \frac{1}{2 \times 0.6^2} \approx 1$$

In this case

$$n = 1 + 5 = 6$$

$$\nu = 1 + 4 - 1 = 4$$

$$m = \frac{1 \times 30.1 + 5 \times 29.2}{6} = 29.35 \text{ MPa}$$

$$s^2 = \left[1 \times 4.4^2 + 4 \times 5.6^2 + \frac{1 \times 5}{6}(30.1 - 29.2)^2\right]/4 = 6.03^2 \text{MPa}^2$$

In this case, the mean increased slightly from 29.2 to 29.35, while the standard deviation increased considerably, from 5.6 to 6.03. However, this is an extreme case, caused by unfavourable estimates of n_1, ν_1 and ν following on from Eqs. (3.20) and (3.21). In practical applications these equations should be applied with caution, particularly in extreme cases similar to the above example. In connection with this warning, an important assumption mentioned at the beginning of this section should be stressed. Only those samples that are evidently taken from the same population can be used for combining or updating statistical data; otherwise the results of the combination of two random samples may lead to incorrect results. On the other hand when two or more samples are taken from one population then their combination is always valuable.

3.7 Note on Terminology and Software Products

It should be mentioned that documents such as ISO 3534 and software products EXCEL, MATHCAD and STATISTICA provide slightly modified terminology and definitions for basic moment characteristics.

In general two modifications are commonly used for the characteristic of dispersion:

– The characteristic called here "the sample standard deviation" is also denoted as "the standard deviation of a sample", or as "the population standard deviation" (when n is the population size), and is given as

$$s_X = \sqrt{\frac{1}{n}\sum\nolimits_1^n (x_i - m_X)^2} \tag{3.23}$$

– The sample estimate of the population standard deviation called here a point estimate of the population standard deviation and denoted by the symbol $\hat{s}_X$ (see also Chap. 8) is sometimes called the sample standard deviation

$$\hat{s}_X = \sqrt{\frac{1}{n-1}\sum\nolimits_1^n (x_i - m_X)^2} \tag{3.24}$$

Expression (3.23) corresponds to Eq. (3.2) for the sample standard deviation. Expression (3.24) represents a point estimate of standard deviation that is derived from the mean of the distribution describing the sample variance (based on the χ^2 random variable and discussed in Chap. 8).

Similar modifications of sample characteristics are also available for the skewness and kurtosis. The "sample skewness" a defined here by Eq. (3.5) can be written in simplified form as

$$a_X = \frac{m_3}{m_2^{3/2}} = \frac{1}{n s_X^3}\sum\nolimits_{i=1}^n (x_i - m_X)^3 \tag{3.25}$$

STATISTICA, EXCEL, MATHCAD and some other software products provide a point estimate of the population skewness $\hat{a}_X$ (see Chap. 8) as

$$\hat{a}_X = \frac{n^2}{(n-1)(n-2)}\frac{1}{n\hat{s}_X{}^3}\sum\nolimits_{i=1}^n (x_i - m_X)^3 = \frac{\sqrt{n(n-1)}}{(n-2)}a_X \tag{3.26}$$

Note that the population estimate $\hat{s}_X$ is used in Eq. (3.26). If the sample standard deviation is used then the estimate of the population skewness would be

$$\hat{a}_X = \frac{n}{(n-1)(n-2)} \frac{1}{s_X^3} \sum_{i=1}^{n} (x_i - m_X)^3 = \frac{n^2}{(n-1)(n-2)} a_X \tag{3.27}$$

The factor enhancing the sample skewness a_X in Eq. (3.27) (the fraction containing the sample size n) is slightly greater than the similar factor in Eq. (3.26) (for $n > 30$ by less than 5 %); the difference diminish with increasing sample size n

Similar modifications of sample characteristics may be found for kurtosis based on the central moment of the fourth order (see Eq. (3.6)). The relevant formulae can be found in the help component of the relevant software products. However, kurtosis is evaluated in practical applications very rarely and only for very large samples ($n > 100$).

3.8 Grouped Data, Histogram

When analyzing large size of statistical data n, it is often useful to group them into a limited number of classes k (usually $7 \leq k \leq 20$) and to determine the number of units belonging to each class n_i ($i = 1,2,...k$), called class frequency ($\Sigma n_i = n$). Each class is represented by class mark x_i^* which is the midpoint of the class interval limited by its lower and upper class limit.

Commonly, the grouped data are presented graphically in the form of a histogram, which is a column diagram showing frequency n_i or relative frequency n_i/n for each class. Histograms are very useful graphical tools providing valuable information about the overall character of the sample. Visual investigation of the histogram is always recommended. It may provide an initial understanding of the sample nature.

The mean m_X is given by the general moment of the first order Eq. (3.7), which for grouped data is written as

$$m_X = m_1^* = \frac{1}{n} \sum_{i=1}^{k} n_i x_i^* \tag{3.28}$$

The central moments (about the mean) of the order l are for grouped data given as

$$m_l = \frac{1}{n} \sum_{i=1}^{k} n_i (x_i^* - m_X)^l \tag{3.29}$$

The moment characteristics of grouped data can be determined using the general formulae (3.10, 3.11, and 3.12). Also the relationships between the general and central moments provided by Eqs. (3.13, 3.14, and 3.15) can be used in the numerical evaluation of grouped data.

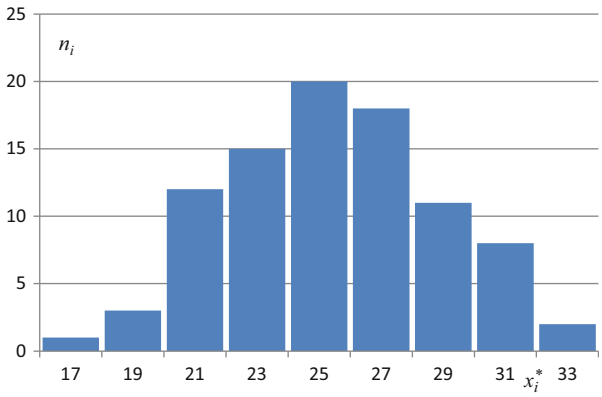


Fig. 3.1 Histogram of the grouped data form Example 3.9 (90 observations of concrete strength)

Example 3.9. Results of $n = 90$ tests of concrete strength are grouped into $k = 9$ classes as indicated in the table below and in the histogram in Fig. 3.1. Visual investigation of the histogram indicates that the sample is well-ordered (without outliers), symmetric (the skewness is expected to be close to zero) and slightly less spiky (more flat) than commonly used normal distribution (a bit of negative kurtosis is expected).

Class i	Class interval in MPa	Class mark x_i^* in MPa	Frequency n_i	Product $n_i x_i^*$	Product $n_i(x_i^* - m_X)^2$
1	16–18	17	1	17	71.309
2	18–20	19	3	57	124.593
3	20–22	21	12	252	237.037
4	22–24	23	15	345	89.630
5	24–26	25	20	500	3.951
6	26–28	27	18	486	43.556
7	28–30	29	11	319	139.062
8	30–32	31	8	248	246.914
9	32–34	33	2	66	114.173
Sum	–	–	90	2,290	1,070.222

The table shows the class intervals, class marks x_i^* (in MPa), frequency n_i and products $n_i x_i^*$ and $n_i(x_i^* - m_X)^2$ used to calculate the general moments of the first order, and the central moment of the second order. The moments of the order 3 and 4 would be necessary for calculation of the skewness a_X and kurtosis e_X.

It follows from Eqs. (3.7) and (3.10) and the numerical results shown in the last row of the above table that the sample mean and standard deviation are

$$m_X = 2290/90 = 25.44\,\text{MPa and } s_X = (m_2)^{0.5} = (1070.222/90)^{0.5} = 3.45\,\text{MPa}$$

The coefficient of variation $v_X = 3.45/25.44 \approx 0.14$ is relatively high and indicates a somewhat low quality of material. The other moment characteristics can be similarly found using the central moments of higher order and general Eqs. (3.11) and (3.12). This way it can be found that the sample skewness is almost zero, $a = 0.03$, and the kurtosis $e = -0.53$. So the sample is really symmetrical and slightly more uniform than the normal distribution.

References

1. Ang, A.H.-S., Tang, W.H.: Probabilistic Concepts in Engineering. Emphasis on Applications to Civil Environmental Engineering. Wiley, New York (2007)
2. Devore, J., Farnum, N.: Applied Statistics for Engineers and Scientists. Thomson, London (2005)
3. Dunin-Barkovskij, I.V., Smirnov, N.V.: The Theory of Probability and Mathematical Statistics in Engineering. (in Russian). Technical and Theoretical Literature, Moscow (1955)
4. Gurskij, E.I.: The Theory Probability with Elements of Mathematical Statistics. (in Russian). Higher School, Moscow (1971)
5. Holicky, M.: Reliability Analysis for Structural Design. SUNN MeDIA, Stellenbosch (2009)
6. ISO 12491: Statistical Methods for Quality Control of Building Materials and Components. ISO, Geneve (1997)
7. ISO 3534-1: Statistics – Vocabulary and Symbols – Part 1: Probability and General Statistical Terms. ISO, Geneve (1993)
8. ISO 3534-2: Statistics – Vocabulary and Symbols – Part 2: Statistical Quality Control. ISO, Geneve (1993)

Chapter 4
Distributions of Random Variables

Two different categories of random variables are commonly used in engineering and scientific applications of probability and statistics: discrete and continuous random variables. Every random variable can be described by distribution function and corresponding probability density function. Commonly used distribution functions are defined by a limited number of parameters. Similarly as in the case of sample characteristics, distribution parameters, summarized in Appendix 1, are used to characterise the location, dispersion, asymmetry and peakedness of a distribution. So called standardized random variables, having the means equal to zero and variances equal to 1, are often applied in numerical methods used to analyse the random properties of engineering and scientific systems.

4.1 Random Variables

Most of the experiments in engineering and science result in random events that can be described by real numbers, for example by the strength of a material, or the content of a specified substance. A set of all these numbers, hypothetically obtained from a given population, form a random variable having a certain probability distribution.

In general, a variable which may take any of the values of a specified set of values, and which is associated with a probability distribution is called a random variable [1, 2]. A comprehensive treatment is provided in [3, 4], a short review in [5]. A correct terminology of basic terms and a description of statistical procedures is provided in standards [6–8].

Two basic types of random variables are recognized. A variable, which may take only isolated values is said to be a "discrete" random variable. A variable which may take any of the values of a specified set of values is called a continuous random variable. These two basic types of random variables are commonly used in engineering and scientific applications.

M. Holický, *Introduction to Probability and Statistics for Engineers*,
DOI 10.1007/978-3-642-38300-7_4, © Springer-Verlag Berlin Heidelberg 2013

4.2 Distribution Function

The distribution function of a random variable X is a function $\Phi(x)$ defined as the probability that X is less than or equal to any real value x of the variable X, thus

$$\Phi(x) = \mathrm{P}(X \leq x) \tag{4.1}$$

Here X denotes a random variable and x any real value. The general properties of distribution function $\Phi(x)$ of a discrete or continuous random variable X follows directly from the definition (4.1):

$$0 \leq \Phi(x) = \leq 1; \quad \Phi(-\infty) = 0; \quad \Phi(\infty) = 1 \tag{4.2}$$

$$\text{If } x_1 \leq x_2, \text{ then } \Phi(x_1) \leq \Phi(x_2) \tag{4.3}$$

$$\mathrm{P}(x_1 < X \leq x_2) = \Phi(x_2) - \Phi(x_1) \tag{4.4}$$

The above Eq. (4.4) illustrates an important relationship between occurrence probability $\mathrm{P}(x_1 \leq X \leq x_2)$ of a random variable X in a given interval $x_1 < X \leq x_2$ and distribution function $\Phi(x)$.

4.3 Discrete Random Variables

A discrete random variable X attains only a countable number of values x_i, say x_1, x_2, x_3, …, for example 0, 1, 2, …. The general form of the distribution function (4.1) is then written as

$$\Phi(x_j) = \mathrm{P}(X \leq x_j) = \sum\nolimits_{x_i \leq x_j} \mathrm{P}(x_i) = \sum\nolimits_{x_i \leq x_j} \mathrm{P}(X = x_i) \tag{4.5}$$

The distribution is fully described by probabilities p_i of individual values x_i

$$\mathrm{P}(x_i) = \mathrm{P}(X = x_i) = p_i \tag{4.6}$$

The distribution function $\Phi(x_j)$ and probabilities p_i of individual values x_i are shown in Fig. 4.1.

Example 4.1. Consider a random variable X that attains the values 1, 2, 3,…, N with a constant probability

$$\mathrm{P}(x) = \mathrm{P}(x = i) = \frac{1}{N}.$$

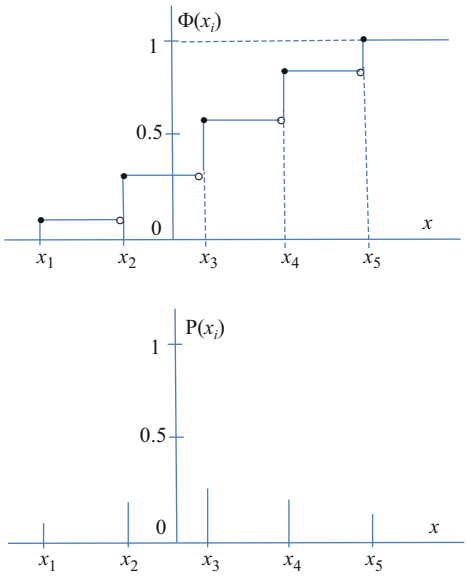


Fig. 4.1 Distribution function and probabilities of a discrete random variable X

Here $i = 1, 2, 3, \ldots, N$. This distribution is called discrete uniform (rectangle) distribution. Its distribution function is

$$\Phi(x) = \frac{i}{N}.$$

Note that if $N = 6$ then the random variable describes the outcomes of throwing a dice, i.e. Numbers 1, 2, 3, 4, 5 and 6.

Example 4.2. A special case of a discrete random variable is the so-called degenerated random variable X that attains only a certain value μ with the probability $\mathrm{P}(x = \mu) = 1$, thus any other value with zero probability. It is, in fact, not a random variable as it attains only one value.

4.4 Continuous Random Variables

A continuous random variable X is fully described by distribution function $\Phi(x)$, for which the full notation $\Phi_X(x)$ is used when necessary, or by probability density function $\varphi(x)$ (the full notation is $\varphi_X(x)$). The distribution function $\Phi(x)$ is the integral of the probability density function $\varphi(x)$, which is a non-negative real function describing the relative frequency of the variable X.

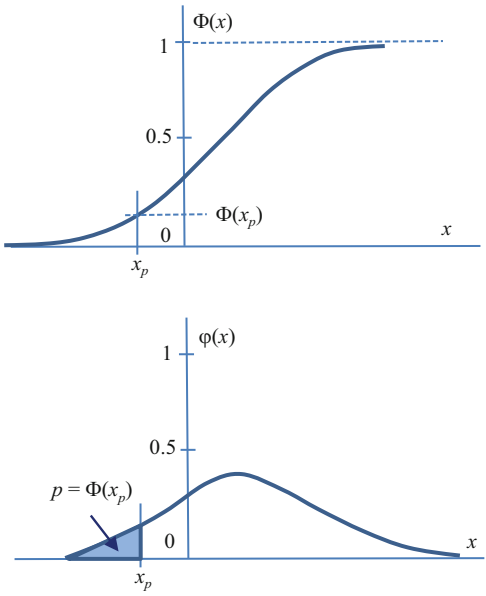


Fig. 4.2 Distribution and probability density function of a continuous random variable X

$$\Phi(x) = \int_{-\infty}^{x} \varphi(x)\mathrm{d}x \tag{4.7}$$

Thus, the probability density function $\varphi(x)$ can be obtained as the derivative of the distribution function (when it exists)

$$\varphi(x) = \frac{\mathrm{d}\Phi(x)}{\mathrm{d}x} \tag{4.8}$$

Their mutual dependence is obvious from Fig. 4.2 (an analogue to Fig. 4.1).

The value x_p, indicated in Fig. 4.2, denotes an important value of the random variable X called a fractile (also called a quantile); it is the value that corresponds to the probability $p = \mathrm{P}(X \leq x_p) = \Phi(x_p)$ that variable X is less than, or equal to, x_p.

It follows from Eqs. (4.4) and (4.7)

$$\mathrm{P}(x_1 < X \leq x_2) = \Phi(x_2) - \Phi(x_1) = \int_{x_1}^{x_2} \varphi(x)\mathrm{d}x \tag{4.9}$$

The above Eq. (4.9) illustrates an important relationship between the distribution function $\Phi(x)$ and probability density function $\varphi(x)$.

Example 4.3. Consider a continuous random variable X having the domain $x\varepsilon<$a, b$>$ and the distribution function (see Fig. 4.3)

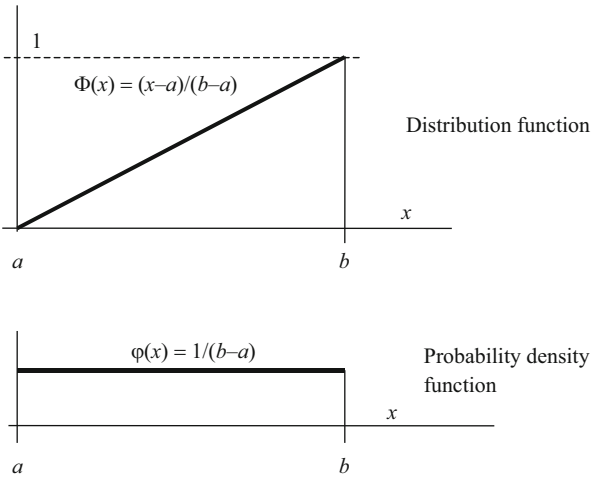


Fig. 4.3 Uniform distribution

$$\Phi(x) = \frac{x-a}{b-a}$$

Then the probability density function follows from Eq. (4.8) as a constant independent of x

$$\varphi(x) = \frac{1}{b-a}$$

This type of distribution of continuous random variables is called uniform or rectangular distribution. It has a number of practical applications in numerical calculations and simulation techniques.

4.5 Parameters of Random Variables

Distribution function and probability density functions are commonly described by distribution parameters. The moment parameters, based on general and central moments, are more often used. The general moments of a discrete and continuous random variable of the order l are defined as follows.

$$\mu_l' = \sum_i x_i^l \mathrm{P}(x_i) \tag{4.10}$$

$$\mu_l' = \int_x x^l \varphi(x)\mathrm{d}x \tag{4.11}$$

The general moment μ_1' of the first order is used to define the mean μ of the random variable as a basic measure of distribution location

$$\mu = \mu_1' \tag{4.12}$$

The central moments of a discrete and continuous random variable of the order l are defined as follows.

$$\mu_l = \sum_i (x_i - \mu)^l \, \mathrm{P}(x_i) \tag{4.13}$$

$$\mu_l = \int_x (x - \mu)^1 \varphi(x) \mathrm{d}x \tag{4.14}$$

The central moment of the first order is obviously zero, $\mu_1 = 0$. The central moments of orders 2, 3 and 4 are used to define other important parameters. The moment of order 2 defines the variance σ^2

$$\sigma^2 = \mu_2 \tag{4.15}$$

The square root of the variance is called the standard deviation σ

$$\sigma = \sqrt{\mu_2} \tag{4.16}$$

Standard deviation is commonly used in all types of applications of mathematical statistics as a basic measure of dispersion. The relative measure of dispersion, used frequently and particularly in engineering, is called the coefficient of variation V. It is defined as a ratio of the standard deviation σ and the mean μ as:

$$V = \frac{\sigma}{\mu} \tag{4.17}$$

Here it is assumed that the mean is not zero, $\mu \neq 0$. If the mean is very small (close to zero), then direct use of the variance or standard deviation is preferable. Nevertheless, the coefficient of variation V is frequently used as an important measure of relative dispersion that is often applied as an indicator of material properties and the quality of production.

The central moments of order 3 and 4 are used to define skewness α (a measure of asymmetry) and kurtosis ε (a measure of peakedness).

$$\alpha = \frac{\mu_3}{\sigma^3} \tag{4.18}$$

$$\varepsilon = \frac{\mu_4}{\sigma^4} - 3 \tag{4.19}$$

If the skewness α is positive then the distribution is skewed to the right, in a case of negative skewness the distribution is skewed to the left. In many engineering applications, the skewness is an important parameter that may significantly affect the results of statistical analysis. However, due to the lack of commonly available data assessment of the skewness may be difficult.

The kurtosis ε is the degree of peakedness relative to normal distribution (that is why the value 3 is subtracted in Eq. (4.19)). If the distribution has a relatively high peak, the kurtosis is positive; if the distribution is flat-topped the kurtosis is negative. In most of the engineering applications only three moment parameters are used (the mean, variance and skewness). As a rule there is not sufficient data to specify the kurtosis.

The central moments μ_2, μ_3, and μ_4 can be expressed using the general moments as follows.

$$\mu_2 = \mu_2' - \mu^2 \tag{4.20}$$

$$\mu_3 = \mu_3' - 3\mu\mu_2' + 2\mu^3 \tag{4.21}$$

$$\mu_4 = \mu_4' - 4\mu\mu_3' + 6\mu^2\mu_2' - 3\mu^4 \tag{4.22}$$

These relationships can be derived from the definitions (4.10, 4.11, 4.12, 4.13, and 4.14). They can be very useful in the practical evaluation of moment parameters as indicated by the following Example 4.2.

Example 4.4. Consider the uniform (rectangular) distribution, described in Example 4.1, having the probability density function $\varphi(x) = 1/(b-a)$. The mean μ and the variance follow from Eqs. (4.12) and (4.15) and Eq. (4.20) as

$$\mu = \int_a^b x \frac{1}{b-a}\,\mathrm{d}x = \frac{a+b}{2}$$

$$\mu_2' = \int_a^b x^2 \frac{1}{b-a}\,\mathrm{d}x = \frac{b^3-a^3}{3(b-a)}$$

$$\sigma^2 = \mu_2' - \mu^2 = \frac{(b-a)^2}{12}, \sigma = \frac{b-a}{2\sqrt{3}}$$

It follows from Eqs. (4.18) and (4.19) that the skewness of a uniform distribution is zero, $\alpha = 0$ (it's a symmetrical distribution), and the kurtosis is negative, $\varepsilon = -1{,}2$ (it's a flat-topped or platykurtic distribution).

Example 4.5. Skewness α_X of any continuous variable X with the probability density function $\varphi_X(x)$ follows from Eqs. (4.14) and (4.18) in an integral form as

$$\alpha_X = \frac{1}{\sigma_X^3} \int_x (x - \mu_X)^3 \varphi_X(x) \mathrm{d}x$$

In practical applications the integration is often done numerically.

4.6 Standardized Random Variable

Standardized random variables are regularly used in tables of random variables, in subsequent numerical calculations and in simulation techniques. The standardized random variable has zero mean and variance equal to 1. Both the original random variable X and the corresponding standardized variable U have the same type of distribution. If the original variable X has the mean μ_X and standard deviation σ_X, then the corresponding standardized random variable U is defined by the transformation formula

$$U = \frac{X - \mu_X}{\sigma_X} \tag{4.23}$$

The inverse transformation of the standardized variable U to the original variable X is

$$X = \mu_X + \sigma_X U \tag{4.24}$$

Equation (4.24) is often used when determining a particular value x_p (for example a fractile) of the original variable X from the corresponding value u_p of the standardized variable U, which is commonly available in tables, in electronic form on the internet or can be obtained from available software products and tools.

Any type of continuous distribution $\Phi_X(x)$ of the original random variable X can be transformed into the standardized distribution $\Phi_U(u)$. As it is a linear transformation, the type of distribution of both variables is the same.

The distribution functions of the original variable X and the transformed variable U (the cumulative probabilities of corresponding values of both variables) must be equal, thus

$$\Phi_U(u) = \Phi_X(x) \tag{4.25}$$

In addition, it is obvious that the differentials of the both the distribution functions must be equal, thus

$$\varphi_U(u)\ \mathrm{d}u = \varphi_X(x)\ \mathrm{d}x \tag{4.26}$$

From the linear transformation (4.24) it follows, for the differentials of random variables X and U, that

$$\mathrm{d}x = \sigma_X \mathrm{d}u \tag{4.27}$$

Substituting Eq. (4.27) into Eq. (4.26) the probability density function $\varphi_U(u)$ of the standardized random variable U can be expressed in terms of density function $\varphi_X(x)$ as

$$\varphi_U(u) = \varphi_X(x)\sigma_X \tag{4.28}$$

The concept of the standardized random variable can be generalized and applied to any type of distribution.

Example 4.6. Consider again the uniform distribution described in Example 4.3: the probability density function $\varphi(x) = 1/(b - a)$, the mean $\mu = (a + b)/2$, the standard deviation $\sigma = (b - a)/(2\sqrt{3})$. The transformation formulas (4.23) and (4.24) become

$$U = \frac{X - \frac{a+b}{2}}{\frac{b-a}{2\sqrt{3}}} = \frac{2X - a - b}{b - a}\sqrt{3}$$

$$X = \frac{b + a}{2} + U\frac{b - a}{2\sqrt{3}}$$

The domain of the standardized variable U is the interval $<-\sqrt{3}, \sqrt{3}>$. The distribution function and probability density functions of the variable U are

$$\Phi(u) = \frac{u}{2\sqrt{3}} + \frac{1}{2}; \quad \varphi(u) = \frac{1}{2\sqrt{3}}$$

A particular value x_p (for example a fractile) of the original variable X can be obtained from the corresponding value u_p of the standardized variable U following transformation formula (4.24)

$$x_p = \frac{b + a}{2} + u_p\frac{b - a}{2\sqrt{3}}$$

This type of relationship between a fractiles x_p of the actual variable X and the corresponding fractile u_p of the standardized variable U (which is usually commonly available) is frequently used in practice.

Example 4.7. A random variable X has the probability density function $\varphi_X(x)$ given as

$$\varphi_X(x) = 1/5, \quad 2 \le x \le 3$$

The mean, standard deviation and skewness of the variable X follow from Eqs. (4.12, 4.13, 4.14, 4.15, and 4.16) as

$$\mu_X = \frac{1}{5}\int_{-2}^{3} x\mathrm{d}x = \frac{1}{2}$$

$$\sigma_X^2 = \frac{1}{5}\int_{-2}^{3} (x-0.5)^2\,\mathrm{d}x = \frac{25}{12}; \quad \sigma_X = \frac{5}{6}\sqrt{3}$$

$$\alpha_X = \frac{1}{\sigma_X^3}\frac{1}{5}\int_{-2}^{3} (x-0.5)^3\,\mathrm{d}x = 0$$

The same results can be obtained from a numerical evaluation of Example 4.4.

References

1. Ang, A.H.-S., Tang, W.H.: Probabilistic Concepts in Engineering. Emphasis on Applications to Civil Environmental Engineering. Wiley, New York (2007)
2. Devore, J., Farnum, N.: Applied Statistics for Engineers and Scientists. Thomson, London (2005)
3. Dunin-Barkovskij, I.V., Smirnov, N.V.: The Theory of Probability and Mathematical Statistics in Engineering. Technical and Theoretical Literature, Moscow (in Russian) (1955)
4. Gurskij, E.I.: The Theory Probability with Elements of Mathematical Statistics. Higher School, Moscow (in Russian) (1971)
5. Holicky, M.: Reliability Analysis for Structural Design. SUNN MeDIA, Stellenbosch (2009)
6. ISO 12491: Statistical Methods for Quality Control of Building Materials and Components. ISO, Geneve (1997)
7. ISO 3534-1: Statistics – Vocabulary and Symbols – Part 1: Probability and General Statistical Terms. ISO, Geneve (1993)
8. ISO 3534-2: Statistics – Vocabulary and Symbols – Part 2: Statistical Quality Control. ISO, Geneve (1993)

Chapter 5
Selected Models of Discrete Variables

Discrete random variables are often applied to engineering and science problems when analysing the number occurrence of a certain event. An elementary but fundamental type of discrete variable that attains only two different values is described by alternative distribution. It can be generalized for a countable number of trial repetitions into binomial and hypergeometric distribution. Time-dependent event are often described by Poisson distribution. The other types of discrete distributions including geometric, negative binomial and multinomial distribution are applied less frequently. In addition to specific distribution parameters, the usual moment parameters, particularly the mean and standard deviation, are used to characterise the distributions. A review of theoretical models provides Appendix 2.

5.1 Alternative Distribution

Discrete random variables are described in detail in books [1–4] including numerical tables. A number of engineering and scientific applications are given particularly in [1, 3].

The basic type of discrete distribution of a random variable X is alternative distribution. The variable X attains only two values 1 and 0 and its probabilistic function is given as

$$\mathrm{P}(x = 1) = p, \quad \mathrm{P}(x = 0) = 1 - p \tag{5.1}$$

Using Eqs. (4.10, 4.12 and 4.15) the mean, variance and standard deviation follows as

$$\mu = 0\,(1 - p) + 1p = p \tag{5.2}$$

$$\sigma^2 = (0 - p)^2 + (1 - p)^2 p = p\,(1 - p) \tag{5.3}$$

M. Holický, *Introduction to Probability and Statistics for Engineers*,
DOI 10.1007/978-3-642-38300-7_5, © Springer-Verlag Berlin Heidelberg 2013

$$\sigma = \sqrt{p\,(1-p)} \tag{5.4}$$

Though it is a simple distribution, its importance in theoretical developments and engineering applications is remarkable. Most of the discrete random variables attaining the values 0, 1, 2, ... can be expressed as the sum of alternative random variables; for example, the number of positive trials in a number of independent experiments. Alternative distribution is also used to develop binomial and Bernoulli distribution.

Example 5.1. A factory device is utilized for 80 % of work time only. The probability that the device is in operation is therefore $\mathrm{P}(x = 1) = p = 0.8$, that the device is out-of-action $\mathrm{P}(x = 0) = 1 - p = 0.2$. Using Eqs. (5.2, 5.3 and 5.4), the mean of the operating time is obviously $\mu = p = 0.8$, its standard deviation $\sigma = [p(1-p)]^{0.5} = 0.4$. It is interesting to note that the coefficient of variation (the relative measure of dispersion) of the operating time is $V = \sigma/\mu = 0.5$.

5.2 Binomial Distribution

Consider n independent random trials carried out under the same (invariant) conditions. In each trial a certain event A may occur with the same probability $\mathrm{P}(A) = p$ (called the probability of success), while the probability of complementary event is $\mathrm{P}(\bar{A}) = 1 - p = q$. (called the probability of failure). The probability function gives probabilities that within n independent trials the number of successful trials (event A occurs) is x. This may occur through a number of different combinations of x successful trials within the total of n trials. The number of such combinations k is given by the combination number

$$k = \binom{n}{x} = \frac{n}{x(n-x)} \tag{5.5}$$

Each combination may occur with the probability $p^x q^{n-x}$ (x successes and $n-x$ failures). Thus, the resulting probability function may be expressed as

$$\mathrm{P}(x, n, p) = \binom{n}{x} p^x q^{n-x} = \binom{n}{x} p^x (1-p)^{n-x} \tag{5.6}$$

Example 5.2. Binomial distribution is often linked to the so-called Bernoulli experiment. In a box there are N balls, X white balls and $N-X$ back ones. The probability that a white ball will be pulled out of the box is

$$p = \frac{X}{N}$$

The probability that a black ball will be chosen is

$$q = \frac{N - X}{N}$$

Obviously $p + q = 1$ (complementary probabilities). In a series of n trials, in which a ball taken from the box will be always returned back, these probabilities do not change (invariant conditions). The number of white balls x taken out of the box is 0, 1, 2, …, n and their probabilities P(0;n;p), P(1;n;p), P(2;n;p) may be determined using Eq. (5.6). For example the probability that in a series of n trials a white boll will never be, or will always be, pulled out is

$$\mathrm{P}(0, n, p) = q^n$$

$$\mathrm{P}(n, n, p) = p^n$$

The mean, variance, standard deviation, skewness and kurtosis of binomial distribution may be derived using the so-called moment developing function [1]

$$\mu = np \tag{5.7}$$

$$\sigma^2 = npq = np(1 - p) \tag{5.8}$$

$$\sigma = \sqrt{npq} \tag{5.9}$$

$$\alpha = \frac{q - p}{\sqrt{npq}} \tag{5.10}$$

$$\varepsilon = \frac{1 - 6pq}{npq} \tag{5.11}$$

The coefficient of variation follows from Eqs. (5.7) and (5.9) as

$$V = \sqrt{\frac{q}{np}} \tag{5.12}$$

Figure 5.1 shows two probability functions P(x; 5; 0.8) and P(x; 10; 0.8) assuming the probability $p = 0.8$ (the mean $\mu = np = 4$ and 8).

Example 5.3. In a factory five independent machines are utilized, each of which is in operation for 80 % of the work time. The mean, the standard deviation and the coefficient of variation are given by Eqs. (5.7, 5.9 and 5.12).

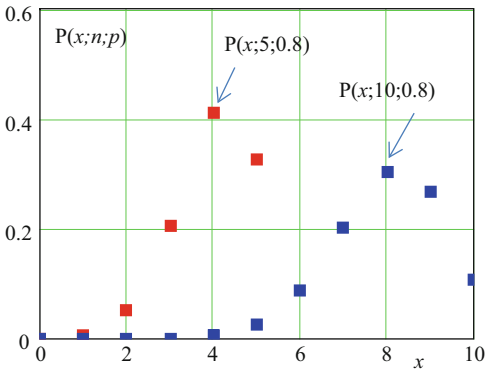


Fig. 5.1 Probability function P(x,n,p) of the binomial distribution

$$\mu = 5 \cdot 0.8 = 4; \quad \sigma = \sqrt{npq} = 0.89; \quad V = 0.22$$

The probability that only two machines are in operation at any one time may be determined using binomial distribution, as follows.

$$P(2;\ 5;\ 0.8) = \binom{5}{2} 0.8^2 0.2^{5-2} \cong 0.05$$

Thus, there is a small probability 0.05 that only two machines will be in operation during work time.

5.3 Hypergeometric Distribution

Consider a population of N elementary events (for example of certain products), X of which belong to the event A (non-conforming products), $N-X$ belong to the complementary event $\bar{A}$ (conforming products). The probability function P(x;n;X;N) describes the probability that x events from n randomly chosen elementary events (note that $\max\{0,n + X-N\} < x < \min\{X,n\}$) belong to the (positive) event A. Using the classical definition of probability, explained in Sect. 2.3, the probability function can be derived as

$$P(x; n; X; N) = \frac{\binom{X}{x}\binom{N-X}{n-x}}{\binom{N}{n}} \tag{5.13}$$

Hypergeometric distribution differs from the Bernoulli experiment, described in the previous section, by the fact that elementary events affect the subsequent

probability of attaining event A and $\bar{A}$ (the products are not returned to the population) as illustrated by the following example.

Example 5.4. In a population of $N = 100$ products there are $X = 30$ non-conforming units. The probability function is given as

$$P(x; n; 70; 100) = \frac{\binom{30}{x}\binom{100-30}{n-x}}{\binom{100}{n}}$$

The probability that in-between 3 randomly chosen units, $n = 3$, is one, $x = 1$, non-conforming is then

$$P(1; 3; 30; 100) = \frac{\binom{30}{1}\binom{70}{3-1}}{\binom{100}{3}} = \frac{30 \cdot 70 \cdot 69 \cdot 3}{100 \cdot 99 \cdot 98} = 0.448$$

So, there is a relatively high probability 0.448 that 1 from 3 chosen products is non-conforming.

Similarly as in case of Bernoulli distribution, the hypergeometric distribution is often characterized by two complementary probabilities

$$p = \frac{X}{N}; \quad q = \frac{N-X}{N} = 1 - p \tag{5.14}$$

The mean, variance, standard deviation, coefficient of variation and skewness follow can be obtained from general expressions (4.10, 4.12, 4.16 and 4.18).

$$\mu = np \tag{5.15}$$

$$\sigma^2 = npq\frac{1-\frac{n}{N}}{1-\frac{1}{N}} = np(1-p)\frac{N-n}{N-1} \tag{5.16}$$

$$\sigma = \sqrt{npq\frac{1-\frac{n}{N}}{1-\frac{1}{N}}} = \sqrt{np(1-p)\frac{N-n}{N-1}}, \quad V = \sqrt{\frac{(1-p)}{np}\frac{N-n}{N-1}} \tag{5.17}$$

$$\alpha = \frac{(N-2X)\sqrt{(N-1)}(N-2n)}{\sqrt{nX(N-X)(N-n)}(N-2)} = \frac{(1-2p)\sqrt{(N-1)}(N-2n)}{\sqrt{np(1-p)(N-n)}(N-2)} \tag{5.18}$$

Obviously for an increasing population N the basic moment characteristics of binomial and hypergeometric distributions are approaching (see also Appendix 2).

5.4 Poisson Distribution

The Poisson distribution is frequently used to describe the time dependent occurrence of random events. Let n independent events belonging to a given result A occurs within an interval T. Thus, an average c of these events occurs within a time unit

$$c = \frac{n}{T} \tag{5.19}$$

A natural question is to assess the probability that x of these events occurs in a given interval t, ($t < T$). It is assumed that n events occurring in the interval T are mutually independent and, consequently, the probability that each of these events occurs in the time interval t is

$$p = \frac{t}{T} \tag{5.20}$$

Using binomial distribution (see Eq. (5.6)) the probability function can be now approximated as

$$\mathrm{P}(x) = \binom{n}{x} p^x (1-p)^{n-x} \tag{5.21}$$

Applying now a limit procedure for P(x) with $T \to \infty$, the probability function (5.21) can be expressed in the usual form

$$P(x, \lambda) = \frac{\lambda^x}{x!} e^{-\lambda} \tag{5.22}$$

Here the parameter λ denotes the average number of events within the time period t given by the mean of Poisson distribution

$$\mu = \lambda = ct = \frac{nt}{T} \tag{5.23}$$

Using the moment developing function, the variance, standard deviation and skewness of the variable X can be derived as

$$\sigma^2 = \lambda, \quad \sigma = \sqrt{\lambda}, \quad \alpha = 1/\sqrt{\lambda} \tag{5.24}$$

Figure 5.2 indicates the probability function P(x,λ) for two average numbers of events λ within the time period t, $\lambda = 0.2 = 0.2$ and 2.

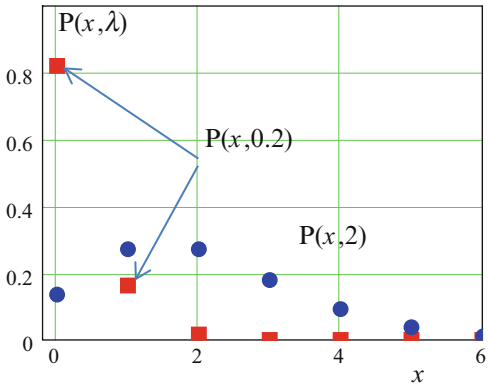


Fig. 5.2 Probability function P(x,λ) of the Poisson distribution

Example 5.5. An automatic machine delives two components in 1 min. During a period of 8 h 38 components do not comply with the requirements. The basic parameters and the probability function given by Eqs. (5.22) and (5.23) are then

$$c = \frac{38}{(8 \times 60)} = 0.08; \quad t = \frac{5}{2} \text{ min}; \quad \lambda = c \times t = 0.2; \quad P(x) = \frac{0.2^x}{x!} e^{-0.2}$$

The probability that in a series of five components two or more components will be non-conforming can be calculate as

$$P(x \geq 2) = 1 - P(0) - P(1) = 0.0175$$

5.5 Geometric Distribution

There are other distributions of discrete random variable that are not so frequently used in engineering applications. These types of distributions include geometric distribution, negative binomial distribution, multinomial distribution and multi hypergeometric distribution. In general, these distributions are particular cases of the previous type of discrete distributions. Description of other types of distribution may be found in specialized literature [1]. Let us consider briefly a geometric distribution that has some important engineering applications.

The geometric distribution in question is related to the binomial distribution described in Sect. 5.2. It describes the probability P(n) that within n trials (n = 1, 2, 3,) the Bernoulli experiment will be successful just once. For example the first $n-1$ trials are unsuccessful (no event A is observed) and subsequent trial number n is successful (event A occurs). The probability of the successful trial is p, probability of the unsuccessful trial is $1 - p = q$. Thus the probability that $n-1$ trials are unsuccessful and then trial number n is successful can be expressed as

$$\mathrm{P}(n,\ p) = p(1-p)^{n-1} = p\ q^{n-1} \tag{5.25}$$

Equation (5.25) describes geometric sequence and that is why the distribution is called geometric distribution. It should be noted that there is an alternative formulation of the geometric distribution when the domain of $n = 0,\ 1,\ 2,\ldots$Then in Eq. (5.25) the exponent of q is n. Here the formulation (5.25) is accepted.

The mean and standard deviation of the variable n is given as

$$\mu_n = 1/p; \quad \sigma_n = \sqrt{q}/p \tag{5.26}$$

For small probability p the second expressions (5.26) may be approximated as

$$\sigma_n \approx 1/p \tag{5.27}$$

Thus, for small p the standard deviation σ_n is approximately equal to the mean μ_n and coefficient variation $V_n = \sigma_n/\mu_n$ approaches one; in fact V_n follows from expressions (5.26) as

$$V_n = q = 1 - p \approx 1 \tag{5.28}$$

The distribution is highly asymmetric having the skewness

$$\alpha_n = \frac{1+q}{\sqrt{1-p}} = \frac{2-p}{\sqrt{1-p}} \approx 2 \tag{5.29}$$

The geometric distribution may have an important engineering or scientific applications. Assume that the time (or space) interval T is discretized into n basic intervals in such a way that the probability p of occurrence of a specified event A in any interval is approximately the same (assumption of the geometric distribution). Then the probability $\mathrm{P}(n, p)$ given by Eq. (5.25) offers a relationship between the number of basic intervals n (reoccurrence time) and the probability p of the event A in one basic interval. The mean reoccurrence time μ_n and its standard deviation σ_n can be assessed by Eq. (5.27) as

$$\mu_n \approx \sigma_n \approx 1/p \tag{5.30}$$

The coefficient of variation $V_n \approx 1$ indicates a great uncertainty in assessment of the reoccurrence time (the number of basic intervals) n.

The probability function $\mathrm{P}(n, p)$ is a monotonously decreasing function shown in Fig. 5.3 for two probabilities $p = 0.02$ and 0.05, having the mean and standard deviation equal to $\mu_n \approx \sigma_n \approx 50$ and 20.

Example 5.6. Assume that the time is discretized into a number of intervals having the same duration of 1 year. The occurrence of annual extremes of a certain climatic actions (for example due temperature, snow or wind) can be assumed to be mutually

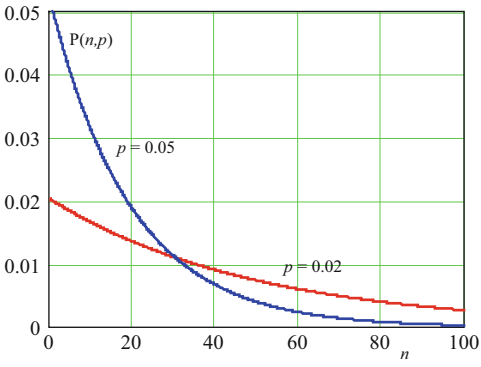


Fig. 5.3 Probability function of the geometric distribution P(*n*,*p*)

independent. Let us define A as an event whereby a particular intensity of the action (the critical or characteristic value) is exceeded by the annual extreme with the probability p. Then the reoccurrence of the critical value may be expected within the time T (number n of time intervals) called the return period. The mean of the return period μ_T may be assessed by Eq. (5.30) as

$$\mu_T = \mu_n \approx 1/p$$

However, it should be emphasized that the return period T is a random variable (assumed here to be described by geometric distribution) that has standard deviation σ_T approximately equal to the mean μ_T (see Eq. (5.30))

$$\sigma_T = \sigma_n \approx 1/p$$

If the probability p of event A is specified by the value $p = 0.02$ then the mean reoccurrence time is assessed as $\mu_T \approx 1/p = 50$ years (Fig. 5.3). The standard deviation σ_T is also approximately equal to 50 years. The coefficient of variation $V_T \approx 1$ confirms a considerable uncertainty in the assessment of the return period T.

It should be mentioned that the probability function P(n, p) is a monotonously decreasing function of n with a high asymmetry (see Fig. 5.3). It indicates that the reoccurrence time T less than the mean $\mu_T = \mu_n \approx 1/p$ is more likely than the right of the mean (see Example 3.6). That is however valid provided that the assumed geometric distribution is applicable (the probability p that event A occurs in any basic interval is the same).

Example 5.7. Codes of practice commonly specify the characteristic value of wind speed as the speed that on average occurs once in 50 years (the so-called 50 years wind speed). It means that the characteristic value can be expected in any 1 year period with the probability 0.02. The probability that a building will be subjected to the characteristic wind speed during the year number $n = 10$ follows from Eq. (5.25) as

$$\mathrm{P}(10,0,02) = 0.02 \times 0.98^{10-1} = 0.017$$

The probability that the structure will be exposed to the characteristic load during the 50 years can then be expressed as the sum of geometric series

$$P(X \leq 50) = \sum_{i=1}^{10} 0.02 \times 0.98^{i-1} = 0.02 \frac{1 - 0.98^{50}}{1 - 0.98} = 0.636$$

Thus, during the first 50 years the probability that the structure will be exposed to the characteristic wind pressure is greater ($\mathrm{P}(X \leq 50) = 0.636$) than the complementary probability that it will occur after 50 years of structural existence ($\mathrm{P}(X > 50) = 0.364$).

References

1. Ang, A.H.-S., Tang, W.H.: Probabilistic Concepts in Engineering. Emphasis on Applications to Civil Environmental Engineering. Wiley, New York (2007)
2. Devore, J., Farnum, N.: Applied Statistics for Engineers and Scientists. Thomson, London (2005)
3. Dunin-Barkovskij, I.V., Smirnov N.V.: The Theory of Probability and Mathematical Statistics in Engineering. Technical and Theoretical Literature, Moscow (in Russian) (1955).
4. Gurskij, E.I.: The Theory Probability with Elements of Mathematical Statistics. Higher School, Moscow (in Russian) (1971)

Chapter 6
Selected Models of Continuous Variables

Most of the random variables used in engineering and scientific applications are described by continuous variables that may attain any value from a given interval. The probability density function of a continuous random variable is often interpreted as the limiting case of a histogram when the number of observations is increasing to infinity. An elementary type of continuous distribution is the uniform distribution describing a variable that may attain any value from a given interval with an equal chance. Frequently used distributions, having the probability density function of a typical bell shape, and applied in engineering and science, include the normal, lognormal, Beta as well as different types of extreme value distributions like the Gumbel, Weibull and Frechet distributions. Other types of continuous distributions are applied less frequently. A review of selected models of continuous random variables is provided in Appendix 3.

6.1 Normal Distribution

From a theoretical and practical point of view the most important type of distribution of a continuous random variable is the well-known normal (Laplace-Gauss) distribution [1–4]. A symmetric normal distribution of a variable X is defined on an unlimited interval $-\infty < x < \infty$ (which can be undesirable in some practical applications) and depends on two parameters only – on the mean μ and standard deviation σ. Symbolically it is often denoted as $N(\mu, \sigma)$. This distribution is frequently used as a theoretical model of various types of random variables describing some loads (self-weight), mechanical properties (strengths), and geometrical properties (outer dimensions). It is convenient for a symmetric random variable with a relatively low variance (a coefficient of variation $V < 0.2$). It may fail for asymmetric variables with a greater variance and a significant skewness $\alpha > 0.3$.

The probability density function of a normal random variable X with the mean μ_X and standard deviation σ_X is given by the exponential expression

M. Holický, *Introduction to Probability and Statistics for Engineers*,
DOI 10.1007/978-3-642-38300-7_6, © Springer-Verlag Berlin Heidelberg 2013

$$\varphi(x) = \frac{1}{\sigma_X \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right] \tag{6.1}$$

The proof that μ_X and σ_X are the mean and standard deviation of the random variable X described by the probability density function (6.1) follows from general Eqs. (4.13, 4.14, 4.15 and 4.16). Furthermore, using Eqs. (4.18) and (4.19) it may be shown that the skewness α_X and kurtosis ε_X of the normal random variable X are zero, $\alpha_X = \varepsilon_X = 0$.

No analytical expression is available for the distribution function $\Phi(x)$. Nevertheless, numerical tables for the probability density function as well as for the distribution function are commonly available in literature [1, 2] and on the internet. A brief numerical table for the distribution function $\Phi(x)$ is also available in Appendix 7. All these tables give the probability density function $\varphi(u)$ and the distribution function $\Phi(u)$ of the standardized variable U that is derived from the actual variable X using the formula (4.23) (applicable for any distribution)

$$U = \frac{X - \mu_X}{\sigma_X} \tag{6.2}$$

Here μ_X and σ_X denote the mean and standard deviation of the actual variable X. The standardized random variable U has a zero mean and a standard deviation equal to one; the normal standardized distribution is symbolically denoted N(0, 1).

The probability density function of the standardized random variable U having the normal distribution N(0, 1) follows from Eqs. (6.1) and (6.2) as

$$\varphi_U(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \tag{6.3}$$

Tabulated values of the probability density function $\varphi_U(u)$ and corresponding distribution function $\Phi_U(u)$, given in Appendix 7, are transformed to the original random variable X using the transformation formula (6.2) in a modified form (4.24).

The probability density function of the normal distribution is a symmetrical function (skewness $\alpha = 0{,}0$) as indicated in Fig. 6.1, where it is shown together with a log-normal distribution (described in the next Section) as having a coefficient of skewness $\alpha = 1{,}0$. Both probability density functions are shown for the standardized random variable U defined by Eq. (6.2) and having zero mean and unit standard deviation.

Note that the probability density function of the standardized normal distribution is plotted for u within the interval $<-3,+3>$. This interval covers a high occurrence probability (0.9973) of the variable U (in technical practice such an interval of the actual variable is sometimes denoted as $\pm 3\sigma$ interval).

Example 6.1. Let us denote as u_p the value of the standardised normal variable for which the distribution function is equal to a specified probability p, thus

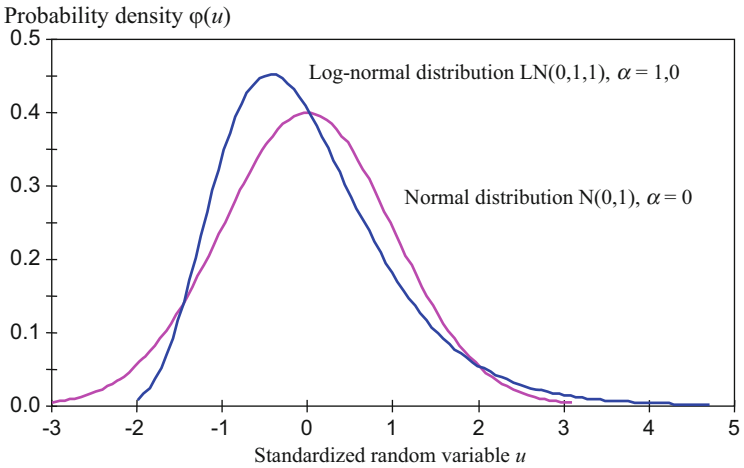


Fig. 6.1 Normal and log-normal distribution (skewness $\alpha = 1.0$)

$$\Phi(u_p) = p$$

The corresponding value x_p of the original variable X, having the mean μ_X and standard deviation σ_X, follows from transformation formula (6.2) as

$$x_p = \mu_X + u_p \, \sigma_X$$

If the probability $p = 0.05$ then it follows from tables of the standardised distribution N(0, 1) that $u_p = -1.645$ (see also Appendix 7) and the corresponding value of the original random variable X is

$$x_p = \mu_X - 1.645 \, \sigma_X$$

If the probability $p = 0.001$ then $u_p = -3.09$ and $x_p = \mu_X - 3.09 \, \sigma_X$.

6.2 Lognormal Distribution

A general three parameter log-normal distribution is defined on a one-sided limited interval $x_0 < x < \infty$ or $-\infty < x < x_0$ [3–6]. It is an asymmetric distribution that partly eliminates one of the undesirable properties of the normal distribution, i.e. the infinite definition domain. A general log-normal distribution is dependent on three parameters, and for that reason is often called the three-parameter log-normal distribution. Commonly, the moment parameters can be applied to define the distribution: the mean μ_X, the standard deviation σ_X and the skewness

α_X. Instead of the skewness α_X (when it is unknown or uncertain), the lower or upper bounds x_0 may be used.

A random variable X has a log-normal (general three-parameter) distribution if the transformed random variable

$$Y = \ln |X - x_0| \tag{6.4}$$

has a normal distribution. In this relation x_0 denotes the lower or upper bound of the variable X, which depends on the skewness α_X. If the variable has a mean μ_X and standard deviation σ_X, then the lower or upper bound can be expressed as

$$x_0 = \mu_X - \sigma_X / c \tag{6.5}$$

Here the coefficient c is given by the value of skewness α_X according to the relation

$$\alpha_X = c^3 + 3c \tag{6.6}$$

from which follows an explicit relation for c [6]

$$c = \left[\left(\sqrt{\alpha_X^2 + 4} + \alpha_X \right)^{1/3} - \left(\sqrt{\alpha_X^2 + 4} - \alpha_X \right)^{1/3} \right] 2^{-1/3} \tag{6.7}$$

The dependence of the limit x_0 on the coefficient α is apparent from Table 6.1, where the lower bound $u_0 = -1/c$ of the standardized variable $U = (X - \mu_X)/\sigma_X$ is given for selected values of the coefficient of skewness $\alpha_X \geq 0$. For $\alpha_X \leq 0$ values of u_0 with the inverse sign (i.e. positive) are considered. A log-normal distribution with the skewness $\alpha_X = 0$ becomes a normal distribution ($u_0 = -1/c \rightarrow \pm \infty$).

When specifying a theoretical model, it is therefore possible to consider either the skewness α_X or, alternatively, the lower or upper bound of the distribution x_0 (in addition to the mean μ_X and standard deviation σ_X). Generally, the first alternative is preferable because more credible information is usually available for the coefficient of skewness than for the lower or upper bound. In general, the moment parameter called the coefficient of skewness provides a better characterisation of the overall distribution of the population (particularly of large populations) than the lower or upper bounds.

The probability density function and distribution function of the general three-parameter log-normal distribution may be obtained from the well-known normal distribution by using a modified (transformed) standardized variable u' obtained from the original standardized random variable $u = (x - \mu_X)/\sigma_X$ as [6]

$$u' = \frac{\ln\left(\left|u + \frac{1}{c}\right|\right) + \ln\left(\, |c| \sqrt{1 + c^2}\right)}{\sqrt{\ln(1 + c^2)}} \operatorname{sign}(\alpha_X) \tag{6.8}$$

Table 6.1 The coefficient u_0 for selected coefficient of skewness $\alpha_X \geq 0$

α_X	0	0.5	1.0	1.5	2.0
$u_0 = -1/c$	$-\infty$	−6.05	−3.10	−2.14	−1.68

Here sign(α_X) equals +1 for $\alpha_X > 0$ and −1 for $\alpha_X < 0$. The probability density function $\varphi_{\mathrm{LN},U}(u)$ and the distribution function $\Phi_{\mathrm{LN},U}(u) = \Phi_{\mathrm{LN},X}(x)$ of the log-normal distribution are given as

$$\varphi_{\mathrm{LN},U}(u) = \frac{\phi(u')}{\left(\left|u + \frac{1}{c}\right|\right)\sqrt{\ln(1+c^2)}} \tag{6.9}$$

$$\Phi_{\mathrm{LN},X}(x) = \Phi_{\mathrm{LN},U}(u) = \Phi(u') \tag{6.10}$$

Here $\varphi(u')$ and $\Phi(u')$ denote the probability density and distribution function of the standardized normal variable.

A special case of the three-parameter log-normal distribution is the popular log-normal distribution with the lower bound at zero ($x_0 = 0$) called here two-parameter log-normal distribution. This distribution depends on two parameters only – the mean μ_X and the standard deviation σ_X (a symbolic notation LN(μ, σ) is then used). In such a case it follows from Eq. (6.5) that the coefficient c is equal to the coefficient of variation V_X. It further follows from Eq. (6.6) that the skewness α_X of the log-normal distribution with the lower bound at zero is given by the coefficient of variation V_X as

$$\alpha_X = 3V_X + V_X^3 \tag{6.11}$$

Thus, the log-normal distribution with the lower bound at zero ($x_0 = 0$) has always a positive skewness. Consequently, applications of the log-normal distribution with the lower bound at zero ($x_0 = 0$) can thus lead to unrealistic theoretical models (usually underestimating the occurrence of negative and overestimating the occurrence of positive deviations from the mean), particularly for higher values of the coefficient of variation V_X. Then the three-parameter log-normal distribution may be used. Although the occurrence of negative values can also be undesirable (unrealistic for most mechanical quantities), it is usually negligible from a practical point of view.

Example 6.2. The skewness may have a relatively high value (greater than 0.5); e.g. for the coefficient of variation equal to 0.30 a coefficient of skewness $\alpha_x = 0.927$ is obtained from Eq. (6.11).

The log-normal distribution is widely applied in the theory of reliability as a theoretical model for various types of random variables [6]. In general it can be used for one-sided limited asymmetric random variables including material properties, actions, and geometrical data. In particular, the log-normal distribution with the lower bound at zero ($x_0 = 0$) is commonly used for resistance properties (strengths) of various materials (concrete, steel, timber, masonry).

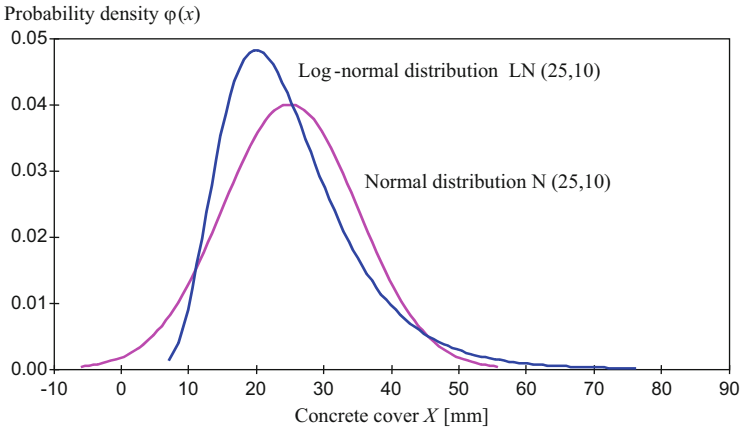


Fig. 6.2 Probability density functions for the concrete cover depth

Example 6.3. A concrete cover depth X of a reinforced concrete cross-section has the mean $\mu = 25$ mm and the standard deviation $\sigma = 10$ mm. The probability density function $\varphi(x)$ for a normal distribution and for a two-parameter log-normal distribution (with the lower bound at zero) is shown in Fig. 6.2.

It follows from Fig. 6.2 that the normal distribution predicts some occurrence of negative values of the concrete cover depth, which may not correspond to reality. On the other hand, the log-normal distribution with the lower bound at zero may overestimate the occurrence of positive deviations, which may not be acceptable and may affect the resulting random variable bending resistance of the cross-section.

The overestimation of the occurrence of extreme positive deviations is due to a high skewness $\alpha = 1.36$ (given by Eq. (6.11)) of the two-parameter log-normal distribution. Note that the available experimental data on a concrete cover depth indicate that in most cases the skewness of the distribution is less than 1, and if no other evidence is available, then the value $\alpha \approx 0.5$ is recommended to be assumed.

6.3 Gamma Distribution

Another popular type of one-sided limited distribution is Pearson's distribution type III. Its detailed description is available in [1]. A special case of Pearson's distribution type III with the lower bound at zero is gamma distribution. The probability density function of this important distribution is dependent on two parameters only: on the mean μ and standard deviation σ. To simplify the notation two auxiliary parameters λ and k are often used

$$\varphi(x) = \frac{\lambda^k \, x^{k-1} \, \exp(-\lambda x)}{\Gamma(k)}, \lambda = \frac{\mu}{\sigma^2}, k = \left(\frac{\mu}{\sigma}\right)^2 \tag{6.12}$$

Here $\Gamma(k)$ denotes the gamma function of the parameter k. The moment parameters of the gamma distribution follow from Eq. (6.12) as

$$\mu = \frac{k}{\lambda}, \quad \sigma = \frac{\sqrt{k}}{\lambda}, \quad \alpha = \frac{2}{\sqrt{k}} = \frac{2\sigma}{\mu} = 2V, \quad \varepsilon = \frac{3\alpha^2}{2} \tag{6.13}$$

The curve is bell shaped for $k > 1$, i.e. for a skewness $\alpha < 2$ (in the inverse case the gamma distribution is a decreasing function of x). For $k \to \infty$, the gamma distribution approaches the normal distribution with parameters μ and σ.

Gamma distribution is applied in much the same way as the log-normal distribution with the lower bound at zero. However, it differs from the log-normal distribution by its skewness, which is equal to the double of the coefficient of variation ($\underline{\alpha} = 2w$) and is considerably lower than the skewness of the log-normal distribution with the lower bound at zero. In accordance with Eq. (6.11) it has the skewness $\alpha_X = 3V_X + V_X^3$. That is why gamma distribution is more convenient for describing some geometrical quantities and variable actions.

Example 6.4. A sample of experimental measurements of a concrete cover depth has the following characteristics: a sample size $n = 157$, $m = 26.8$ mm, $s = 11.1$ mm, and $a = 0.40$. It is a relatively large sample, which can be used for assessing skewness (long-term experience may be available to verify the obtained value). A histogram of the experimental measurements and theoretical models of the normal distribution, log-normal distribution with the origin at zero, gamma distribution and beta distribution (described in the following Section) is shown in Fig. 6.3. It appears that the gamma and beta distributions are the most suitable theoretical models. However, it follows from Eq. (6.13) that the skewness of the gamma distribution is $2 \times 11.1/26.3 = 0.83$, thus about double the value assessed from the measurements. Obviously, the beta distribution would be the most suitable model.

To choose an appropriate theoretical model for experimental data is, in general, a complicated task. Information about theoretical methods (the so-called goodness of fit tests) provided by mathematical statistics can be found in literature [1–3]. In this textbook only some practical aspects and procedures will be indicated.

6.4 Beta Distribution

Beta distribution (also called Pearson's distribution type I) is defined on a two-sided interval $<a, b>$ (this interval can be arbitrarily extended and then the distribution approaches the normal distribution). Generally, the beta distribution depends on four parameters and is used mainly in those cases where the domain of the random variable is evidently limited (some actions and geometrical data, e.g. the weight of

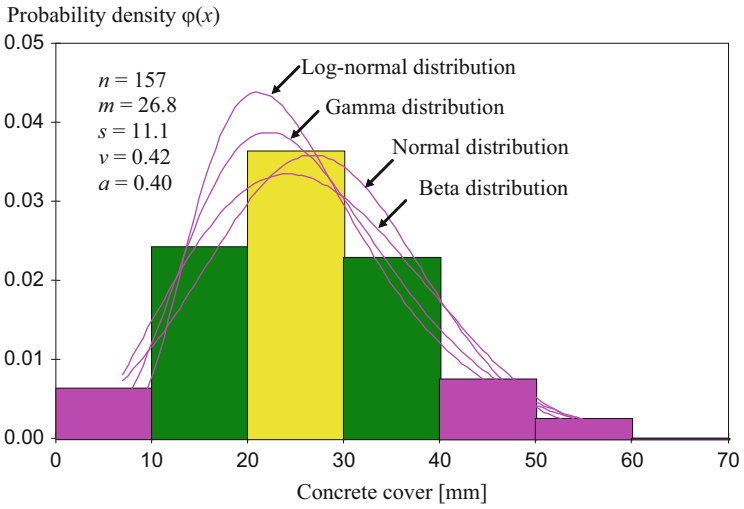


Fig. 6.3 Histogram and theoretical models for concrete cover depth of reinforcement

a subway car, fire load intensity, a concrete reinforcement cover depth). The principal difficulty in a practical application of the beta distribution is the necessity of estimating four parameters, for which credible data may not always be available [6].

The beta distribution is usually written in the form

$$\varphi(x) = \frac{(x-a)^{c-1}(x-b)^{d-1}}{\mathrm{B}(c,d)(b-a)^{c+d-1}} \tag{6.14}$$

where c and d are the so-called shape parameters and B(c,d) is the beta function (also called the Euler integral). The lower and upper bounds are given as

$$a = \mu - c\ g\,\sigma, \quad b = \mu + d\ g\,\sigma, \quad g = \sqrt{\frac{c+d+1}{cd}} \tag{6.15}$$

where g is an auxiliary parameter. The parameters c and d can be derived from Eq. (6.15) as

$$c = \frac{\mu - a}{b-a}\left(\frac{(\mu-a)(b-\mu)}{\sigma^2} - 1\right), \quad d = \frac{b-\mu}{b-a}\left(\frac{(\mu-a)(b-\mu)}{\sigma^2} - 1\right) \tag{6.16}$$

The moment parameters of the beta distribution can be expressed in terms of the parameters a, b, c and d as

$$\mu = \frac{a + (b-a)c}{(c+d)}, \quad \sigma = \frac{(b-a)}{(cg+dg)} \tag{6.17}$$

$$\alpha = \frac{2g\,(d-c)}{(c+d+2)}, \quad \varepsilon = \frac{3g^2(2(c+d)^2 + cd(c+d-6))}{(c+d+2)(c+d+3)} - 3 \tag{6.18}$$

Note that the skewness α and kurtosis ε are dependent on only the parameters c and d (they are independent of the limits a and b). That is why the parameters c and d are called the shape parameters. In practical applications the distribution is used for $c > 1$ and $d > 1$ (otherwise the curve is J or U shaped); for $c = d = 1$ it becomes a uniform distribution; for $c = d = 2$ it is the so-called parabolic distribution on the interval $<a, b>$. When $c = d$, the curve is symmetric around the mean. When $d \to \infty$, the curve becomes type III Pearson's distribution (see Sect. 3.5). If $c = d \to \infty$, it approaches the normal distribution. Depending on the shape parameters c and d the beta distribution thus covers various special types of distribution. The location of the distribution is given by the parameters a and b.

Beta distribution can be defined in various ways. If all four parameters a, b, c and d are given, it is possible to assess the moment parameters μ, σ, α and ε from Eqs. (6.15, 6.16, 6.17, and 6.18). In practical applications, however, two other combinations of input parameters are likely to be applied [6]:

1. The input parameters are μ, σ, a and b. The remaining parameters c and d will be assessed from Eqs. (6.15) and (6.16), the moment parameters α and ε from Eqs. (6.17) and (6.18).
2. The input parameters are μ, σ, α and one of the limits a (for $\alpha > 0$) or b (for $\alpha < 0$). The remaining parameters of distributions b (or a), c and d will be assessed by means of Eqs. (6.15, 6.16 and 6.17).

The beta distribution with the lower bound $a = 0$ is often used in practical applications. It can be shown that in such a case the beta distribution is defined as

$$\alpha \leq 2V \tag{6.19}$$

where $V = \sigma / \mu$ is the coefficient of variation. For $\alpha = 2\ V$ the curve becomes type III Pearson's distribution (see Sect. 3.5). Therefore, if the input parameters are the mean μ, the standard deviation σ and the skewness $\alpha \leq 2\ V$, the beta distribution with a lower limit at zero ($a = 0$) is fully described. The upper limit of the beta distribution with the lower bound at zero follows from the relation (6.15)

$$b = \frac{\mu(c+d)}{c} = \frac{\mu(1 + w\,(2 + \alpha w))}{(2w - \alpha)} \tag{6.20}$$

In Eq. (3.32) the parameters c and d are given as

$$c = -\frac{\alpha}{2w}\,\frac{(2w-\alpha)^2-(4+\alpha^2)}{(w\alpha+2)^2-(4+\alpha^2)} \tag{6.21}$$

$$d = \frac{\alpha}{2}\,\frac{(2w-\alpha)^2-(4+\alpha^2)}{(w\alpha+2)^2-(4+\alpha^2)}\,\frac{2+\alpha w}{\alpha-2w} \tag{6.22}$$

Equations (6.21) and (6.22) follow from the general Eqs. (6.13, 6.14, 6.15, 6.16 and 6.17) for the lower bound $a = 0$.

Example 6.5. Given a mean $\mu = 25$ mm, a standard deviation 10 mm ($V = 0.40$), and a skewness $\alpha = 0.5$, let us assess the parameters of a beta distribution with an origin at zero ($a = 0$) for a reinforcement cover layer. The inequality in Eq. (6.19) is thereby satisfied ($0.5 < 2 \times 0.4$). From Eqs. (6.21) and (6.22) it follows that

$$c = -\frac{0.5}{2\times 0.4}\,\frac{(2\times 0.4-0.5)^2-(4+0.5^2)}{(0.4\times 0.5+2)^2-(4+0.5^2)} = 4.406$$

$$d = \frac{0.5}{2}\,\frac{(2\times 0.4-0.5)^2-(4+0.5^2)}{(0.4\times 0.5+2)^2-(4+0.5^2)}\,\frac{2+0.5\times 0.4}{0.5-2\times 0.4} = 12.926$$

For the upper bound of the distribution it follows from Eq. (6.20) that

$$b = \frac{25(4.407+12.926)}{4.407} = 98.326$$

Figure 6.4 shows the beta distribution with the parameters assessed above together with the corresponding normal, log-normal and Gamma distributions that have the same mean μ and standard deviation σ. Obviously, there are considerable differences between the distributions indicated in Fig. 6.4.

The normal distribution (skewness $\alpha = 0$) predicts the occurrence of negative values, which may not comply with the real conditions for the reinforcement cover depth. The log-normal distribution with the lower bound at zero has a skewness $\alpha = 1.264$ (given by Eq. (6.11)), which does not correspond to experimental results and leads to an overestimation of the occurrence of positive deviations (which may further lead to unfavourable consequences for the resistance of the reinforced concrete element). The gamma distribution has a skewness $\alpha = 2\,V = 0.8$ (Eq. (6.13)) and is closer to the experimental value 0.5. The most convenient model seems to be the beta distribution with a skewness $\alpha = 0.5$ obtained from experimental data.

The above discussion can be supplemented by statistical tests (see Chap. 10 and books [1–3]). On the other hand, it should be mentioned that goodness of fit tests

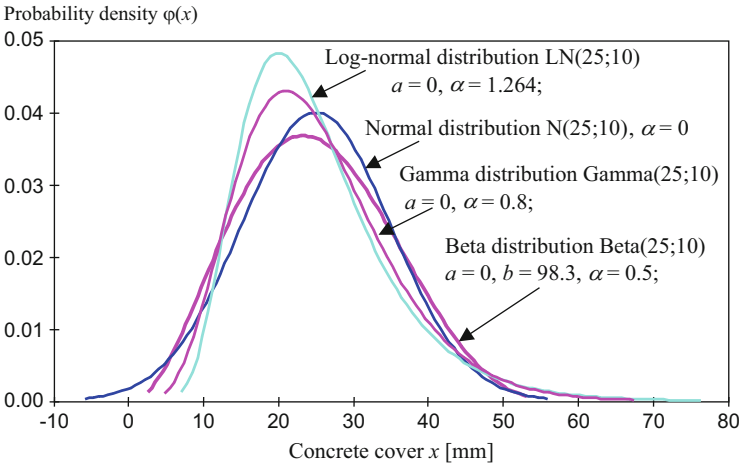


Fig. 6.4 Normal, log-normal, gamma and beta distributions for the concrete cover depth of reinforcement in a reinforced concrete element

often fail and do not lead to an unambiguous conclusion. In such a case the selection of a convenient model depends on the character of the basic variable, on available experience and on common experience.

6.5 Gumbel Distribution

The extreme values (maximal or minimal) in a population of a certain size are random variables and their distribution is extremely important for the theory of structural reliability. Three types of distribution of extreme values are usually covered in specialised literature, and are denoted as types I, II and III. Each of the types has two versions – one for the distribution of minimal values, the second for the distribution of maximal values. All these types of distribution have a simple exponential shape and are convenient to work with. The extreme value distribution of type I, which is commonly called the Gumbel distribution, is described in detail. Descriptions of the other types of distribution can be found in specialised literature [1, 2].

The distribution function of type I for the maximal values distribution version (the Gumbel distribution of maximum values) has the form [6]

$$\Phi(x) = \exp(-\exp(-c(x - x_{\text{mod}}))) \tag{6.23}$$

It is a distribution defined within an infinite interval, which depends on two parameters: on the mode x_{mod} and the parameter $c > 0$. By differentiating the distribution function we obtain the probability density function in the form

$$\varphi(x) = c \ \exp(-c(x - x_{\text{mod}}) - \exp(-c(x - x_{\text{mod}}))) \tag{6.24}$$

Both these parameters are related to the mean μ and standard deviation σ

$$x_{\text{mod}} = \mu - 0.577\sigma\frac{\sqrt{6}}{\pi} \tag{6.25}$$

$$c = \frac{\pi}{\sigma\sqrt{6}} \tag{6.26}$$

The skewness and kurtosis of the distribution are constant: $\alpha = 1.14$, $\varepsilon = 2.4$.

An important feature of the Gumbel distribution is the easy transformation of the distribution function $\Phi(x)$ of an original random variable having the mean μ and standard deviation σ to the distribution function $\Phi_N(x)$ for the maxima of populations that are N times greater than the original population. If individual original populations constituting a new N times greater population are mutually independent, then the distribution function $\Phi_N(x)$ is given as

$$\Phi_N(x) = (\Phi(x))^N \tag{6.27}$$

By the substitution of Eq. (6.23) into Eq. (6.27) $\Phi_N(x)$ can be written as

$$\Phi_N(x) = \exp(-\exp(-c(x - x_{\text{mod}} - \ln N/c))) \tag{6.28}$$

It follows from Eqs. (6.23) and (6.27) that the mean μ_N and standard deviation σ_N of the maxima of the new N times greater population are

$$\mu_N = \mu + \ln(N/c) = \mu + 0.78 \ \ln(N\sigma), \sigma_N = \sigma \tag{6.29}$$

Thus the standard deviation of the original population does not change and $\sigma_N = \sigma$, but the mean μ_N is greater than the original value μ by $\ln(N/c)$.

The distribution function of type I, for the minimal values distribution (Gumbel distribution of minimum values) has the form

$$\Phi(x) = 1 - \exp(-\exp(-c(x_{\text{mod}} - x))) \tag{6.30}$$

This distribution is symmetric to the distribution of maximal values given by Eq. (6.23). It is therefore also defined within an open interval and depends on two parameters: on the mode x_{mod} and parameter $c > 0$. By differentiating the distribution function we obtain the probability density function in the form

$$\varphi(x) = c \ \exp(-c\,(x_{\text{mod}} - x) - \exp(-c\,(x_{\text{mod}} - x))) \tag{6.31}$$

Both these parameters can be assessed from the mean μ and standard deviation σ

$$x_{\text{mod}} = \mu + 0.577\sigma\frac{\sqrt{6}}{\pi} \tag{6.32}$$

$$c = \frac{\pi}{\sigma\sqrt{6}} \tag{6.33}$$

Example 6.6. One-year maxima of wind pressure are described by a Gumbel distribution with a mean $\mu_1 = 0.35$ kN/m^2, $\sigma_1 = 0.06$ kN/m^2. The corresponding parameters of 50-year maximum value distribution, i.e. parameters μ_{50} and σ_{50}, follow from Eq. (6.29)

$$\mu_{50} = 0.35 + 0.78 \times \ln(50 \times 0.06) = 0.53 \text{ kN/m}^2, \quad \sigma_{50} = 0.06 \text{ kN/m}^2$$

Figure 6.5 shows both the distributions of 1-year and 50-year maxima of wind pressure described by a Gumbel distribution.

The probability density functions of the minimum values are symmetric to the shape of maximal values relative to the mode x_{mod}, as is apparent from Fig. 6.6.

In a similar way, type II distribution, the so-called Fréchet distribution, and type III distribution, the so-called Weibull distribution, are defined. All three types of distribution complement each other with regard to possible values of the skewness α. Each type covers a certain area of skewness, as shown in Fig. 6.7.

Types I and II of the extreme values distribution are often applied in the description of quantities of which the maximal values are studied (actions), and type III distribution is applied for quantities of which the minimal values are studied (e.g. strength and other material properties).

Types I and II of the extreme values distribution are often applied in the description of quantities of which the maximal values are studied (actions), and type III distribution is applied for quantities of which the minimal values are studied (e.g. strength and other material properties).

6.6 Basic Rules for Selecting Distribution

Commonly used continuous distributions (normal, two- and three-parameter log-normal distribution, gamma and beta distribution) may be selected using a simple guide based on two basic parameters: relative measure of variance – the coefficient of variation V; and a measure of asymmetry – the coefficient of skewness α. The basic rules may be summarized as follows:

1. If the skewness α is close to zero, $\alpha \approx 0$ (the distribution is symmetric) then most likely the best distribution to use would be the normal distribution. However, it should be remembered that the definition domain of the normal distribution is infinite $<-\infty, \infty>$ and, generally there is a nonzero probability

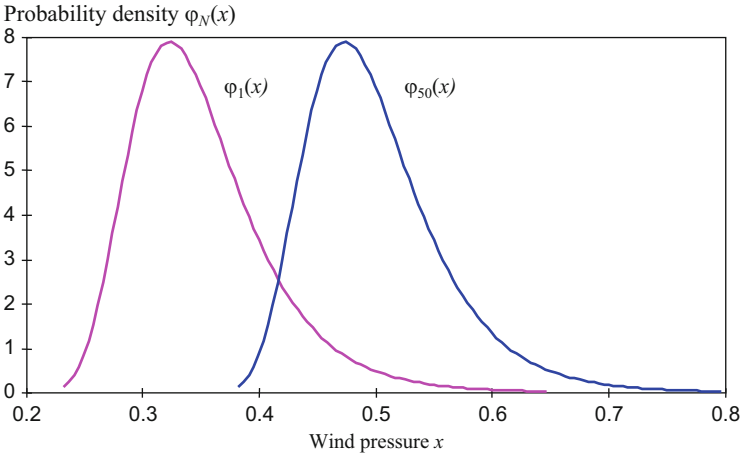


Fig. 6.5 Distribution of maximum wind pressure over the periods of 1 year and 50 years

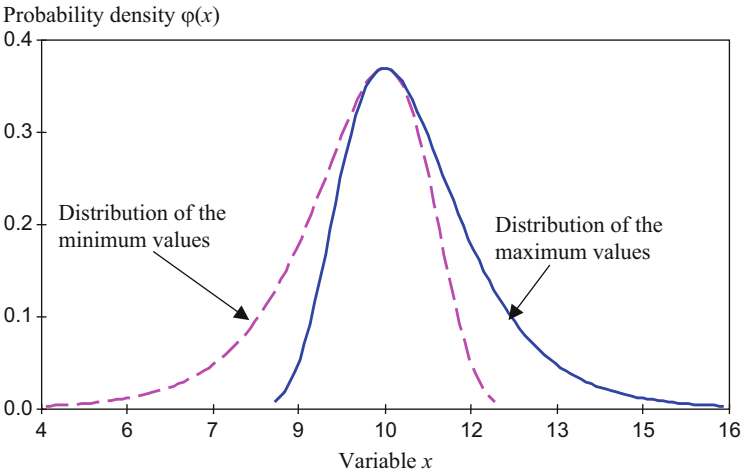


Fig. 6.6 The Gumbel distribution of the minimum and maximum values

of occurrence of negative values (not negligible when the coefficient of variation $V > 0.2$). Then truncated (one sided or two sided limited) normal distribution or beta distribution may be applied.

2. If the skewness α is not negligible, $\alpha > 0$ (the distribution is asymmetric) then several distribution could be used: the two-parameter log-normal, the three-parameter log-normal, the gamma and beta distributions. The two-parameter log-normal distribution, the three-parameter log-normal distribution and gamma distributions are one sided limited distribution, whereas the beta distribution is two sided limited distribution. The two-parameter log-normal and gamma

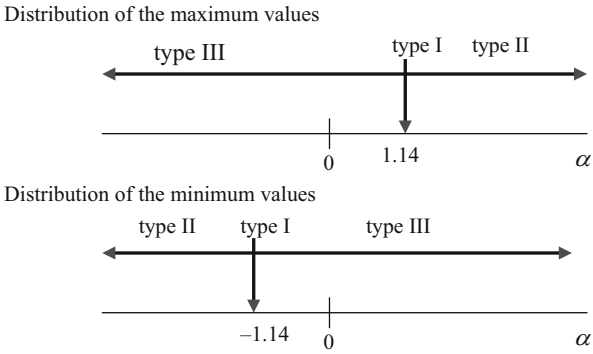


Fig. 6.7 Types of distribution of extreme values versus the skewness α

distributions are limited by zero, the three-parameter lognormal distribution by lower bound (if the skewness is positive) or upper bound (if the skewness is negative).

Figure 6.8 indicates possible combinations of the coefficient of variation V (horizontal axis) and the positive coefficient of skewness α (vertical axis) that could be accepted by selected continuous distributions: two- and three-parameter log-normal distribution, gamma and beta distribution. In addition to these distributions the Gumbel distribution is also included in Fig. 6.8; its skewness $\alpha = 1.14$ (independent of the coefficient of variation V) is indicated by a horizontal dashed line.

Other types of distribution less frequently applied in engineering and science may be found in books [3, 4, 7, 8]. A brief review of conventional distributions provides Appendix 6.

Example 6.7. Consider a non-negative random variable described in Example 6.5. For some physical reasons the random variable has only positive values limited by zero and an unknown upper bound. Assume the mean $\mu = 25$ mm, a standard deviation 10 mm (coefficient of variation $V = 0.40$), and a skewness $\alpha = 0.5$. As the skewness α is significant (not negligible), however less than $2\ V$, $\alpha < 2\ V$, it follows from Fig. 6.8 that the beta distribution with lower bound at the origin seems to be the appropriate type of continuous distribution. Its upper bound is 98.326 (given by Eq. (6.22), see also Example 6.5).

A second possible distribution is the three-parameter log-normal distribution, which is very general and certainly can take account of the given parameters. However in that case the lower bound is a negative value. It follows from Table 6.1 that the standardised value of the lower bound is $u_0 = -6.05$ that therefore $x_0 = 25 - 6.05 \cdot 10 = -35.5$. Consequently, there is some probability of occurrence of negative values X that can be calculated from Eqs. (6.8) and (6.10) as $\mathrm{P}(X < 0) = \Phi_X(0) \approx 0.0014$.

Another possibility is to use two-parameter log-normal distribution (with the lower bound at zero). However, due to its high coefficient of variation $V_X = 0.4$ this distribution has a relatively high skewness equal to $3\ V_X + {V_X}^3 = 1.264$

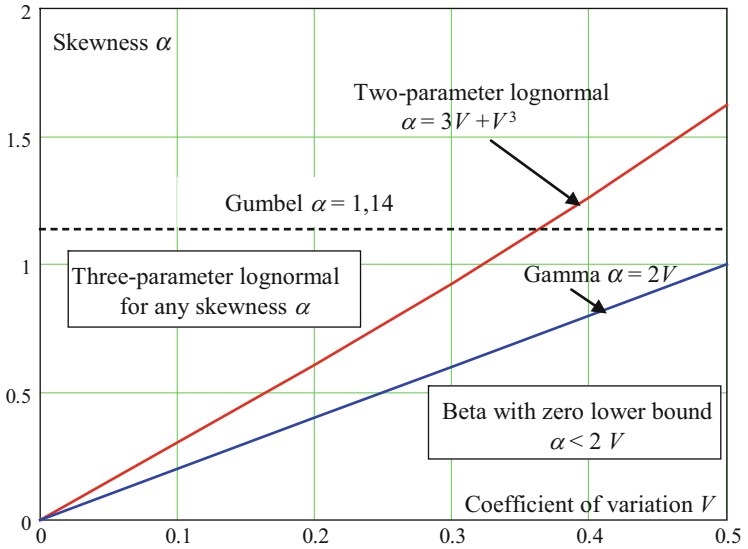


Fig. 6.8 Combinations of the coefficient of variation V and skewness α allowed for by various continuous distributions

(Eq. (6.11)). This one sided limited distribution may well fit the left tail of the distribution but at the same time it amplifies the positive deviations of the random variable from its mean (there is no upper bound of the distribution).

References

1. Ang, A.H.-S., Tang, W.H.: Probabilistic Concepts in Engineering. Emphasis on Applications to Civil Environmental Engineering. Wiley, New York (2007)
2. Devore, J., Farnum, N.: Applied Statistics for Engineers and Scientists. Thomson, London (2005)
3. Dunin-Barkovskij, I.V., Smirnov, N.V.: The Theory of Probability and Mathematical Statistics in Engineering. Technical and Theoretical Literature, Moscow (in Russian) (1955)
4. Gurskij, E.I.: The Theory Probability with Elements of Mathematical Statistics. Higher School, Moscow (in Russian) (1971)
5. Holický, M., Vorlíček, M.: Distribution Asymmetry in Structural Reliability. Acta Polytechnica **35**(3), 75–85 (1995)
6. Holicky, M.: Reliability Analysis for Structural Design. SUNN MeDIA, Stellenbosch (2009)
7. Stewart, M.G., Melchers, R.E.: Probabilistic Risk Assessment of Engineering Systems. Chapman & Hall, London (1997)
8. Melchers, R.E.: Structural Reliability: Analysis and Prediction. Wiley, New York (1999)

Chapter 7
Functions of Random Variables

Functions of random variables defining resulting random variables as functions of several input random variables regularly enter many engineering and scientific applications. The elementary functions of a single continuous variable and two or more independent variables, reviewed in Appendix 4, are supplemented by functions several random variables. A special function of a single random variable is the extreme value of samples taken from a population described by various types of so-called extreme value distributions. These distributions play a substantial role in a number of practical applications. Another important function of a random variable is the updating of its probability distribution when newly obtained information is taken into account. This procedure is developed as an extension of Bayes´ theorem. Finally, the distribution of a sum of several random variables is discussed in conjunction with the central limit theorem.

7.1 Function of a Single Random Variable

The functions of random variables enter many engineering and scientific applications. The distributions of the resulting random variables and their parameters are derived in detail in publications [1–3]. The following short review of basic rules and computational procedures are adapted from the description provided in book [4], paper [5] (for mutually independent variables) paper [6] (for dependent variables).

The general form of a function Z (resulting variable) of a single random variable X is expressed as

$$Z = \mathrm{f}(X) \tag{7.1}$$

Assuming that $\mathrm{f}(X)$ is a single value function (one-to one mapping), then a given value z of the random variable Z corresponds to a particular value $x = \mathrm{f}^{-1}(z)$, where $\mathrm{f}^{-1}(z)$ denotes the inverse function to $\mathrm{f}(x)$ or inverse mapping of z to x.

M. Holický, *Introduction to Probability and Statistics for Engineers*,
DOI 10.1007/978-3-642-38300-7_7, © Springer-Verlag Berlin Heidelberg 2013

If the transformation function f(x) is an increasing function of x, then for a given value $z = \mathrm{f}(x)$ the probabilities $\mathrm{P}(Z \leq z)$ and $\mathrm{P}(X \leq x)$ are equal, and the probability distribution function $\Phi_Z(z)$ of the transformed variable Z may be expressed in terms of the distribution function $\Phi_X(x)$ simply as

$$\Phi_Z(z) = \Phi_Z(\mathrm{f}(x)) = \Phi_X(x) = \Phi_X(\mathrm{f}^{-1}(z)) \tag{7.2}$$

This equation can be written in the integral form (see Eq. (4.7)) as

$$\Phi_Z(z) = \int_{-\infty}^{z} \varphi_z(z)\mathrm{d}z = \int_{-\infty}^{\mathrm{f}^{-1}(z)} \varphi_X(x)\mathrm{d}x = \int_{-\infty}^{(z)} \varphi_X(\mathrm{f}^{-1}(z))\mathrm{d}x \tag{7.3}$$

It follows from Eqs. (7.2) and (7.3) that the probability elements $\varphi_Z(z)\mathrm{d}z$ and $\varphi_X(x)\mathrm{d}x$ are also equal

$$\varphi_Z(z)\mathrm{d}z = \varphi_X(x)\mathrm{d}x \tag{7.4}$$

Equation (7.4) clearly indicates that the relationship between the probability density functions $\varphi_Z(z)$ and $\varphi_X(x)$ depends on the differentials $\mathrm{d}z$ and $\mathrm{d}x$ of continuous variables Z and X. Consequently the probability density $\varphi_Z(z)$ of the transformed variable Z is related to the probability density functions $\varphi_X(x)$ as follows

$$\varphi_Z(z) = \varphi_X(x)\frac{\mathrm{d}x}{\mathrm{d}z} \tag{7.5}$$

Equation (7.4) may be generalised for a monotonous (both increasing and decreasing) transformation function f(X) as follows

$$\varphi_Z(z) = \varphi_X(x)\left|\frac{\mathrm{d}x}{\mathrm{d}z}\right| = \varphi_X(\mathrm{f}^{-1}(z))\left|\frac{\mathrm{df}^{-1}(z)}{\mathrm{d}z}\right| \tag{7.6}$$

Example 7.1. A random variable X has a normal distribution $\mathrm{N}(\mu,\sigma)$ having the mean μ and standard deviation σ:

$$\varphi_X(x)\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

A new random variable U (standardized normal variable) is introduced by transformation formula defining standardised random variables

$$U = \mathrm{f}(X) = \frac{X-\mu}{\sigma}$$

The differential $du = dx/\sigma$ and the ratio of the differentials $dx/du = \sigma$. Then it follows from Eq. (7.5) that the new probability density function $\varphi_U(u)$ of the transform variable U is

$$\varphi_U(u) \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}u^2\right]$$

This formula is already known from Chap. 6 (Eq. (6.3)) for a standardised random variable having the normal distribution N(0,1) (the mean equals zero and the standard deviation equals 1).

Example 7.2. The wind pressure P can be expressed in terms of wind speed S as

$$P = \mathrm{f}(S) = ks^2$$

Here k denotes a quantity dependent on several characteristics of structure and its surroundings, but independent of the velocity S. The ratio of differentials ds/dp follows from the transformation formula as

$$\frac{ds}{dp} = \frac{1}{2\sqrt{kp}}$$

Then the probability density function $\varphi_P(p)$ of the wind pressure can be derived from the density of wind velocity $\varphi_S(s)$ using Eq. (7.5) as

$$\varphi_P(p) = \frac{\varphi_S\left(\sqrt{p/k}\right)}{2\sqrt{kp}}$$

The moment parameters of the transformed random variable P may be obtained by integration considering the above derived probability density function $\varphi_P(p)$. It may, however, require approximations using numerical procedures.

7.2 Function of Two Random Variables

The probability distribution of a function of two variables may be derived in much the same way in the case of one random variable. Consider a function of two mutually independent random variables X and Y:

$$Z = \mathrm{f}(X, Y) \tag{7.7}$$

The probability density function $\varphi_Z(z)$ of the resulting random variable Z can be expressed as

$$\varphi_Z(z) = \int_{-\infty}^{\infty} \varphi_{X,Y}(\mathrm{f}^{-1}(z,y),y)\left|\frac{\partial \mathrm{f}^{-1}(z,y)}{\partial z}\right|\mathrm{d}y$$
$$= \int_{-\infty}^{\infty} \varphi_{X,Y}(x,\mathrm{f}^{-1}(x,z))\left|\frac{\partial \mathrm{f}^{-1}(x,z)}{\partial z}\right|\mathrm{d}x \tag{7.8}$$

In general, Eq. (7.8) may not be easily applied and usually numerical integration has to be used. That is why different approximate techniques are frequently applied in engineering applications. A simple procedure for assessing moment parameters of the resulting random variable Z is described in the following Section.

Example 7.3. As an example consider the sum of two random variables

$$\mathrm{Z} = X + Y$$

The inverse functions and their derivatives are

$$x = z - y \quad \text{and} \quad \mathrm{y} = z - x$$

$$\frac{\partial x}{\partial z} = \frac{\partial y}{\partial z} = 1$$

The probability density function $\varphi_Z(z)$ follows from Eq. (7.8) as

$$\varphi_Z(z) = \int_{-\infty}^{\infty} \varphi_{X,Y}(z-y,y)\,\mathrm{d}y = \int_{-\infty}^{\infty} \varphi_{X,Y}(x,z-x)\mathrm{d}x$$

Furthermore, if X and Y are mutually independent random variables, then the probability joint density of the two random variables equals to the product of densities of each random variable, thus

$$\varphi_{X,Y}(x,y) = \varphi_X(x)\varphi_Y(y)$$

The resulting probability density $\varphi_Z(z)$ can then be written as

$$\varphi_Z(z) = \int_{-\infty}^{\infty} \varphi_X(z-y),\varphi_Y(y)\,\mathrm{d}y = \int_{-\infty}^{\infty} \varphi_X(x)\varphi_Y(z-x)\,\mathrm{d}x$$

7.3 Parameters of Functions of Independent Random Variables

Another approach to investigating the functions of mutually independent random variables, and to assessing the probability distribution of the resulting random variable Z, is to estimate the basic moment parameters (the mean μ_Z, standard deviation σ_Z and skewness α_Z) using the Taylor expansion of the transformation function f(X, Y,. . .) into a power series [5].

Thus instead of deriving probability distribution of the transformed random variable Z, moment parameters of the resulting random variable are estimated first and then used to approximate the distribution of the variable Z. Three basic moment parameters are considered for all random variables in the following: the mean μ, the standard deviation σ and the skewness α. (supplemented by appropriate subscripts).

Consider a function of independent random variables X, Y, . . ., resulting in the variable Z given by a general relationship

$$Z = \mathrm{f}\ (X,\ Y, \ldots) \tag{7.9}$$

The variable Z is therefore also a random variable having moment parameters μ_Z, σ_Z, α_Z, for which (using the Taylor expansion of f (X, Y,. . .) into a power series) approximate relationships may be found

$$\mu_Z = \mathrm{f}_1(\mu_X, \mu_Y, \ldots \sigma_X, \sigma_Y, \ldots, \alpha_X, \alpha_Y, \ldots) \tag{7.10}$$

$$\sigma_Z = \mathrm{f}_2(\mu_X, \mu_Y, \ldots \sigma_X, \sigma_Y, \ldots, \alpha_X, \alpha_Y, \ldots) \tag{7.11}$$

$$\alpha_Z = \mathrm{f}_3(\mu_X, \mu_Y, \ldots \sigma_X, \sigma_Y, \ldots, \alpha_X, \alpha_Y, \ldots) \tag{7.12}$$

Appendix 4 provides relationships (7.10, 7.11 and 7.12) for elementary forms of functions (7.9) considering one or two independent random variables X and Y. These relationships may be effectively applied to simplify a number of common expressions describing the behaviour of the resulting random variable Z, such as a capacity of structural members.

Example 7.4. Consider a simple product of two random variables X and Y. Equation (7.9) is then written as

$$Z = aX + bY + c$$

Here symbols a, b and c denote constants. Using formulae in Appendix 4 the following relationships for the basic moment parameters of Z can be found

$$\mu_Z = a\mu_X + b\mu_Y + c$$

$$\sigma_Z^2 = a^2\sigma_X^2 + b^2\sigma_Y^2$$

$$\alpha_z = \frac{a^3\sigma_X^3\alpha_X + b^3\sigma_Y^3\alpha_Y}{\sigma_Z^3}$$

For a numerical illustration let us consider a difference of two random variables

$$\mathrm{Z} = X - Y$$

Parameters of X (lognormal distribution) are $\mu_X = 100$, $\sigma_X = 10$, $\alpha_X = 0.301$. Parameters of X (Gumbel distribution) are: $\mu_Y = 50$, $\sigma_Y = 10$, $\alpha_Y = 1.14$.

$$\mu_Z = 100 - 50 = 50$$

$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 = 10^2 + 10^2 = 14.14^2$$

$$\alpha_Z = \frac{\sigma_X^3\alpha_X + \sigma_Y^3\alpha_Y}{\sigma_Z^3} = \frac{10^3 \times 0.301 - 10^3 \times 1.14}{14.14^3} = -0.30$$

Note that due to the difference of input variables the resulting variable Z has a negative skewness -0.30.

Example 7.5. Consider a simple product of two random variables X and Y. Equation (7.9) is then written as

$$\mathrm{Z} = X \times Y$$

Using Appendix 4 the following relationships for the basic parameters may be found

$$\mu_Z = \mu_X \times \mu_Y$$

$$V_Z^2 = V_X^2 + V_Y^2 + V_X^2V_Y^2$$

$$\alpha_Z = \frac{V_X^3\alpha_X + V_Y^2\alpha_Y + 6\,V_X^2w_Y^2}{(V_X^2 + V_Y^2 + V_X^2V_Y^2)^{3/2}} = \frac{V_X^3\alpha_X + V_Y^2\alpha_Y + 6\,V_X^2V_Y^2}{V_Z^3}$$

Note that the product Z of two random variables X and Y having the normal distribution ($\alpha_X = \alpha_Y = 0$) is not a normal variable. If, for example, $V_X = 0.1$ and $V_Y = 0.2$, then it follows, using the above formulae, that the resulting coefficient of variation $V_Z = 0.22$ and the skewness $\alpha_Z = 0.11$.

Example 7.6. Considering wind pressure $P = ks^2$ described in Example 7.2, the moment parameters of pressure P may be well approximated as follows:

$$\mu_P \cong k\left(\mu_S^2 + \sigma_S^2\right)$$

$$\sigma_P \cong k2\sigma_S\left(\mu_S^2 + \mu_S\sigma_S\alpha_S\right)^{1/2}$$

$$\alpha_P \cong \frac{8\mu_S^3\sigma_S^3(\alpha_S + 3V_S)}{\sigma_P^3}$$

The above expression may be effectively used to approximate a theoretical model for distribution of the pressure P. If, for example, the mean $\mu_S = 30$ m/s, standard deviation $\sigma_S = 3$ m/s (coefficient of variation $V_S = 0.1$) and skewness $\alpha_S = 1.14$ (Gumbel distribution), then

$$\mu_P \cong 909k$$

$$\sigma_P \cong 190k$$

$$\alpha_P \cong 1.22$$

Compared with the original variable S, the coefficient of variation and skewness of the transformed variable P increased: $V_P \approx 0.21$ and $\alpha_P \approx 1.22$.

7.4 Parameters of Functions of Dependent Random Variables

Parameters of functions of two mutually dependent variables can be found in a paper [6]. Considering again the transformation function (7.10), basic moment parameters of the resulting variable Z formulae are provided for basic moment parameters (see Eqs. (3.9, 3.10 and 3.11)):

- The mean μ_Z
- The central moments of the second power (variance) $\mu_{Z2} = \sigma_Z{}^2$
- The central moment of the third power μ_{Z3}, from which the skewness is derived as $\alpha_Z = \mu_{Z3}/\sigma_Z{}^3$

The above moments of the variable Z are expressed as function of the relevant moments of variables X and Y:

- The means μ_X and μ_Y
- The central moments μ_{Xi} and μ_{Yi} (for $i = 1,2,3,4$)
- The product moment $\mu_{Xi,\,Yj}$ (for $i = 1,2,3$ and $j = 1,2,3$)

An example of the transformation function and the resulting moments taken from the paper [6] is shown below

$$Z = aX + bY + c \tag{7.13}$$

$$\mu_Z = a\ \mu_X + b\ \mu_Y + c \tag{7.14}$$

$$\mu_{Z,2} = a^2\mu_{X2} + b^2\mu_{Y2} + 2a\ b\mu_{X1,Y1} \tag{7.15}$$

$$\mu_{Z3} = a^3\mu_{X3} + b^3\mu_{Y3} + 3a^2b\mu_{X2,Y1} + 3ab^2\mu_{X1,Y2} \tag{7.16}$$

Similar expressions are available for a number of linear and nonlinear functions commonly encountered in engineering applications.

Example 7.7. Consider a sum of two normally distributed and dependent random variables X and Y given by Eq. (7.13). The central moments of the third power and skewness of both these variables is zero. Then the moment parameters of the variable Z follow from Eqs. (7.14, 7.15 and 7.16) as

$$\mu_Z = a\mu_X + b\mu_Y + c$$

$$\sigma_Z{}^2 = a^2\sigma_X{}^2 + b^2\sigma_Y{}^2 + 2a\ b\rho\sigma_X\sigma_Y$$

Here $\rho = \mu_{X1,Y1}/(\sigma_X\ \sigma_Y)$ denotes the coefficient of correlation. Note that in this case of normally distributed random variables X and $Y\mu_{Z3} = \mu_{X3} = \mu_{Y3} = \mu_{Z2,Y1} = \mu_{X1,Y2} = 0$. Consequently the skewness of the variable Z is zero.

In case of a simple difference $Z = X-Y$ ($a = 1$, $b = -1$ and $c = 0$), the above expressions become

$$\mu_Z = \mu_X - \mu_Y$$

$$\sigma_Z{}^2 = \sigma_X{}^2 + \sigma_Y{}^2 - 2\rho\sigma_X\sigma_Y$$

7.5 Updating of Probability Distributions

A special case of transformation of a random variable X is the updating of its probability distribution. The prior probability density function $\varphi_X(x)$ (provided by previous experience) may be updated using new information expressed by a likelihood function $L(I/x)$. If the prior probabilities are described by a continuous probability density function $\varphi_X(x)$ of a random variable X likelihood by a function $L(I/x)$, where I denotes the outcomes of additional investigation I, then *a posteriori* (updated) probability density $\varphi_X(x|I)$ may be derived from (7.17) by using integration instead of the summation as

$$\varphi_X(x|I) = \frac{\varphi_X(x)L(I|x)}{\int \varphi_X(x)L(I|x)\mathrm{d}x} \tag{7.17}$$

Note that the likelihood $L(I|x)$ is a function describing the potential (it may not be probability) that the outcome of the updating investigation I (information obtained from I) is due to the occurrence of x. Formulae (7.17) can be used for the updating of distribution functions when additional experimental investigations are used for assessing new or existing structures.

Example 7.8. Assume that a variable X has a normal prior distribution with probability density function $\varphi_X(x)$ having the mean μ and the standard deviation σ. Additional investigation indicated that the likelihood function $L(I|x)$ is described by a general three-parameter log-normal distribution having the same standard deviation σ but the mean equal to $\mu + 0.5\,\sigma$ and the skewness $\alpha = 1$. Using a numerical integration it follows that the updated distribution $\varphi_X(x|I)$ has the following moment parameters

$$\mu_{X|I} = \mu_X + 0.18\sigma_X$$

$$\sigma_{X|I} = 0.64\sigma_X$$

$$\alpha_{X|I} = 0.39$$

Figure 7.1 shows the prior probability density $\varphi(u)$, likelihood $L(I|u)$ and the updated probability density function $\varphi(u|I)$ using standardized random variable U.

It follows from Fig. 7.1 that the updated distribution has considerably lower variability than the prior distribution. Obviously updating of probability distributions may be extremely effective when assessing the characteristic values of the resistance variables using additional tests.

7.6 Central Limit Theorem

The central limit theorem has a number of variants. The following description is devoted only to practical applications of the theorem devoted to the sum of a number of random variables. In its classical form the central limit theorem states [7] that the mean of a sufficiently large number of independent observations, each taken from a certain population with a finite mean and variance, is approximately normally distributed. Moreover the distribution has the same mean as the parent distribution and a variance equal to the variance of the parent distribution divided by the sample size.

Thus, in this variant, the central limit theorem considers a sample taken from one distribution (generally non-normal) of the mean μ and variance σ^2. It can be

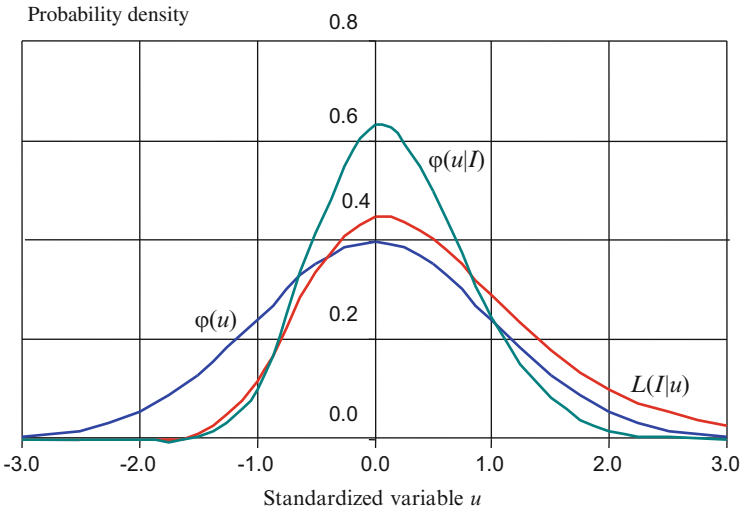


Fig. 7.1 Prior probability density $\varphi(u)$, likelihood $L(I|u)$ and the updated probability density function $\varphi(u|I)$

shown [3] that with increasing sample size n, distribution of the sample mean m approaches the normal distribution $N(\mu,\sigma/\sqrt{n})$ with the same mean μ as the parent distribution and with the reduced standard deviation $\sigma/\sqrt{n}$. The formal representation of this finding may be written like this:

$$m = \frac{\sum_1^n x_i}{n} \sim \mathrm{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \tag{7.18}$$

In other variants of the central limit theorem, convergence of the mean to the normal distribution also occurs for non-identical parent distributions, as long as they comply with certain conditions. This outcome holds even when the parent distributions are non-normal. So, although the distribution of the sample mean reflects the properties of the parent distribution (particularly its location μ), the shape of this sampling distribution is symmetrical (normal) and primarily affected by the sample size n.

In general, the distribution of the means tends to be normal as the sample size increases regardless of the distribution from which the mean is taken, except when the moments of the distribution do not exist. However, all practical distributions applied in engineering and science have definite moments, and thus the central limit theorem applies. Because of that remarkable result the central limit theorem plays an important role in many statistical procedures, including the estimation of the population parameters, testing of statistical hypothesis and quality control.

The central limit theorem may be interpreted in a broad sense as follows: most natural phenomena are dependent on a number of random variables X_i and may be approximately described by a sum $Y = \sum X_i$. If the random variables X_i are

mutually independent variables of an identical distribution with the mean μ, variance σ^2 and with an existing third moment, then the variable Y approaches a normal distribution [3] of the mean $n\mu$ and variance $n\sigma^2$.(the standard deviation $\sigma\sqrt{n}$). The formal representation may be written as

$$Y = \sum_1^n X_i \sim \mathrm{N}\left(n\mu, \sqrt{n}\sigma\right) \tag{7.19}$$

Many natural phenomena in the real world may be approximated by a sum of other random variables and its distribution, as indicated in Eq. (7.19). Consequently, such variables are expected to follow some kind of normal distribution (with different means and standard deviations) depending on a number of random variables X_i. This finding seems to be an extremely important piece of information for the general understanding of the natural phenomena, one that depends on many uncertainties.

However, practical experience from engineering and science clearly indicates that some random variables follow somewhat asymmetric (non-normal) distribution patters; for example, variables (like strength of materials) that are dependent on the product rather than on a sum of other variables. Then the resulting variable follows asymmetric log-normal distribution. Obviously some natural phenomena are the results of more complex relationships, and need to be described by theoretical models based on experimental evidence. Nevertheless, in many cases a normal distribution is a good approximation and should be considered whenever there is a lack of convincing statistical data.

Example 7.9. Consider a population (generally non-normal, say two parameter log-normal LN2) of the mean $\mu = 100$ and variance $\sigma^2 = 225$ (the standard deviation $\sigma = 15$, coefficient of variation $V = \sigma/\mu = 0.15$). If the sample size is limited to $n = 9$, then in accordance with Eq. (7.18) the sample mean tends to approach normal distribution

$$m = \frac{\sum_1^9 x_i}{9} \sim \mathrm{N}(100, 5)$$

The standard deviation of the resulting normal distribution is given as $\sigma/\sqrt{n} = 15/3 = 5$.

Figure 7.2 shows both distributions, the parent distribution LN2(100,15) and the distribution of the sample mean N(100,5). Note that the parent distribution LN2 (100,15) is asymmetrical (positive skewness $\alpha = 3\ V + V^3 = 0.453$), the sampling distribution of the mean is symmetrical (normal) N(100,5)

Example 7.10. Consider the sum of $n = 4$ independent random variables X_i

$$Y = \sum_1^4 X_i, \quad \mu_{X_i} = 0.5, \quad \sigma_{X_i} = 0.1$$

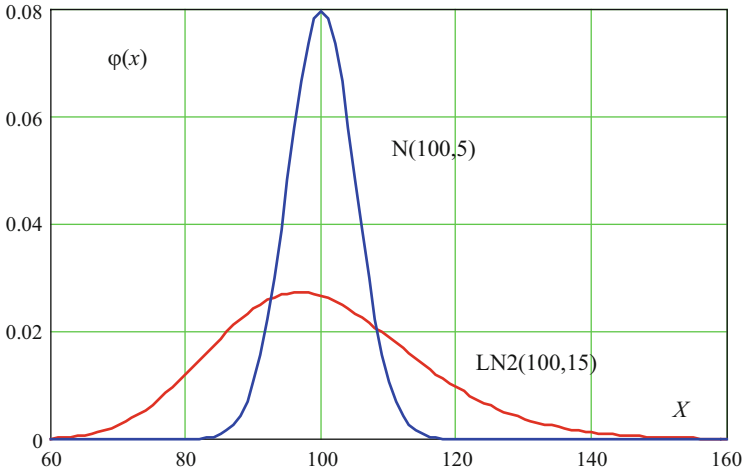


Fig. 7.2 Probability density $\varphi(x)$ of the parent distribution LN2(100,15) and sampling distribution N(100,5) of the mean m for the sample size $n = 9$

The mean μ_Y and standard deviation σ_Y follow from Eq. (7.19) as

$$\mu_Y = 4 \times 0.5 = 2, \quad \sigma_Y = \sqrt{4} \times 0.1 = 0.2$$

Note the difference between the sum of mutually independent variables X_i and the multiple of a single variable $Y = 4 \times X$, where X is a random variable having the same distribution as the input variables X_i which in the sum are considered to be identical (or perfectly dependent). It follows from Annex 4 that in this case the mean μ_Y is not changed ($\mu_Y = 4 \times \mu_Y$), but the standard deviation σ_Y is different. The standard deviation σ_Y of the resulting variable Y is now given simply as the product $n \times \sigma_X$ of the multiplication factor n and the standard deviation $\sigma_X = \sigma_{X_i}$ (not $\sqrt{n} \times \sigma_X$ as in case of independent variables X_i). Thus

$$\mu_Y = 4 \times 0.5 = 2, \quad \sigma_Y = 4 \times 0.1 = 0.4$$

In the first case a sum of independent random variables X_i the coefficient of variation $V_Y = 0.2/2 = 0.1$, in the second case $V_Y = 0.4/2 = 0.2$. There is also a difference in distribution of the variable Y. In the first case the sum Y tends to be normally distributed, in the second case the simple multiple Y has the same type of distribution as X (Y has the same skewness as X, see Annex 4). This example clearly illustrates the significance of mutual dependency of input random variables X_i on the distribution of the resulting variable Y.

7.7 Extreme Value Distribution

The extreme values (the maximum and minimum values) of samples are of great interest and importance to many engineering and scientific applications. In particular several natural phenomena (flooding, snow and temperature extremes) can be well described by one of the extreme value distributions. This Section is devoted to a short description of the classical extreme value theory and models.

Consider a set of independent and identically distributed random variables (X_1, X_2, …, X_n) having the probability density function $\varphi_X(x)$ and distribution function $\Phi_X(x)$. Samples (x_1, x_2, …, x_n) of the size n are created in such a way that each x_i of each sample is taken from the corresponding population of variables X_i,. A hypothetical infinite number of the samples (x_1, x_2, …, x_n) represents set of random variables (X_1, X_2, …, X_n). The maximum values of the samples may be then expressed as the maximum of the random variables (X_1, X_2, …, X_n), thus

$$Y_n = \max(X_1, X_2, \ldots X_n) \tag{7.20}$$

The distribution function $\Phi_{Yn}(y)$ of the maximum value Y_n is therefore defined as

$$\Phi_{Y_n}(y) = \mathrm{P}(X_1 \leq y, X_2 \leq \mathrm{y}, \ldots X_n \leq \mathrm{y}) = [\Phi_X(y)]^n \tag{7.21}$$

Here the assumption of independent random variables (X_1, X_2, …X_n) is taken into account (see Eq. (2.22) for the probability of intersection of independent events).

The probability density function $\varphi_{Y_n}(y)$ is derived from Eq. (7.21)

$$\varphi_{Y_n}(y) = \frac{\mathrm{d}\Phi_{Y_n}(y)}{dy} = n[\Phi_X(y)]^{n-1}\varphi_X(y) \tag{7.22}$$

Obviously, the resulting distribution depends on the sample size n and the initial distribution of the variable X.

Similarly the minimum value of samples (x_1, x_2, …, x_n) of the size n, when each x_i is taken from the corresponding population X_i may be written as

$$Y_1 = \min(X_1, X_2, \ldots X_n) \tag{7.23}$$

The distribution function $\Phi_{Y1}(y)$ can now be derived from the complementary (survival) function as

$$1 - \Phi_{Y_1}(y) = \mathrm{P}(X_1 > y, X_2 > y, \ldots X_n > y) = [1 - \Phi_X(y)]^n \tag{7.24}$$

Thus the distribution function is

$$\Phi_{Y_1}(y) = 1 - [1 - \Phi_X(y)]^n \tag{7.25}$$

The probability density function is then

$$\varphi_{Y_1}(y) = \frac{d\Phi_{Y_1}(y)}{dy} = n[1 - \Phi_X(y)]^{n-1}\varphi_X(y) \tag{7.26}$$

Equations (4.22) and (4.26) provide an exact solution for probability density functions of the maximum and minimum value of a sample of n observations taken from any type of initial population (including a normally distributed population). These relationships have been used to derive so-called extreme value distributions, including Gumbel, Weibull and Frechet distributions, as introduced in Sect. 6.5.

Example 7.11. Consider the exponential distribution

$$\Phi_X(x) = 1 - \exp\,(-\lambda x)$$

The corresponding distribution function is

$$\varphi_X(x) = \frac{d\Phi_X(x)}{dx} = \lambda \exp(-\lambda x)$$

The distribution function $\Phi_{Y_n}(y)$ follows from Eq. (7.21) as

$$\Phi_{Y_n}(y) = [\Phi_X(y)]^n = [1 - \exp(-\lambda x)]^n$$

The corresponding probability density function is

$$\varphi_{Y_n}(y) = \frac{\Phi_{Y_n}(y)}{dy} = \lambda n[1 - \exp(-\lambda y)]^{n-1} \exp(-\lambda y)$$

Note that the exponential functions are convenient for modeling extreme value distributions.

Example 7.12. Consider the Gumbel distribution of maximum values defined in Sect. 6.5 as

$$\Phi(x) = \exp\,(-\exp(-c(x - x_{\text{mod}})))$$

The probability density function is given as

$$\varphi_X(x) = \frac{d\Phi_X(x)}{dx} = c \exp(-c(x - x_{\text{mod}})) - \exp(-c(x - x_{\text{mod}}))$$

Using Eq. (7.21) the distribution function of the maximum value of a sample of N observations (capital N is used as in Sect. 6.5) may be derived from Eq. (7.21) as

$$\Phi_{Y_N}(y) = [\Phi_X(y)]^N = \exp(-\exp(-c(x - x \bmod -\ln N/c)))$$

This expression is already provided in Sect. 6.5 by Eq. (6.28). The probability density function follows from Eq. (7.21) as

$$\varphi_{Y_N}(y) = \frac{\mathrm{d}\Phi_N(y)}{dy} = N[\Phi_X(y)]^{N-1}\varphi_X(y)$$

For practical use of this equation the previous two expressions for $[\Phi_X(y)]^{N-1}$ and $\varphi_X(y)$ should be adjusted and substituted.

References

1. Ang, A.H.-S., Tang, W.H.: Probabilistic Concepts in Engineering. Emphasis on Applications to Civil Environmental Engineering. Wiley, New York (2007)
2. Dunin-Barkovskij, I.V., Smirnov, N.V.: The Theory of Probability and Mathematical Statistics in Engineering. (in Russian). Technical and Theoretical Literature, Moscow (1955)
3. Gurskij, E.I.: The Theory Probability with Elements of Mathematical Statistics. (in Russian). Higher School, Moscow (1971)
4. Holický, M.: Reliability Analysis for Structural Design. SUNN MeDIA, Stellenbosch (2009)
5. Vorlíček, M.: Statistical Quantities for Functional Relations used in Building and Construction Research. Stavebnícky časopis IX/8, SAV, Bratislava (1961)
6. Vorlíček, M.: Statistical Parameters of the Function of Mutually Dependent Quantities. Stavebnícky časopis 27/3, SAV, Bratislava (1979)
7. Devore, J., Farnum, N.: Applied Statistics for Engineers and Scientists. Thomson, London (2005)

Chapter 8
Estimations of Population Parameters

The estimation of population parameters from limited sample data is an indispensable part of any engineering and scientific application of probability and mathematical statistics. Based on appropriate sampling distributions, two types of estimate are commonly applied: point and interval estimates. Point estimates of the population mean and variance are obtained as the mean of relevant sampling distributions evaluated for the sample mean and variance. Interval estimates are obtained as the intervals of relevant sampling distribution that cover the population parameters with a given probability called confidence level. Guidance is given on how to specify the sample size of an experimental investigation with the accuracy required for the estimate of the population mean. Notes on estimating the population skewness are also provided. A review of basic formulae used for point estimates of the population mean, variance and skewness is provided in Appendix 1.

8.1 Sampling Distributions

The concept of population and samples has already been introduced in Chap. 3, with references to more detailed descriptions provided in [1–4]. It is necessary to be reminded here that the population is the totality of items under consideration. Depending on the actual conditions, it may have a limited or an unlimited number of items or units. The term "item" or "unit" denotes an actual or conventional object on which a set of observations can be made [5, 6]. When bulk material or continuous material is considered then the unit is a defined quantity of material having a physical or hypothetical boundary (container, time interval)

A precise and comprehensive definition of a population is the first important step in any statistical investigation. The population is commonly characterized by a number of relevant aspects, including time, geographic region, production technology, producer, etc. These aspects are substantial for the correct interpretation of obtained results.

M. Holický, *Introduction to Probability and Statistics for Engineers*,
DOI 10.1007/978-3-642-38300-7_8, © Springer-Verlag Berlin Heidelberg 2013

A sample is one or more units taken from a population and intended to provide information on the population and possibly to serve as the basis for a decision on the population. The number of sampling units is called sample size. If the sampling units are taken from the population in such a way that each unit has the same chance to be taken, then the sample is called a random sample. In what follows only random samples are considered.

The constants providing information about the population are called parameters and denoted here mostly by Greek letters; the corresponding quantities obtained from a sample are called characteristics and are denoted by Roman letters. Obviously sample characteristics (discussed in Chap. 3) are not constants but random variables (called often statistics) that differ from sample to sample and that are described by special types of distributions, called sampling distributions. The most important sampling distributions are available in literature [1–4] and their numerical representations are provided by several software products (EXCEL, STATISTCA, MATHCAD, MATLAB etc.) or from statistical tables available as a public domain on the internet. A concise table of standardized normal distribution is provided in Appendix 7.

The sampling distributions are therefore introduced here very briefly. The distribution of the sample means is described as the normal distribution (this important sampling distribution has been introduced in Chap. 6 and a table of its distribution function is given in Appendix 7). The sample variance is described by χ^2-distribution, *t*-distribution is used for estimation of the means when the population variance is unknown, and *F*-distribution is used for the testing of two sample variances. A short introduction of these sampling distributions is given below; the numerical values used in the examples are obtained from the above-mentioned sources.

8.1.1 χ^2-Distribution

The random variable χ^2 is the sum of the squares of normalised random variables U_i having normal distribution.

$$\chi^2 = \sum\nolimits_{i=1}^{v} U_i^2 \tag{8.1}$$

The distribution depends on one parameter only, $\nu = 1,2,3\ldots$, called the degree of freedom (the number of independent summands). It can be shown that the moment parameters are as follows

$$\mu_{\chi^2} = \nu, \quad \sigma_{\chi^2} = \sqrt{(2\nu)}, \quad \alpha_{\chi^2} = 2\sqrt{(2/\nu)}, \quad \varepsilon_{\chi^2} = 12/\nu \tag{8.2}$$

For example, if the number of the independent variables U_i is $\nu = 9$ then

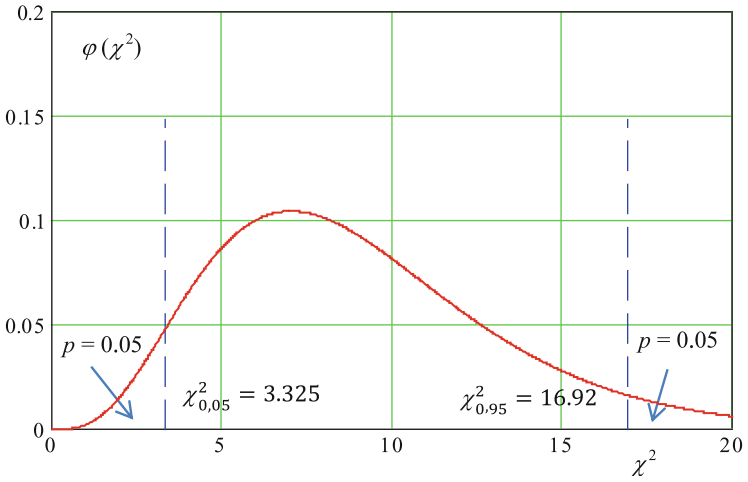


Fig. 8.1 Probability density $\varphi(\chi^2)$ for $\nu = 9$ degree of freedom

$$\mu_{\chi^2} = 9, \quad \sigma_{\chi^2} = 4.24, \quad \alpha_{\chi^2} = 0.47, \quad \varepsilon_{\chi^2} = 1.33.$$

Probability density function $\varphi(\chi^2)$ for $\nu = 9$ is shown in Fig. 8.1.

8.1.2 t-*Distribution*

The random variable t is a ratio of two random variables: the normalised normal variable U and a function of the χ^2 random variable degree of freedom ν:

$$t = \frac{U}{\sqrt{\frac{\chi^2}{\nu}}} \tag{8.3}$$

It can be shown that the moment parameters are as follows

$$\mu_t = 0, \quad \sigma_t = \sqrt{\frac{\nu}{\nu - 2}}, \quad \text{for} \quad \nu > 2, \quad \alpha_t = 0, \quad \varepsilon_t = \frac{6}{\nu - 4} \quad \text{for} \quad \nu > 4 \tag{8.4}$$

Obviously it is a symmetrical distribution around zero ($\mu_t = \alpha_t = 0$). As indicated in Fig. 8.2 with an increasing degree of freedom it approaches the standardised normal distribution N(0,1) (the distribution function is provided in Appendix 7).

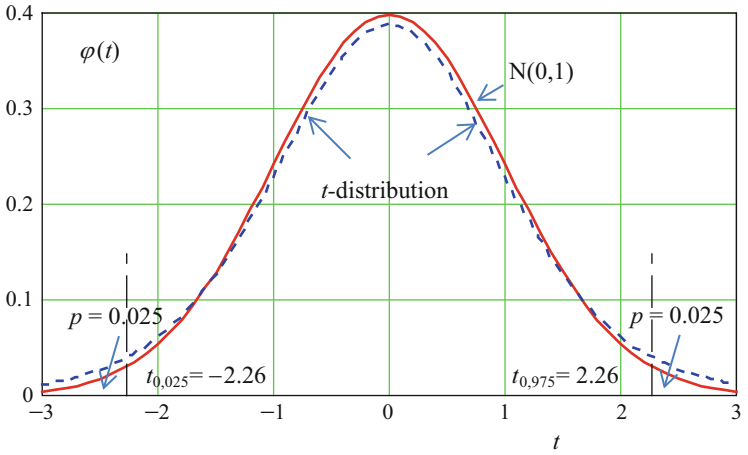


Fig. 8.2 Normal N(0,1) and t-distribution for degree of freedom $\nu = 9$

8.1.3 F-*Distribution*

Random variable F is a fraction of two scaled random variables having χ^2-distribution defined as

$$F = \frac{\frac{\chi_1^2}{\nu_1}}{\frac{\chi_2^2}{\nu_2}} \tag{8.5}$$

It is an asymmetrical distribution having the mode

$$\tilde{F} = \frac{\nu_2}{\nu_1}\frac{\nu_1 - 2}{\nu_2 + 2} \quad \text{for} \quad \nu_1 > 2 \tag{8.6}$$

The moment parameters of F-distribution are as follows

$$\mu_F = \frac{\nu_2}{\nu_2 - 2} \quad \text{for} \quad \nu_2 > 2 \tag{8.7}$$

$$\sigma_F = \frac{\nu_2}{\nu_2 - 2}\sqrt{\frac{2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 4)}} \quad \text{for} \quad \nu_2 > 4 \tag{8.8}$$

$$\alpha_F = \frac{2(2\nu_1 + \nu_2 - 2)}{\nu_2 - 6}\sqrt{\frac{2(\nu_2 - 4)}{\nu_1(\nu_1 + \nu_2 - 2)}} \quad \text{for} \quad \nu_2 > 6 \tag{8.9}$$

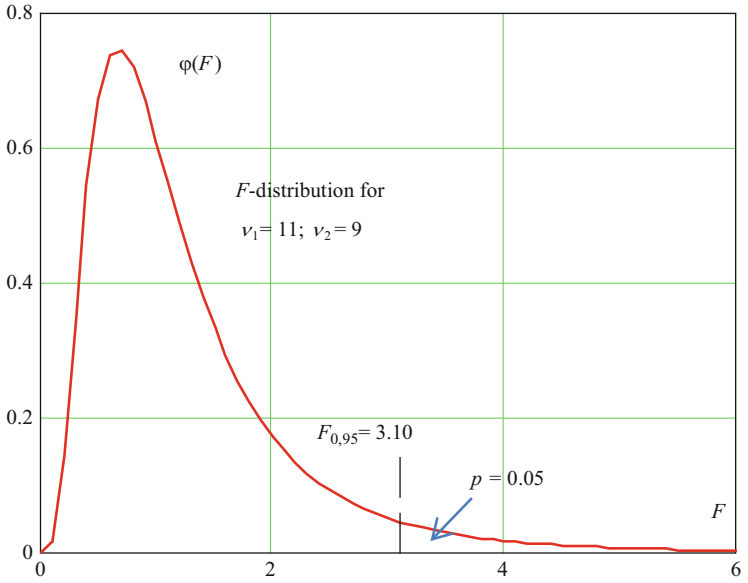


Fig. 8.3 Probability density function of F – distribution

$$\varepsilon_F = 3\left[(\nu_2 - 4)\frac{(\nu_1+\nu_2-2)(\nu_1\nu_2+6\nu_1+6\nu_2-8)+4\nu_1}{\nu_1(\nu_1+\nu_2-2)(\nu_2-8)(\nu_2-6)} - 1\right] \quad \text{for} \quad \nu_2 > 8 \tag{8.10}$$

This distribution is used for testing statistical hypothesis concerning the difference between two sample standard deviations. The probability density distribution $\varphi(F)$ for $\nu_1 = 11$ and $\nu_2 = 9$ (used in Chap. 10) is shown in Fig. 8.3.

8.2 Point Estimate of the Mean

The sample mean m given by Eq. (3.1) has in general normal distribution $N(\mu,\sigma/\sqrt{n})$ having the mean and standard deviation

$$\mu_m = \mu, \quad \sigma_m = \sigma/\sqrt{n} \tag{8.11}$$

Here μ and σ denote the mean and standard deviation of the population. Equation (8.11) holds approximately for any population distribution (due to the theorem of central tendency). Thus the unbiased estimate $\hat{m}$ of the population mean μ is simply equal to the sample mean m, thus

$$\hat{m} = m \tag{8.12}$$

This trivial result for the point estimate of the population mean is independent of the sample size n. However, it follows from Eq. (8.11) that the standard deviation σ_m of the sample mean m may be relatively large, particularly for small samples and a great variance of the population. For example, if the sample size is only $n = 4$, then the standard deviation $\sigma_m = \sigma/2$ and the point estimate (8.12) may suffer from a significant statistical uncertainty.

8.3 Point Estimate of the Variance

The sample variance s^2 given by Eq. (3.2) can be described by χ^2 distribution with $\nu = n-1$ degree of freedom. The random variable χ^2 follows from Eqs. (8.1) and (3.2) (substituting $\sum_1^n (x_i - m)^2 = ns^2$) as

$$\chi^2 = \frac{\sum_1^n (x_i - m)^2}{\sigma^2} = n\frac{s^2}{\sigma^2} \tag{8.13}$$

Here, the degree of freedom arises from the residual sum-of-squares in the numerator, and in turn the $n - 1$ degree of freedom of the underlying residual vector $\{x_i - m\}$. The sum of the residuals is necessarily 0. If $n-1$ values are known, then the last one can be thus found. That means the residuals are constrained to lie in a space of dimension $n-1$. One says that "there are $n-1$ degree of freedom for the residual."

The unbiased estimate $\hat{s}^2$ of the population variance σ^2 corresponds to the mean $\mu_{\chi 2}$ of the random variable χ^2 defined by Eq. (8.2), that is

$$\mu_{\chi^2} = n\frac{s^2}{\hat{s}^2} = n - 1 \tag{8.14}$$

The best estimate $\hat{s}$ of the population standard deviation σ corresponding to the mean μ_{χ^2} of the χ^2 distribution of the sample variance s^2 follows from Eq. (8.14) as

$$\hat{s} = s\sqrt{\frac{n}{n-1}} = \sqrt{\frac{\sum_1^n (x_i - m)^2}{n-1}} \tag{8.15}$$

So, the denominator $n-1$ in the estimate (Eq. (8.15)) is due to the fact that the best estimate is derived from the mean of the χ^2-distribution describing the sample variance s^2.

Example 8.1. Consider the sample of measurements of a steel tensile strength having the standard deviation $s = 20.51$ MPa that is determined from a sample of the size $n = 100$. Obviously the best estimate $\hat{s}^2$ of the population variance σ^2 follows from Eq. (8.15)

$$\hat{s} = s\sqrt{\frac{n}{n-1}} = 20.51\sqrt{\frac{100}{99}} = 20.61 \text{ MPa}$$

However, if the sample size would be limited to $n = 10$ and the sample standard deviation would be the same $s = 20.51$ MPa (as before for $n = 100$), then the point estimate of the population standard deviation would be greater

$$\hat{s} = s\sqrt{\frac{n}{n-1}} = 20.51\sqrt{\frac{10}{9}} = 21.62 \text{ MPa}$$

8.4 Interval Estimate of the Mean

In general the interval estimates provide better information about the possible range of the population parameters. The interval estimates that cover the population parameters always correspond to some confidence level (probability close to 1) that the parameter will be covered by the interval. Commonly, the confidence level $1-2p = 0.90$ or 0.95 is accepted, where the probability $p = 0.05$ or 0.025 denotes a one sided probability that the estimation limits will be exceeded.

The interval estimates of the population mean depends on whether the population standard deviation σ is known or unknown. If the population standard deviation is unknown then instead of σ the sample standard deviation s and appropriate sampling distribution is to be considered.

8.4.1 Known σ

The interval that covers the population mean with the confidence level 1–2p follows from Eqs. (8.11) and (8.12) as

$$m + u_p \frac{\sigma}{\sqrt{n}} < \mu < m + u_{1-p} \frac{\sigma}{\sqrt{n}} \tag{8.16}$$

Here u_p and u_{1-p} denotes the fractiles of standardised normal variable corresponding to the probabilities p and $1-p$.

Example 8.2. The sample mean of the yield point determined from a sample of 100 observations is 250.5 MPa. The standard deviation of the population is known from previous experience as $\sigma = 20.51$ MPa. The interval estimate corresponding to the confidence level 0.95 ($p = 0.025$) is then

$$250.5 - 1.96\frac{20.51}{\sqrt{100}} < \mu < 250.5 + 1.96\frac{20.51}{\sqrt{100}}$$

$$246.5\,\text{MPa} < \mu < 254.5\ \text{MPa}$$

Here the values $-u_p = u_{1-p} = 1.96$ can be obtained from any commonly available software products (EXCEL, STATISTICA MATHCAD, MATLAB).

8.4.2 Unknown σ

If the population standard deviation σ is unknown then the sample standard deviation s has to be used. The interval estimate corresponding to the confidence level 1–2p is given as

$$m + t_p\frac{s}{\sqrt{n-1}} < \mu < m + t_{1-p}\frac{s}{\sqrt{n-1}} \tag{8.17}$$

Here t_p and t_{1-p} denotes the fractiles of t-distribution corresponding to the the $n-1$ degree of freedom and probabilities p and $1-p$.

Example 8.3. Consider a similar case as Example 8.2. The sample mean of the yield point and its standard deviation are now determined from a sample of 10 observations only. Assume the same numerical values yielding the sample mean $m = 250.5$ MPa, the sample standard deviation $s = 20.51$ MPa. The interval estimate of the population mean corresponding to the confidence level $1 - 2p = 0.95$ ($p = 0.025$) and to the degree of freedom $\nu = 10 - 1 = 9$ ($-t_{0.025} = t_{0.975} = 2.262$ follows from Eq. (8.17) as

$$250.5 - 2.262\ \frac{20.51}{\sqrt{9}} < \mu < 250.5 + 2.262\ \frac{20.51}{\sqrt{9}}$$

$$235.0\,\text{MPa} < \mu < 266.0\,\text{MPa}$$

Comparing this interval estimate with the previous one in Example 8.2, it is clear that due to the unknown standard deviation σ the interval is slightly broader.

8.5 Interval Estimate of the Variance

The interval estimate of unknown population variance and standard deviation is derived from χ^2-distribution introduced in Sect. 8.1. The confidence level is now expressed as $1–p_1–p_2$, where p_1 and $1–p_2$ denote the probabilities corresponding to the lower and upper fractiles $\chi^2_{p_1}$ and $\chi^2_{1-p_2}$ specified for $\nu = n–1$ degree of freedom. It follows from Eq. (8.13) that the interval covering the population standard deviation is

$$\sqrt{\frac{n}{\chi^2_{1-p_2}}}s < \sigma < \sqrt{\frac{n}{\chi^2_{p_1}}}s \tag{8.18}$$

As a rule the confidence level $1–p_1–p_2$ is 0,90 or 0,95 usually assuming the probabilities $p_1 = p_2 = 0.05$ or 0.025.

Example 8.4. Consider the sample standard deviation from previous example $s = 20.51$ MPa determined from a sample of 10 observations. Considering the confidence level $1–p_1–p_2 = 0.90$ and $p_1 = p_2 = 0.05$, $\nu = n–1 = 9$, then $\chi^2_{0.05} = 3.325$ and $\chi^2_{0.95} = 16.92$ (see Fig. 8.1). It follows from Eq. (8.18)

$$\sqrt{\frac{10}{16.92}}20.51 < \sigma < \sqrt{\frac{10}{3.325}}20.51$$

$$15.77 < \sigma < 35.57$$

This example clearly indicates that the interval covering the population standard deviation σ with the probability 0.90 may be significantly broad and that the point $\hat{s} = 21.62$ MPa obtained for sample size $n = 10$ in Example 8.1 suffer from a great statistical uncertainty.

8.6 Specification of the Sample Size

It is expected that random samples obtained by experimental investigation will represent well the populations that are being studied. In particular it is often required that the estimated population mean is satisfactorily accurate and has a limited error. It clearly follows from the Sect. 8.1 that the relative error in the estimated population mean μ depends on the sample size n. Determination of an appropriate sample size n is therefore an important step in any experimental investigation. The adequate statistical procedure depends on whether the population standard deviation σ is known or unknown.

8.6.1 Known σ

If the population coefficient of variation σ is known (from previous experience or from similar populations) then the maximum deviation of the population mean estimate follows from Eq. (8.16) as

$$\left| u_p \frac{\sigma}{\sqrt{n}} \right| \tag{8.19}$$

Then the relative error δ can be obtained by dividing the above absolute error by the population mean μ. Then

$$\delta = \left| u_p \frac{\sigma}{\mu} \frac{1}{\sqrt{n}} \right| = \left| u_p \frac{V}{\sqrt{n}} \right| \tag{8.20}$$

Here $V = \sigma/\mu$ denotes the population coefficient of variation. The required sample size can then be expressed in terms of the coefficient V as

$$n > \left(\frac{u_p V}{\delta} \right)^2 \tag{8.21}$$

Here u_p denotes the normalised normal random variable corresponding to the probability p; Thus, the probability (confidence level) that the relative error δ will not be exceeded is $1-2p$ (commonly equal to 0.90 or 0.95).

Example 8.5. It is known that the concrete strength has a coefficient of variation $V = 0.13$. If the relative error in estimating the population mean μ should be at the most 10 % (that corresponds to $\delta < 0.10$), then it follows from Eq. (8.21) that the number of specimens to be investigated should be

$$n > \left(\frac{1.96 \times 0.13}{0.1} \right)^2 = 6.5$$

So, at least 7 specimens should be used.

8.6.2 Unknown σ

If the population standard deviation σ is unknown, then the sample standard deviation s and t-distribution are to be considered. Then Eqs. (8.19, 8.20 and 8.21) become

$$\left| t_p \frac{s}{\sqrt{n-1}} \right| \tag{8.22}$$

$$\delta = \left| t_p \frac{s}{m} \frac{1}{\sqrt{n-1}} \right| = \left| t_p \frac{v}{\sqrt{n-1}} \right| \tag{8.23}$$

$$n > \left(\frac{t_p v}{\delta} \right)^2 + 1 \tag{8.24}$$

Here t_p denotes the p-fractile of the t- distribution corresponding to the probability p and to the sample coefficient of variation $v = s/m$.

Example 8.6. Assume that the sample coefficient of variation $v = 0.13$ is determined from a small sample of the size $n = 7$ (degree of freedom $\nu = 6$). If the relative error in estimating the population mean μ should be at the most 10 % (that corresponds to $\delta < 0.10$), then it follows from Eq. (8.24) that the number of specimens to be investigated should be

$$n > \left(\frac{2.447 \times 0.13}{0.1} \right)^2 + 1 = 11.1$$

Here $t_p = 2.447$ is determined for $\nu = 6$ and $p = 0.025$. So, at least 12 specimens should be tested, more than in the case of known σ when only 7 specimens are required.

8.7 Estimation of the Skewness

The sample skewness a (the coefficient of asymmetry without subscrip) is a very sensitive sample characteristic substantially affected by deviations and possible errors of observed data, particularly in the case of small samples. It is strongly recommended that the sample skewness (as well as kurtosis) be evaluated with the utmost caution. In particular dubious data and outliers should be carefully verified, tested and if need be deleted from further investigation.

The corresponding sampling distribution is complicated and its description is beyond the scope of this introductory text. As already indicated in Chap. 3 by Eqs. (3.26) and (3.27) there are alternative expressions commonly used for an unbiased point estimate of the population skewness. STATISTICA software products provide another expression, assuming that the population standard deviation σ is known

$$\hat{a} = \frac{n}{(n-1)(n-2)\sigma^3} \sum_1^n (x_i - m)^3 = \frac{n^2}{(n-1)(n-2)} \frac{m_3}{\sigma^3} \tag{8.25}$$

Here m_3 denotes the third central moment given by Eq. (3.8). If the population standard deviation σ is unknown then it should be substituted by the estimate $\hat{s}$ given by Eq. (8.15). The resulting expression for the point estimate of the population skewness $\hat{a}$ then becomes

$$\hat{a} = \frac{n}{(n-1)(n-2)\hat{s}^3} \sum_1^n (x_i - m)^3 = \frac{\sqrt{n(n-1)}}{(n-2)s^3} \frac{1}{n} \sum_1^n (x_i - m)^3$$
$$= \frac{\sqrt{n(n-1)}}{(n-2)} a \tag{8.26}$$

Here the sample skewness a is given by Eq. (3.5). The expression (Eq. (8.26)) gives the same result as Eq. (3.26), and is commonly used by software products (EXCEL, MATLAB and MATHCAD and other statistical packages).

The enhanced factor of the sample skewness a in Eq. (8.26) is greater than 1 and with increasing sample size n the factor decreases and ultimately approaches 1. It should be noted that the population standard deviation σ may surprisingly increase the estimate $\hat{a}$ given by Eq. (8.25) compared with the expression Eq. (8.26) (the population standard deviation σ may be less than the population estimate $\hat{s}$). However, the differences between expressions Eqs. (8.25) and (8.26) are significant only for small samples (sample size $n < 30$) and diminish with increasing n; if the sample size $n > 30$ then the enhanced factor according to expression Eq. (8.25) is less than 1,10, the enhanced factor according Eq. (8.26) less than 1,05 (see also discussion in Sect. 3.7).

Uncertainty in evaluating skewness is usually characterized by the variance $\sigma_{\hat{a}}^2$ (the standard deviation $\sigma_{\hat{a}}$) of the point estimate $\hat{a}$ made from limited sample data taken from the normal population

$$\sigma_{\hat{a}}^2 = \frac{6n(n-1)}{(n-2)(n+1)(n+3)} \tag{8.27}$$

An approximate value of the variance Eq. (8.27) is $6/n$ but this is inaccurate for small samples of the size $n < 30$. For the sample size $n > 30$ (recommended for estimating population skewness) the standard deviation $\sigma_{\hat{a}}$ of the point estimate $\hat{a}$ can be approximated as

$$\sigma_{\hat{a}} \cong \sqrt{6/n}. \tag{8.28}$$

Approximately, it can be stated that if the point estimate is, in absolute value, greater than a double of this value, then the skewness is considered to be significant and the population is assumed to be asymmetric (not normal). The above procedure for estimating population skewness $\hat{a}$ is illustrated by the following numerical example.

Example 8.7. Consider the sample skewness $a = 1{,}00$ determined from a sample of 30 observations.

The estimated population skewness $\hat{a}$, in accordance with Eq. (8.25) used in STATISTICA, assuming the population standard deviation σ is known, and the ratio $\frac{m_3}{\sigma^3}$ is equal to 1, is given as

$$\hat{a} = \frac{30^2}{29 \times 28} 1.00 = 1.11$$

The formula Eq. (8.26) yields assuming the sample skewness $a \approx \frac{m_3}{\sigma^3} = 1$

$$\hat{a} = \frac{\sqrt{30 \times 29}}{28} 1.00 = 1.05$$

Obviously, the enhanced factor obtained from formula Eq. (8.25) (involving the population standard deviation σ) is slightly greater than that obtained from Eq. (8.26) (in which the estimate of population standard deviation $\hat{s}$ is considered).

If the population is normal, then the standard deviation of the estimate can be assessed using Eq. (8.28)

$$\sigma_{\hat{a}} = \sqrt{6/n} = 0.45$$

The skewness is to be considered as significant because

$$a = 1.00 > 2 \times 0.45$$

Thus, the population from which the sample is taken cannot be considered as symmetric (normally distributed).

References

1. Ang, A.H.-S., Tang, W.H.: Probabilistic Concepts in Engineering. Emphasis on Applications to Civil Environmental Engineering. Wiley, New York (2007)
2. Devore, J., Farnum, N.: Applied Statistics for Engineers and Scientists. Thomson, London (2005)
3. Dunin-Barkovskij, I.V., Smirnov, N.V.: The Theory of Probability and Mathematical Statistics in Engineering. Technical and Theoretical Literature, Moscow (in Russian) (1955)
4. Gurskij, E.I.: The Theory Probability with Elements of Mathematical Statistics. Higher School, Moscow (in Russian) (1971)
5. ISO 3534–1: Statistics – Vocabulary and Symbols – Part 1: Probability and General Statistical Terms. ISO, Geneve (1993)
6. ISO 3534–2: Statistics – Vocabulary and Symbols – Part 2: Statistical Quality Control. ISO, Geneve (1993)

Chapter 9
Fractiles of Random Variables

A fractile is the value of a random variable corresponding to a given probability of occurrence of values smaller than the fractile. It is an important concept used in many engineering and scientific applications. If a random variable is defined by a known theoretical model then the fractile is simply the point at which the distribution function attains the specified probability. However, estimation of fractiles from limited sample data without having a theoretical model of the random variable is a more complicated task. Two different methods are commonly used: the classical coverage method and the prediction method. Operational techniques are provided for both methods and their comparison, taking into account the confidence level of the coverage method offered. In addition, the Bayessian approach to fractile estimation is explained, by way of updating prior data with newly obtained information. A review of fundamental procedures provides Annex 5.

9.1 Fractiles of Theoretical Models

One of the most important keywords in the theory of structural reliability is the term "fractile" of a random variable X (or of its probability distribution). In some publications and software products the term "quantile" [1, 2] is used, but more frequently the term fractile [3–5] is accepted (used also in this book). For a given probability p, the p-fractile x_p denotes such a value of the random variable X, for which it holds that values of the variable X smaller than or equal to x_p occur with the probability p. If $\Phi(x)$ is the distribution function of the random variable X, then it follows from Eq. (4.1) that the value $\Phi(x_p)$ is equal to the probability p, thus the fractile x_p can be defined as

$$\mathrm{P}(X \leq x_p) = \Phi(x_p) = p \tag{9.1}$$

M. Holický, *Introduction to Probability and Statistics for Engineers*,
DOI 10.1007/978-3-642-38300-7_9, © Springer-Verlag Berlin Heidelberg 2013

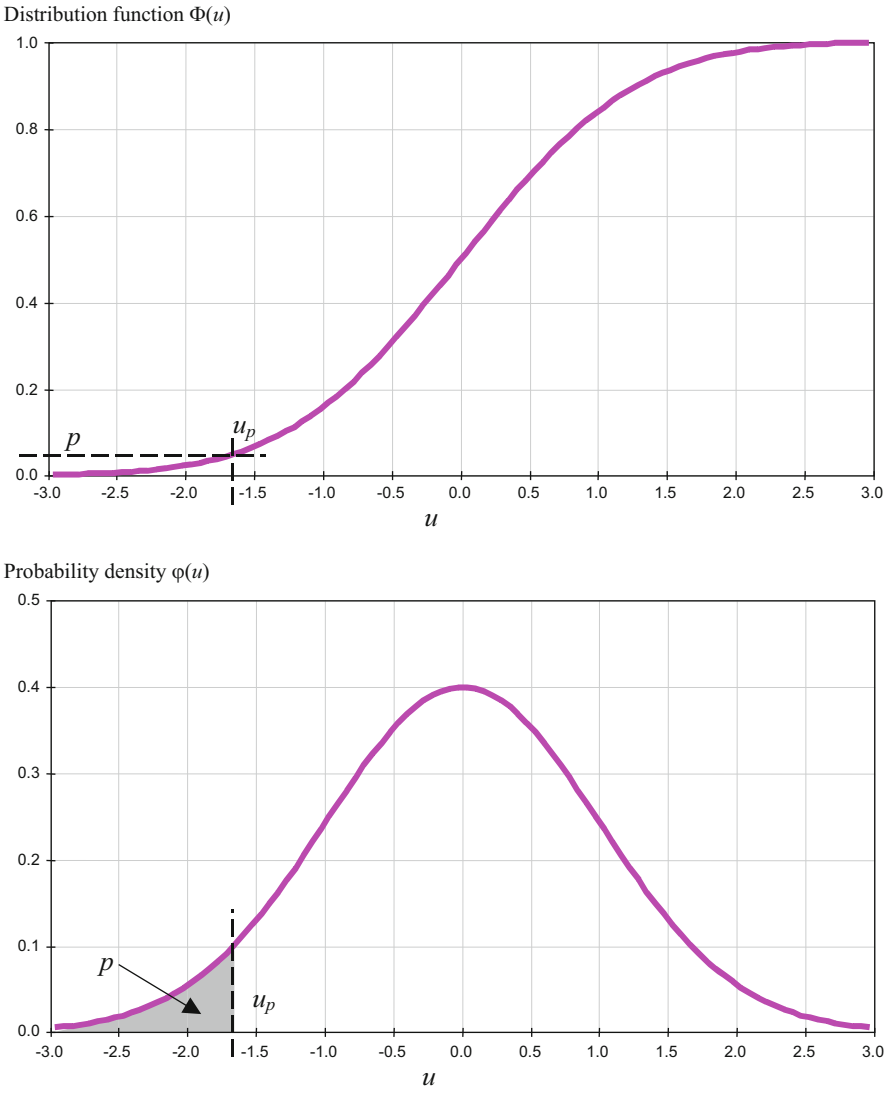


Fig. 9.1 Definition of the fractile for a standardised random variable U

The same definition holds for a standardised random variable U (given by the transformation Eq. (4.23)), when in Eq. (9.1) U is substituted for X and u_p is substituted for x_p. Figure 9.1 illustrates the definition given in Eq. (9.1).

Fractiles u_p of standardised random variables U are commonly available in tables. Figure 9.1 illustrates the definition of the fractile described by Eq. (9.1) for a standardised random variable U; it shows the distribution function $\Phi(u)$, the probability density function $\varphi(u)$, the probability p (equal to 0.05) and the fractile u_p

(equal to -1.645) for the distribution of a standardised variable U having the normal distribution.

In general, the fractile x_p of an original random variable X may be calculated using tables for u_p available for standardised random variables U with a relevant type of distribution. It follows from the transformation Eq. (4.23) that the fractile x_p may be determined from the standardised random variable u_p (found in available tables) using the relationship

$$x_{\mathrm{p}} = \mu + u_p \sigma = \mu(1 + u_p V) \tag{9.2}$$

where μ denotes the mean, σ the standard deviation and V the coefficient of variability of the observed variable X.

If the probability $p < 0.5$, then the value x_p is often called the lower fractile, for $p > 0.5$ the x_p is called the upper fractile. Figure 9.2 shows the lower and upper fractiles u_p of a standardised random variable U with a normal distribution for probabilities $p = 0.05$ and 0.95, and thus denoted $u_{0.05}$ and $u_{0.95}$.

The values u_{p} of the lower fractile of a standardised random variable U having a normal distribution for selected probabilities p are given in Table 9.1. Considering the symmetry of the normal distribution, the values u_{p} of the upper fractile can be assessed from Table 9.1 by the substitution of p with $1-p$ and by changing the sign of values u_{p} (from negative to positive). Detailed tables can be found, for example in textbooks [1, 2], in the standard ISO 12491 [5], and in specialised literature.

For a standardised random variable having a general three-parameter log-normal distribution the value u_{p} of the standardised random variable is dependent on the skewness α. The values u_{p} for selected skewnesses α and probabilities p are given in Table 9.2.

The fractile corresponding to the probability $p = 0.05$ is usually applied for an assessment of the characteristic value of material properties (strength of concrete, yield strength of steel, masonry strength). However, the design values of dominant variables are fractiles which correspond to a lower probability ($p \cong 0.001$), the design values of non-dominant variables are fractiles corresponding to a greater probability ($p \cong 0.10$).

In the case of a log-normal distribution with the lower bound at zero, which is described in Sect. 6.2, it is possible to calculate the fractile from the value of the fractile $u_{\mathrm{norm},p}$ of a standardised random variable having the normal distribution using the relation

$$x_p = \frac{\mu}{\sqrt{1+V^2}} \exp\left(u_{norm,p}\sqrt{\ln(1+V^2)}\right) \tag{9.3}$$

where $u_{\mathrm{norm,p}}$ is the fractile of a standardised random variable with a normal distribution, μ is the mean and V the coefficient of variation of the variable X. An approximation of Eq. (9.3) is often applied in the form

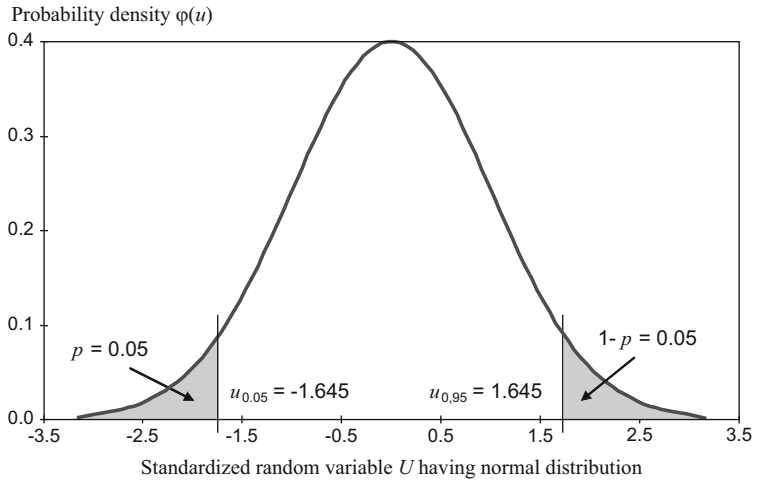


Fig. 9.2 The lower and upper fractiles of a standardised random variable U having a normal distribution

Table 9.1 Fractile u_p of a standardised random variable having the normal distribution

p	10^{-7}	10^{-6}	10^{-5}	10^{-4}	0.001	0.010	0.050	0.100	0.200	0.500
$-u_p$	5.199	4.753	4.265	3.719	3.091	2.327	1.645	1.282	0.841	0.000

$$x_p \cong \mu \exp(u_{\text{norm},p} \times V) \tag{9.4}$$

whose accuracy is fully satisfying for $V < 0.2$, but is commonly used for greater V as well.

Example 9.1. Let us assess the fractile x_p of a normal and two parameter log-normal distribution (with the lower limit at zero) for probabilities $p = 0.001$; 0.01; 0.05 and 0.10, assuming the coefficient of variation $V = 0.3$. We know that the log-normal distribution with the lower limit at zero has, in this case, a positive skewness $\alpha = 0.927$ (according to Eq. (6.11)), which needs to be known for interpolation in Table 9.2. The resulting values x_p are given in the following table in the form of dimensionless ratios x_p/μ (the ratio of the fractile to the mean), which were assessed in different ways for the normal and for the log-normal distribution.

Table of the fractions x_p/μ.

	Probability p			
Fraction x_p/μ for	0.001	0.010	0.050	0.100
Normal distribution, Equation (9.2), Table 9.1	0.073	0.302	0.506	0.615
Log-normal distribution, Equation (9.2), Table 9.2	0.385	0.483	0.591	0.658
Log-normal distribution, Equation (9.3), Table 9.1	0.387	0.484	0.591	0.657
Log-normal distribution, Equation (9.4), Table 9.1	0.396	0.496	0.610	0.681

Table 9.2 Fractile u_p of a standardised random variable having three parameter log-normal distribution

	Probability p												
α	10^{-4}	10^{-3}	0.01	0.05	0.10	0.20	0.50	0.80	0.90	0.95	0.99	$1-10^{-3}$	$1-10^{-4}$
−2.0	−9.52	−6.24	−3.52	−1.89	−1.24	−0.61	0.24	0.77	0.97	1.10	1.28	1.42	1.49
−1.5	−7.97	−5.51	−3.31	−1.89	−1.29	−0.68	0.20	0.81	1.04	1.21	1.45	1.65	1.77
−1.0	−6.40	−4.70	−3.03	−1.85	−1.32	−0.74	0.15	0.84	1.13	1.34	1.68	1.99	2.19
−0.5	−4.94	−3.86	−2.70	−1.77	−1.32	−0.80	0.08	0.85	1.21	1.49	1.98	2.46	2.81
0.0	−3.72	−3.09	−2.33	−1.65	−1.28	−0.84	0.00	0.84	1.28	1.65	2.33	3.09	3.72
0.5	−2.81	−2.46	−1.98	−1.49	−1.21	−0.85	−0.08	0.80	1.32	1.77	2.70	3.86	4.94
1.0	−2.19	−1.99	−1.68	−1.34	−1.13	−0.84	−0.15	0.74	1.32	1.85	3.03	4.70	6.40
1.5	−1.77	−1.65	−1.45	−1.21	−1.04	−0.81	−0.20	0.68	1.29	1.89	3.31	5.51	7.97
2.0	−1.49	−1.42	−1.28	−1.10	−0.97	−0.77	−0.24	0.61	1.24	1.89	3.52	6.24	9.52

The above table of ratios x_p/μ shows the expected difference between the fractiles of normal and log-normal distributions. The lower fractile of the normal distribution is significantly lower than the corresponding fractile of the log-normal distribution, particularly for small probabilities p. The table also shows that the approximate formula Eq. (9.4) provides satisfactory results for computation of the fractile of the log-normal distribution (the error will decrease by decreasing the coefficient of variation V).

The fractile of the gamma distribution can be calculated from the available tables for type III Pearson distribution [5, 6]. To calculate the fractile of the beta distribution, the available tables of an incomplete beta function may be used, or it can be assessed by an integration of the probability density function according to definition Eq. (9.1). However, when it is needed (and neither appropriate tables nor software products are available), the fractile of the beta distribution, which is bell shaped (for shape parameters $c > 2$ and $d > 2$), may be assessed approximately from Eq. (9.2) using the table values of u_p for a standardised log-normal distribution, having the same skewness α as the beta distribution. An analogical procedure may be also used for other types of distribution.

The fractile x_p can be easily assessed for the Gumbel distribution. From Eq. (6.23) and definition Eq. (9.1) follows an explicit relation for x_p directly dependent on the probability p

$$x_p = x_{\mathrm{mod}} - \frac{1}{c}\ln(-\ln(p)) \cong \mu - (0.45 + 0.78\ln(-\ln(p)))\sigma \tag{9.5}$$

where the mode x_{mod} and parameter c are substituted by relations Eqs. (6.25) and (6.26).

Example 9.2. Let us determine the upper fractile of the wind pressure from Example 6.6 described by a Gumbel distribution when a probability $p = 0.98$ is considered. It is known from Example 6.6 that for the 1-year maximum $\mu_1 = 0.35$ kN/m^2, $\sigma_1 = 0.06$ kN/m^2. The fractile $x_{0.98}$ for such parameters follows from Eq. (9.5)

$$x_{0.98} = 0.35 - (0.45 + 0.78 \times \ln(-\ln(0.98))) \times 0.06 = 0.51 \text{ kN/m}^2$$

The corresponding fractile of the maximum for a period of 50 years (as shown in Example 6.6 where $\mu_{50} = 0.53$ kN/m^2, $\sigma_{50} = 0.06$ kN/m^2) is

$$x_{0.98} = 0.53 - (0.45 + 0.78 \times \ln(-\ln(0.98))) \times 0.06 = 0.69 \text{ kN/m}^2$$

Simple mathematical operations with the Gumbel distribution, including the computation of fractiles, are the main reasons why this distribution is so popular. The Gumbel distribution is frequently used as a theoretical model of random variables describing climatic and other variable actions that are defined on the

basis of the maximal values in a given period of time (for example in one or several years).

9.2 Fractile Estimation from Samples: Coverage Method

Theoretical models are very rarely known precisely in practical applications. In civil engineering, it is often necessary to assess the fractile of a random variable (for example, the strength of a new or unknown material) from a limited sample, the size n of which may be very small ($n < 10$). Furthermore, random variables may have a high variability (the coefficient of variation is sometimes greater than 0.30). The assessment of the fractile of a population from a very small sample is then a serious problem, which is solved in mathematical statistics by various methods of estimation theory. In the following, three basic methods are briefly described: the coverage method, the prediction method and the Bayesian method for the estimate of the population fractile.

The keyword of the coverage method for fractile estimation from a sample of limited size n is confidence γ, i.e. the probability (usually 0.75, 0.90 or 0.95) that the estimated value covers the population fractile (that is why the method is called the coverage method). The estimator $x_{p,\text{cover}}$ of the lower fractile x_p is determined by the coverage method in such a way that

$$\mathrm{P}(x_{p,\text{cover}} < x_p) = \gamma \tag{9.6}$$

Thus, the estimator $x_{p,\text{cover}}$ is lower (on the safe side) than the unknown fractile x_p with the probability (confidence) γ.

In the following summary practical formulas are given without being derived, assuming that the population has a general three-parameter distribution characterised by the skewness α, which is assumed to be known from previous experience. Besides that, it is assumed that the mean μ of the population is never known in advance and that the estimate is based on the average m obtained from a sample. The standard deviation s of the population is assumed to be either known, in which case it is used, or unknown, in which case the sample standard deviation s is used instead.

If the standard deviation σ of the population is known from previous experience, the estimator $x_{p,\text{cover}}$ of the lower p-fractile is given by the relation

$$x_{p,\text{cover}} = m - \kappa_p \sigma \tag{9.7}$$

If the standard deviation of the population σ is unknown, then the sample standard deviation s is considered

$$x_{p,\text{cover}} = m - k_p s \tag{9.8}$$

The coefficients of estimation $\kappa_p = \kappa(\alpha, p, \gamma, n)$ and $k_p = k(\alpha, p, \gamma, n)$ depend on the skewness α, on the probability p corresponding to the fractile x_p, which is estimated, on the confidence γ and on the size n of the population. The knowledge of the confidence γ that the estimate $x_{p,\text{cover}}$ will be on the safe side of the real value is the greatest advantage of the classic coverage method. In documents [7, 8] the confidence γ is recommended by the value 0.75. In cases of the demands of increased reliability, when a detailed reliability analysis is required, a higher value of confidence, say 0.95 may be more appropriate [4].

9.3 Fractile Estimation from Samples: Prediction Method

According to the prediction method [5] the lower p-fractile x_p is estimated by the so-called prediction limit $x_{p,\text{pred}}$, for which it holds that a new value x_{n+1} randomly drawn from the population will be lower than the estimate $x_{p,\text{pred}}$ with only the probability p, i.e. it holds that

$$\mathrm{P}(x_{n+1} < x_{p,\text{pred}}) = p \tag{9.9}$$

It can be shown that for a growing n the estimator $x_{p,\text{pred}}$ defined in this way is asymptomatically approaching the unknown fractile x_p. It can also be shown that the estimator $x_{p,\text{pred}}$ corresponds approximately to the estimator obtained by the coverage method $x_{p,\text{cover}}$ for a confidence $\gamma = 0.75$.

If the standard deviation σ of the population is known, then the lower p-fractile is estimated by the value $x_{p,\text{pred}}$ according to the relation

$$x_{p,\text{pred}} = m + u_p(1/n + 1)^{1/2}\sigma \tag{9.10}$$

where $u_p = u(\alpha, p)$ is the p-fractile of a standardised log-normal distribution, having the skewness α.

If, however, the standard deviation of the population is unknown, then the sample standard deviation s must be considered instead of σ

$$x_{p,\text{pred}} = m + t_p(1/n + 1)^{1/2}s \tag{9.11}$$

where $t_p = t(\alpha, p, \nu)$ is the p-fractile of the generalised Student's t-distribution for $\nu = n-1$ degrees of freedom, which has a skewness α (information about the Student's distribution and about the number of degrees of freedom may be obtained from Sect. 8 and from other specialised sources [1, 2])

Table 9.3 Coefficients κ_p and $-u_p(1/n + 1)^{1/2}$ from Eqs. (9.7) and (9.10) for $p = 0.05$ and a normal distribution of the population (when σ is known)

Coefficients		Sample size n								
		3	4	5	6	8	10	20	30	∞
κ_p	$\gamma = 0.75$	2.03	1.98	1.95	1.92	1.88	1.86	1.79	1.77	1.64
	$\gamma = 0.90$	2.39	2.29	2.22	2.17	2.10	2.05	1.93	1.88	1.64
	$\gamma = 0.95$	2.60	2.47	2.38	2.32	2.23	2.17	2.01	1.95	1.64
$-u_p(1/n + 1)^{1/2}$		1.89	1.83	1.80	1.77	1.74	1.72	1.68	1.67	1.64

9.4 Comparison of the Coverage and Prediction Methods

The coverage and predictive methods represent two basic procedures of estimation of the population's fractile from an available sample of a limited size n. If the standard deviation of the population σ is known, then Eqs. (9.7) and (9.10) are applied, in which two analogical coefficients κ_p and $-u_p(1/n + 1)^{1/2}$ appear. Both of these coefficients depend on the sample size n; the coefficient κ_p of the coverage method depends more on the confidence γ. Table 9.3 shows the coefficients κ_p and $-u_p(1/n + 1)^{1/2}$ for $p = 0.05$ and selected values of n and γ when a normal distribution of the population is assumed.

It follows from Table 9.3 that both the coefficients approach with $n \to \infty$ the value 1.64, valid for a theoretical model of normal distribution (see Table 9.1). The coefficient κ_p of the coverage method increases with increasing confidence γ. Note that for a confidence $\gamma = 0.75$ it holds that $\kappa_p \cong -u_p(1/n + 1)^{1/2}$. Thus, for $\gamma = 0.75$ the coverage method leads to approximately the same estimate as the prediction method, $x_{p,\text{cover}} \cong x_{p,\text{pred}}$ (for $\gamma > 0.75$ the $x_{p,\text{cover}} < x_{p,\text{pred}}$).

If the standard deviation of the population σ is unknown, Eqs. (9.8) and (9.11) are applied, in which two analogical coefficients k_p and $-t_p(1/n + 1)^{1/2}$ appear. Both of these coefficients depend again on the sample size n, but the coefficient k_p of the coverage method depends more on the confidence γ. Table 9.4 and Fig. 9.3 show the values of coefficients k_p and $-t_p(1/n + 1)^{1/2}$ for $p = 0.05$ and selected values of n and γ when a normal distribution of the population is assumed.

It is obvious from Table 9.4 and Fig. 9.3 that with increasing the sample size n both the coefficients k_p and $-t_p(1/n + 1)^{1/2}$ approach the value 1.64, which is valid for a theoretical model of normal distribution (see Table 9.1). In the case of the coverage method, the coefficient k_p increases with increasing confidence γ and the relevant estimates $x_{p,\text{cover}}$ of the lower fractile decrease (on the safe side). Note that as in the case of the known standard deviation σ both coefficients are approximately equal, $k_p \cong -t_p(1/n + 1)^{1/2}$ and for the confidence $\gamma = 0.75$ the coverage method leads to approximately the same estimate, $x_{p,\text{cover}} \cong x_{p,\text{pred}}$, as the prediction method.

Also the skewness (asymmetry) of the population α may significantly affect the estimate of the population's fractile. Tables 9.5 and 9.6 show the coefficients k_p from Eq. (9.8) for three values of the skewness $\alpha = -1.0$, 0.0 and 1.0, a probability

Table 9.4 Coefficients k_p and $-t_p(1/n + 1)^{1/2}$ from Eqs. (9.8) and (9.11) for $p = 0.05$ and a normal distribution of the population (when σ is unknown)

Coefficient		Sample size n								
		3	4	5	6	8	10	20	30	∞
k_p	$\gamma = 0.75$	3.15	2.68	2.46	2.34	2.19	2.10	1.93	1.87	1.64
	$\gamma = 0.90$	5.31	3.96	3.40	3.09	2.75	2.57	2.21	2.08	1.64
	$\gamma = 0.95$	7.66	5.14	4.20	3.71	3.19	2.91	2.40	2.22	1.64
$-t_p(1/n + 1)^{1/2}$		3.37	2.63	2.33	2.18	2.00	1.92	1.76	1.73	1.64

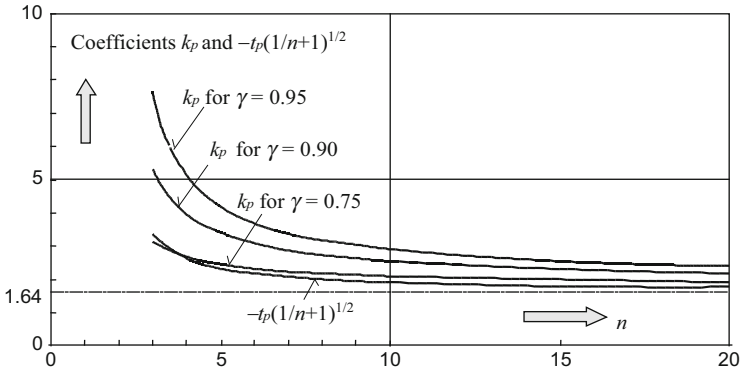


Fig. 9.3 Coefficients k_p and $-t_p(1/n + 1)^{1/2}$ for $p = 0.05$ and a normal distribution of the population (when σ is unknown)

Table 9.5 Coefficient k_p from Eq. (9.8) for $p = 0.05$, $\gamma = 0.75$ and a log-normal distribution having the skewness α (when σ is not known)

Skewness	Sample size n								
	3	4	5	6	8	10	20	30	∞
$\alpha = -1.00$	4.31	3.58	3.22	3.00	2.76	2.63	2.33	2.23	1.85
$\alpha = 0.00$	3.15	2.68	2.46	2.34	2.19	2.10	1.93	1.87	1.64
$\alpha = 1.00$	2.46	2.12	1.95	1.86	1.75	1.68	1.56	1.51	1.34

Table 9.6 Coefficient k_p from Eq. (9.8) for $p = 0.05$, $\gamma = 0.95$ and a log-normal distribution having the skewness α (when σ is not known)

Skewness	Sample size n								
	3	4	5	6	8	10	20	30	∞
$\alpha = -1.00$	10.9	7.00	5.83	5.03	4.32	3.73	3.05	2.79	1.85
$\alpha = 0.00$	7.66	5.14	4.20	3.71	3.19	2.91	2.40	2.22	1.64
$\alpha = 1.00$	5.88	3.91	3.18	2.82	2.44	2.25	1.88	1.77	1.34

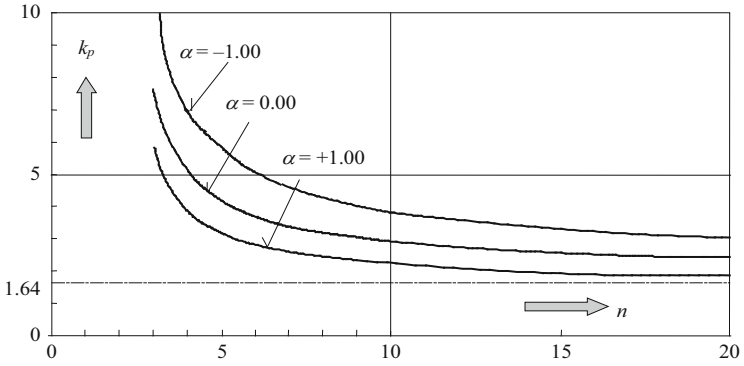


Fig. 9.4 Coefficient k_p for $p = 0.05$ and a confidence $\gamma = 0.95$ (when σ is unknown)

$p = 0.05$ and confidences $\gamma = 0.75$ (Table 9.5) and $\gamma = 0.95$ (Table 9.6). Values of the coefficients from Table 9.6 are shown in Fig. 9.4.

It is evident from Tables 9.5 and 9.6 that as the sample size n increases, the coefficients k_p approach the values of u_p, which are valid for a theoretical model of log-normal distribution (see Table 9.2). Thus, the influence of the skewness does not disappear when $n \to \infty$, and it is especially significant for small samples and a greater confidence $\gamma = 0.95$ (see Fig. 9.4).

A similar dependence on the skewness may be observed in the case of the generalised Student's t-distribution for which the fractiles t_p are given in Table 9.7. These values t_p are applied with the prediction method using formula (9.11) and further in the Bayes method. That is why Table 9.7 gives the values of fractiles t_p directly depending on the number of degrees of freedom ν. As in Tables 9.6 and 9.7, the probability $p = 0.05$ and three skewnesses $\alpha = -1.0$; 0.0 and 1.0 are considered.

It follows from Table 9.7 that as the size of the sample n increases, the values of t_p approach the theoretical values of u_p, which are valid for a model of the log-normal distribution with the appropriate skewness, and are given in Table 9.2. Therefore, the influence of the skewness again (as in the case of k_p) does not disappear for $n \to \infty$, but it is especially significant for small samples (it increases with a decreasing sample size n).

Example 9.3. A sample of size $n = 5$ measuring the strength of concrete has an average $m = 29.2$ MPa and a standard deviation $s = 4.6$ MPa. It can be assumed that the population is normal and that its standard deviation σ is unknown. The characteristic strength $f_{ck} = x_p$, where $p = 0.05$ is firstly assessed by the coverage method. If the confidence is $\gamma = 0.75$, then it follows from Eq. (9.8) and Table 9.4 that

$$x_{p,\text{cover}} = 29.2 - 2.46 \times 4.6 = 17.9 \text{ MPa}$$

Table 9.7 Coefficient $-t_p$ from Eq. (9.11) for $p = 0.05$ and a log-normal distribution with the skewness α (when σ is unknown)

	Coefficient $-t_p$ for $\nu = -1$ degrees of freedom								
Skewness	3	4	5	6	8	10	20	30	∞
$\alpha = -1.00$	2.65	2.40	2.27	2.19	2.19	2.04	1.94	1.91	1.85
$\alpha = 0.00$	2.35	2.13	2.02	1.94	1.86	1.81	1.72	1.70	1.64
$\alpha = 1.00$	1.92	1.74	1.64	1.59	1.52	1.48	1.41	1.38	1.34

If a higher confidence $\gamma = 0.95$ is required, then

$$x_{p,\text{cover}} = 29.2 - 4.20 \times 4.6 = 9.9 \text{ MPa}$$

If the predictive method is used, then it follows from Eq. (9.11) and Table 9.4 that

$$x_{p,\text{pred}} = 29.2 - 2.33 \times 4.6 = 18.5 \text{ MPa}$$

The characteristic strength obtained by the predictive method is only a little greater than the value according to the coverage method with the confidence $\gamma = 0.75$. However, if a higher confidence $\gamma = 0.95$ is required, then the predictive method leads to a value which is almost twice as great as the value obtained by the coverage method.

If the sample comes from a population with a log-normal distribution and a positive skewness $\alpha = 1$, then the coverage method with the confidence $\gamma = 0.75$ (Table 9.5) gives an estimator

$$x_{p,\text{cover}} = 29.2 - 1.95 \times 4.6 = 20.2 \text{ MPa}$$

which is a value that is 13 % greater than when the skewness is zero.

Similarly, it follows for the predictive method from Eq. (9.11) and Table 9.7 that

$$x_{p,\text{pred}} = 29.2 - 1.74 \times \sqrt{\frac{1}{5} + 1} \times 4.6 = 20.4 \text{ MPa}$$

where the value $t_p = -1.74$ is given, in Table 9.7, for $\alpha = 1.0$ and $\nu = 5-1 = 4$. The resulting strength is in this case is 10 % greater than the value which corresponds to the normal distribution ($\alpha = 0$).

9.5 Bayesian Estimation of Fractiles

If previous experience is available for a random variable (for example, in the case of a long-term production), it is possible to use the so-called Bayes method, which generally follows the idea of updating the probabilities described in Sect. 2.7. The Bayes method of fractile estimation is described here without deriving any important relations. A more detailed description is given in documents ISO [6, 8] and other specialised literature [1].

Let us assume that a sample of a size n with an average m and a standard deviation s is available. In addition, that an average m' and a sample standard deviation s' assessed from an unknown sample of an unknown size n' are known from previous experience. It is, however, assumed that both the samples come from the same population having a mean μ and standard deviation σ. The two samples may then be combined. This would be a simple task if the individual values of the previous set were known, but it is not the case. However, the Bayes method must still be used.

The characteristics of the combined sample are generally given by relations [6, 8]

$$n'' = n + n' \tag{9.12}$$

$$\nu'' = \nu + \nu' - 1 \text{ if } n' \geq 1, \nu'' = \nu + \nu' \text{ if } n' = 0$$

$$m'' = (mn + m'n')/n''$$

$$s''^2 = (\nu s^2 + \nu' s'^2 + nm^2 + n'm'^2 - n''m''^2)/\nu''$$

The unknown values n' and ν' may be assessed using the relations for the coefficients of variation of the mean and standard deviation $v(\mu)$ and $v(\sigma)$, (parameters μ and σ are considered as random variables in the Bayes concept) for which it holds [6, 8]

$$n' = [s'/(m'v(\mu))]^2, \nu' = 1/\left(2v(\sigma)^2\right) \tag{9.13}$$

Both the unknown variables n' and ν' are assessed independently (generally $\nu' \neq n'-1$), depending on previous experience concerning the degree of uncertainty of the estimate of the mean μ, and the standard deviation σ of the population.

The next step of the procedure applies the prediction method of fractile estimation. The Bayes estimator $x_{p,\text{Bayes}}$ of the fractile is given by a relationship similar to Eq. (9.11) for a predictive estimator, assuming that the standard deviation σ of the population is not known

$$x_{p,\text{Bayes}} = m'' + t''_p(1/n'' + 1)^{1/2}s'' \tag{9.14}$$

where $t_p'' = t_p''(\alpha, p, \nu'')$ is a fractile of a generalised Student's t-distribution having an appropriate skewness α, for ν'' degrees of freedom (which is generally different from the value $n''-1$).

If the Bayes method is applied for an assessment of material strength, advantage may be taken of the fact that the long-term variability is constant. Then the uncertainty of an assessment of σ and the value $v(\sigma)$ are relatively small, and the variables ν' assessed according to Eq. (9.13) and ν'' assessed according to Eq. (9.12) are relatively high. This factor may lead to a favourable decrease of the value t_p'' and to an augmentation of the estimate of the lower fractile of $x_{p,}$ according to Eq. (9.14). On the other hand, uncertainties in assessment of the mean μ and the variable $v(\mu)$ are usually great, and previous information may not significantly affect the resulting values n'' and m''.

If no previous information is available, then $n' = \nu' = 0$ and the resulting characteristics m'', n'', s'', ν'' equal the sample characteristics m, n, s, ν. In this case the Bayes method is reduced to the prediction method and Eq. (9.14) becomes Eq. (9.11); if σ is known, Eq. (9.10) is used. This particular form of the Bayes method, when no previous information is available, is considered in international documents CEN [7] and ISO [8].

Example 9.4. If previous experience were available for Example 9.3, the Bayes method could be used. Let us suppose that the information is

$$m' = 30.1 \text{ MPa}, \; v(\mu) = 0.50, \; s' = 4.4 \text{ MPa}, \; v(\sigma) = 0.28.$$

It follows from Eq. (9.13) that

$$n' = \left(\frac{4.4}{30.1}\frac{1}{0.50}\right)^2 < 1, \nu' = \frac{1}{2 \times 0.28^2} \approx 6$$

Further on these values are thus considered: $n' = 0$ and $\nu' = 6$. Because $\nu = n-1 = 4$, it follows from Eq. (9.12)

$n'' = 5$, $\nu'' = 10$, $m'' = 29.2$ MPa, $s'' = 4.5$ MPa.

From Eq. (9.14) the fractile estimate follows as

$$x_{p,Bayes} = 29.2 - 1.81 \times \sqrt{\frac{1}{5} + 1} \times 4.5 = 20.3 \text{ MPa}$$

where the value $t_p'' = 1.81$ is given in Table 9.7 for $\alpha = 0$, and $\nu'' = 10$. The resulting strength is thus greater (by 10 %) than the value obtained by the predictive method.

If the population has a log-normal distribution with the skewness $\alpha = 1$, then it follows from Eq. (9.14) considering the value $t_p'' = 1.48$, given in Table 9.7, that

$$x_{p,Bayes} = 29.2 - 1.48 \times \sqrt{\frac{1}{5} + 1} \times 4.5 = 21.9 \text{ MPa}$$

which is a value greater by 8 % than the Bayes estimator for $\alpha = 0$.

Examples 9.3 and 9.4 clearly show that the estimate of characteristic strength (a fractile with the probability $p = 0.05$) assessed from one sample may be expected within a broad range (in Examples 9.3 and 9.4 from 9.9 to 21.9 MPa), depending on the applied method, required confidence, previous information, and on assumptions concerning the population. Note that besides the alternatives considered in Examples 9.3 and 9.4 concerning confidence level and skewness, knowledge of the standard deviation σ of the population and the assumption of a normal distribution or even a negative skewness (in the case of some high- strength materials) may be applied.

In general, more significant differences in the resulting fractiles may occur when the design values of strength are estimated, i.e. fractiles corresponding to a small probability, than when characteristic values ($p \cong 0.001$) are considered. However, a direct estimate of such fractiles from very small ($n < 10$) or small samples ($10 < n < 30$) of the population can be made only if a sufficient amount of information concerning the distribution of the relevant random variable is available. In such a case, it is advisable to compare the results of a direct assessment of the design value with an indirect assessment when the characteristic value is estimated first as a 5 % fractile – then the design value is determined using material partial factors.

References

1. Ang, A.H.-S., Tang, W.H.: Probabilistic Concepts in Engineering. Emphasis on Applications to Civil Environmental Engineering. Wiley, New York (2007)
2. Devore, J., Farnum, N.: Applied Statistics for Engineers and Scientists. Thomson, London (2005)
3. Holický, M., Vorlíček, M.: Distribution asymmetry in structural reliability. Acta Polytechnica **35**(3), 75–85 (1995)
4. Holicky, M.: Reliability Analysis for Structural Design. SUNN MeDIA, Stellenbosch (2009)
5. ISO 3534–1: Statistics – Vocabulary and Symbols – Part 1: Probability and general statistical terms. ISO, Geneve (1993)
6. ISO 12491: Statistical Methods for Quality Control of Building Materials and Components. ISO, Geneve (1997)
7. EN 1990: Eurocode – Basis of Structural Design. CEN, Brussels (2002)
8. ISO 2394: General Principles on Reliability for Structures. ISO, Geneve (1998)

Chapter 10
Testing of Statistical Hypotheses

The testing of statistical hypotheses is one of the essential topics of mathematical statistics, and is often used in engineering and scientific applications. In general, a given hypothesis about a population based on limited sample data is verified specifying a certain high probability (0.95) that the hypothesis is accepted. The complementary small probability (0.05), called significance level, is the probability that the hypothesis will be rejected, even though it is correct (Type I error). Another error may occur when the hypothesis is accepted, although incorrect (Type II error). Operational techniques are provided for testing the deviation of a sample mean from the population mean, testing the deviation of a sample variance from the population variance, testing the difference between two sample means, and testing difference between two sample variances. Two additional frequently applied tests are included: tests of good fit of a given theoretical model, and the testing of outliers in a sample.

10.1 Statistical Tests

Statistical hypotheses are statements, assumptions, or guesses about populations that are to be verified using limited sample data. The procedures of testing the hypotheses consist of specific rules enabling a decision about the population to be made on the basis of sample information. Such decisions are called statistical decisions [1, 2]. The additional terminology used here also follows the standards [3–5]. An important statistical technique concerns tests of outliers [6].

A typical testing procedure may be described as follows: a certain random variable characterizing the hypothesis, called tested variable x, is defined and its probability distribution is investigated. In particular its critical values x_p (one or two values), are specified in such a way that an unfavourable value of x occurs only with a small probability α, called the significance level (commonly between 0.01 and 0.05). The critical values x_p are specified fractiles of tested variables that define the

M. Holický, *Introduction to Probability and Statistics for Engineers*,
DOI 10.1007/978-3-642-38300-7_10, © Springer-Verlag Berlin Heidelberg 2013

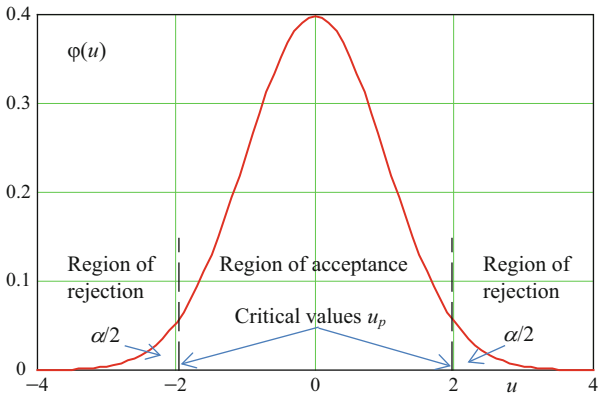


Fig. 10.1 Area of acceptance and rejection with the critical values $u_{p1} = u_{\alpha/2} = -1.96$ and $u_{p2} = u_{1-\alpha/2} = 1.96$ for the significance level $\alpha = 0.05$

acceptance and rejection region (see Fig. 10.1). As a rule there are two critical values x_{p1} and x_{p2} corresponding to the probabilities $p_1 = \alpha/2$ and $p_2 = 1-\alpha/2$ [1, 2].

Let the data available from a sample yield the value of the tested variable equal to x_0, called the test value. Then, if the test value x_0 is within the margin of acceptance delineated by the critical values x_p, the hypothesis is accepted; if test value x_0 is outside the margin of acceptance (i.e. inside the rejection area), the hypothesis is rejected. The whole test procedure may be summarized by the following steps:

1. Tested variable x is defined.
2. The critical values x_p are specified for the significance level α.
3. The test value x_0 is evaluated using sample information.
4. The statistical decision is made:

   (a) If x_0 is inside the acceptance margin, the hypothesis is accepted.
   (b) If x_0 is inside the rejection area the hypothesis is rejected.

It should be emphasised that statistical decisions are made with limited information, and will therefore show some errors regarding any decision that is made. There will always be some non-zero probability that a hypothesis will be rejected when it should have been accepted; this error is called Type I error. The Type I error occurs with the probability equal to the significance level α. Commonly, the significance level is a small probability, typically $\alpha = 0.05$ (a more strict significance, level) or $\alpha = 0.01$ (less strict significance level), that the hypothesis will be rejected incorrectly. So, with an increasing significance level (probability) α the decision becomes more strict (i.e. the chance of rejecting the hypothesis increases).

However, there is another error which cannot be avoided, and this is called Type II error. It refers to the decision that a hypothesis is accepted when it should be rejected. Analysis of the probability of Type II error is more complicated than just the specification of the significant level α (which is the probability of a Type I error) and will not be discussed here in detail. Nevertheless, it should be mentioned that with a decreasing probability of Type I error (significance level α), the probability

of Type II error increases. Note that both types of errors (Type I and Type II) may generally occur in statistical decisions.

Regions of acceptance and rejection for tests based on the standardised random variable u having normal distribution are shown in Fig. 10.1, together with the critical values $u_{p1} = u_{\alpha/2} = -1.96$ and $u_{p2} = u_{1-\alpha/2} = 1.96$ corresponding to a significance level of $\alpha = 0.05$.

10.2 Deviation of Sample Mean from Population Mean

The hypothesis that a sample having the mean m is taken from a population with the mean μ is to be tested. The sample size is denoted by n. The testing procedure depends on whether the standard deviation of the population σ is known or unknown and the sample standard deviation s must be used instead of σ.

10.2.1 Known Standard Deviation

If the population standard deviation σ is known, the mean m is a normally distributed random variable with the mean m and the variance σ^2/n. Then the difference of the sample mean m from the population mean μ is tested using a variable expressed as

$$u = \frac{m - \mu}{\sigma}\sqrt{n} \tag{10.1}$$

Here u is the normally distributed standardised random variable. The tested value u_0, evaluated for a particular sample mean m, is compared with the critical value u_p specified for a given significance level α.

Example 10.1. Consider a test based on the standardised normal variable u. If the significance level $\alpha = 5$ % (2.5 % on each side of the distribution), then the critical values u_p are taken from the standardised normal distribution (numerical values determined from table in Annex 7 or by a software product) as

$$u_{0.025} = -1.96; \quad u_{0.975} = 1.96)$$

Thus the critical values may be expressed as $u_p = \pm 1.96$. Note that if the significance level $\alpha = 1$ % (0.5 % on each side) then the critical values are $u_p = \pm 2.576$.

Example 10.2. A sample of $n = 16$ tests of concrete strength yields the mean value $m = 28.8$ MPa. From long term production it follows that the population mean is $\mu = 31.0$ MPa and standard deviation $\sigma = 4.20$ MPa. The hypothesis that

the sample mean is equal to the population mean is to be tested. From Eq. (10.1) the test value follows as

$$u_0 = \frac{28.8 - 31.0}{4.20}\sqrt{16} = -2.095$$

It follows from Example 10.1 that for the significance level $\alpha = 5$ % the critical values $-u_{\alpha/2} = u_{1-\alpha/2} = 1.96$. Obviously the test value $u_0 = -2.095$ is outside the acceptance region delineated by the critical value ± 1.96 (see also Fig. 10.1) and the hypothesis that the sample is taken from the population having the mean $\mu = 31.0$ MPa is to be rejected.

However, if the significance level $\alpha = 1$ % (with decreasing error of Type I and with increasing error of the Type II), the critical values are ± 2.576 and the hypothesis that a sample having the mean $m = 28.8$ MPa is taken form the population having the mean $\mu = 31.0$ MPa is to be accepted.

10.2.2 Unknown Standard Deviation

If the population standard deviation σ is unknown, then it should be substituted by its point estimate s, and instead of the tested variable u the variable t should be used. Then Eq. (10.1) becomes

$$t = \frac{m - \mu}{s}\sqrt{n - 1} \tag{10.2}$$

The variable t has t-distribution with $n-1$ degrees of freedom. For large samples ($n > 30$) the t-distribution can be approximated by the standardized normal distribution.

Example 10.3. Consider again the sample of $n = 16$ tests from Example 10.2 with the mean $m = 28.8$ MPa and standard deviation $s = 4.20$ MPa. The hypothesis that the sample is taken from the population having the mean is $\mu = 31.0$ MPa is to be tested. From Eq. (10.2) the test value follows as

$$t_0 = \frac{28.8 - 31.0}{4.20}\sqrt{15} = -2.029$$

For the significance level $\alpha = 5$ % the critical value $-t_{\alpha/2} = t_{1-\alpha/2} = 2.131$ for $\nu = 16-1 = 15$ degrees of freedom (taken from tables or determined by using a software products). The test value $u_0 = -2.029$ is then inside the acceptance area ± 2.131 and the hypothesis is to be accepted.

10.3 Deviation of Sample Variance from Population Variance

The deviation of the variance s^2 of a sample of n units taken from a population having the variance σ^2 is to be tested. The tested variable is defined as χ^2 variable that has already been introduced in Chap. 8.

$$\chi^2 = \frac{ns^2}{\sigma^2} \tag{10.3}$$

Here the variable χ^2 should be considered with $n-1$ degrees of freedom. The critical values of the variable χ^2 should be defined separately for the critical values given by the lower boundary χ^2_{p1} and the upper boundary χ^2_{p2} corresponding to the significance levels $\alpha = p_1 + 1-p_2$ (χ^2-distribution is asymmetrical).

Example 10.4. The variance of the sample of $n = 16$ tests in Example 10.3 is 5.1 MPa. Its deviation from the population standard deviation $\sigma = 4.2$ should be tested on the significance level $\alpha = 1-p_2 = 2.5$ % (only the upper boundary is considered). It follows from Eq. (10.3) that the test value

$$\chi^2_0 = \frac{16 \times 5.1^2}{4.2^2} = 23.6$$

The critical value can be determined from tables or by software products for $\nu = 16-1 = 15$ degrees of freedom as $\chi^2_{p2} = \chi^2_{0.975} = 27.5$. As $\chi^2_0 < \chi^2_{p2}$ the deviation of the sample variance from the population variance is insignificant. The probability density function $\varphi(\chi^2)$ for $\nu = 15$ degrees of freedom is shown in Fig. 10.2.

10.4 Difference Between Two Sample Means

Consider two samples of sizes n_1 and n_2, the means m_1 and m_2 and variances σ_1^2 and σ_2^2. The difference between the means m_1 and m_2 (it is assumed that $m_1 > m_2$) is to be tested. The statistical tests have two different procedures, depending on the following circumstances:

- The variances are known but generally different, $\sigma_1^2 \neq \sigma_2^2$
- The variances are unknown and sample variances s_1^2 and s_2^2 must be considered

10.4.1 Variances are Known

If the variances are known but generally different, $\sigma_1^2 \neq \sigma_2^2$, then it could be shown (see also Sect. 10.2) that the following tested variable is a standardised normal variable

$$u = \frac{m_1 - m_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{10.4}$$

The critical value $u_{\alpha/2}$ and $u_{1-\alpha/2}$ (the lower and upper values) are determined for a specified significance level α, similarly as in Sect. 10.2.

Example 10.5. There are two producers of bricks with the following characteristics of tests concerning strength (in MPa):

- $n_1 = 12, m_1 = 10.73, \sigma_1 = 1.512$
- $n_2 = 10, m_2 = 9.84, \sigma_2 = 1.603$

The test value follows from Eq. (10.4) as

$$u_0 = \frac{10.73 - 9.84}{\sqrt{\frac{1.512^2}{12} + \frac{1.603^2}{10}}} = 1.330$$

Considering the significance level $\alpha = 0.05$ the critical values u_p are $u_{\alpha/2} = -1.96$ and $u_{1-\alpha/2} = 1.96$. The difference between the means is insignificant (the test value is within the margin of acceptance as $u_{\alpha/2} < u_0 < u_{1-\alpha/2}$) and the hypothesis that the mean values of the brick strengths is the same in both factories can be accepted.

10.4.2 Variances are Unknown

If the variances are unknown sample variances s_1^2 and s_2^2 must be considered then it could be shown (see also Sect. 10.2) that the following tested variable has the t-distribution

$$t_0 = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}} \tag{10.5}$$

The critical values $t_{\alpha/2}$ and $t_{1-\alpha/2}$ are based on the combination of two critical values $t_{1,\alpha/2}$ and $t_{2,\alpha/2}$ of t-distribution and degrees of freedom $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ appropriate to the samples involved and can be expressed as

$$t_{\alpha/2} = -t_{1-\alpha/2} = \frac{t_{1,\alpha}\frac{s_1^2}{n_1-1} + t_{2,\alpha}\frac{s_2^2}{n_2-1}}{\frac{s_1^2}{n_1-1} + \frac{s_2^2}{n_2-1}} \tag{10.6}$$

Here the symmetry of t-distribution is taken into account.

Example 10.6. There are two producers of bricks with the following characteristics of brick testings:

- $n_1 = 12, m_1 = 10.73, s_1 = 1.492$
- $n_2 = 10, m_2 = 9.84, s_2 = 0.813$

The test value follows from Eq. (9.4) as

$$t_0 = \frac{10.73 - 9.84}{\sqrt{\frac{1.492^2}{11} + \frac{0.813^2}{9}}} = 1.695$$

The critical value follows from Eq. (10.6) considering $\nu_1 = 11$, $\nu_2 = 9$, $\alpha = 0.05$, $t_{1,\alpha/2} = -2.201$ and $t_{2,\alpha/2} = -2.262$ (determined by tables or using software products):

$$-t_{\frac{\alpha}{2}} = t_{1-\frac{\alpha}{2}} = \frac{2.201\frac{1.492^2}{11} + 2.262\frac{0.813^2}{9}}{\frac{1.492^2}{11} + \frac{0.813^2}{9}} = 2.217$$

As in the previous Example 10.4, the difference between the means is insignificant, as the test value is within the margin of acceptance $t_{\alpha/2} < t_0 < t_{1-\alpha/2}$ and the hypothesis that the mean value of the strength in both factories is the same can be accepted.

10.5 Difference Between Two Sample Variances

Consider again two samples of the sizes n_1 and n_2, the means m_1 and m_2 and variances s_1^2 and s_2^2. The statistical hypothesis that both samples are taken from populations with the same variance σ^2 (its value is not needed) is to be tested. In this case the tested variable is defined using sampling distribution F for $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ degrees of freedom:

$$F = \frac{\frac{n_1}{n_1-1}s_1^2}{\frac{n_2}{n_2-1}s_2^2} \tag{10.7}$$

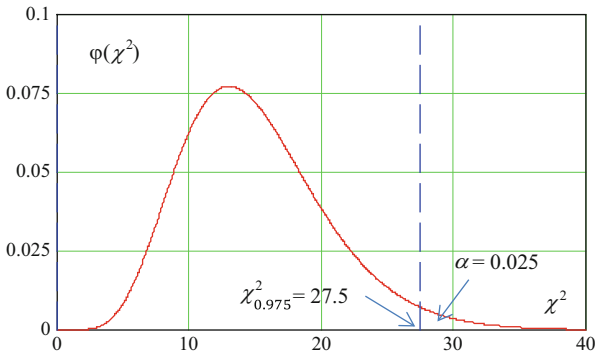


Fig. 10.2 Probability density function $\varphi(\chi^2)$ for $\nu = 15$ degrees of freedom

The subscripts referring to samples are commonly chosen in such a way that $s_1^2 > s_2^2$. So the critical value $F_p > 1$ (a one-sided critical area for the significance level $\alpha = 0.05$ or 0.01).

Example 10.7. Consider the two samples from Example 10.6:

- $n_1 = 12$, $s_1 = 1.492$
- $n_2 = 10$, $s_2 = 0.813$

The statistical hypothesis that variances of corresponding populations are equal is to be tested on the significance level $\alpha = 0.05$. The test value F_0 follows from Eq. (10.7):

$$F_0 = \frac{\frac{12}{11}1.492^2}{\frac{10}{9}0.813^2} = 3.31$$

The critical value $F_p = F_{1-\alpha} = F_{0.95}$ (the upper fractile 0.95) for $n_1 - 1 = 11$ and $n_2 - 1 = 9$ degrees of freedom is 3.10 (see Fig. 8.3) and the hypothesis is therefore rejected (the difference between the variances is significant as $F_0 > F_p = 3.10$).

10.6 Tests of Good Fit

The statistical hypothesis that a given sample of n observations is taken from a population of a certain type of distribution $\Phi(x)$ can be examined using different tests. The most general seems to be the χ^2-test that can be used for both discrete and continuous random variables. Another popular test is called the K-test (developed by A.N. Kolmogorov).

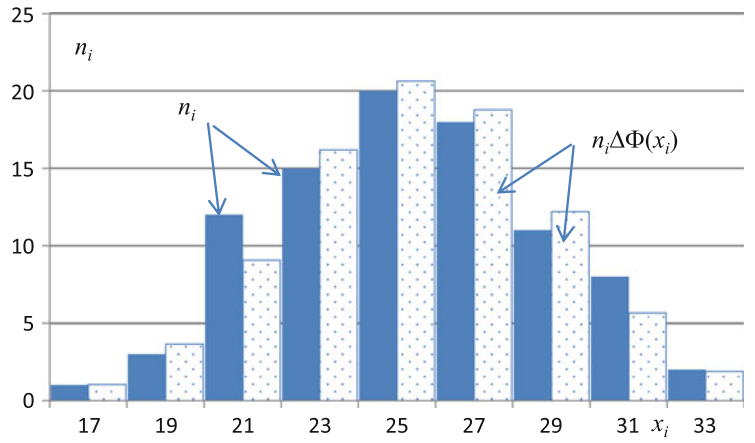


Fig. 10.3 Frequencies n_i, and the theoretical frequencies $n_i\Delta\Phi(x_i)$ assuming normal distribution, $\mu = 25.44$ MPa and standard deviation $\sigma = 3.45$

The definition domain of a random variable x is discretised by the values x_i ($i = 1, 2, \ldots k$) into k classes and the class frequencies n_i are compared with the relevant theoretical values $n\,[\Phi(x_{i+1}) - \Phi(x_i)] = n\,\Delta\Phi(x)$. The classes should be adjusted to the condition that the minimum theoretical class frequency is 5, $n\,\Delta\Phi(x) \geq 5$. The tested variable is then given as

$$\chi^2 = \sum_1^k \frac{[n_i - n\Delta\Phi(x_i)]^2}{n\Delta\Phi(x_i)} \tag{10.8}$$

The critical value $\chi_p^2 = \chi_{1-\alpha}^2$ is determined for specified significance level α. Typically $\alpha = 0.05$ as a strict level or $\alpha = 0.01$ for less strict level. The critical value χ_p^2 corresponds to the upper fractile having the probability $\alpha = 1-p$ (0.05 and 0.01) of being exceeded. The degree of freedom is in this case given as $\nu = k-c-1$, where c denotes number of distribution parameters that are determined using sample data ($c = 2$ for normal distribution).

Example 10.8. Consider the test results described in Example 3.9. The following table summarises the χ^2-test and illustrates the use of Eq. (10.8). It shows classes 1–9, class marks x_i (in MPa), the frequency n_i, and the theoretical frequencies $n_i\Delta\Phi(x_i)$ assuming the normal distribution having the mean $\mu = 25.44$ MPa and standard deviation $\sigma = 3.45$. The table also indicates the calculation of the test variable χ^2.

The frequencies n_i, and $n_i\Delta\Phi(x_i)$ are also shown in Fig. 10.3. For each class the full columns represent the observed frequencies n_i and the transparent columns the theoretical frequencies $n_i\Delta\Phi(x_i)$. It appears from a visual investigation of Fig. 10.3

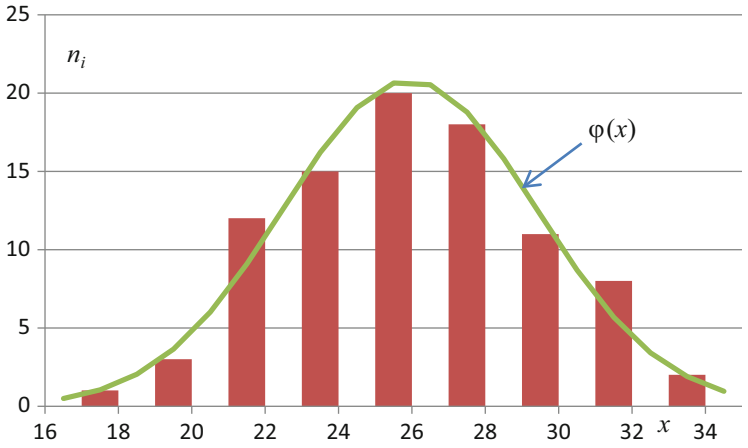


Fig. 10.4 Histogram and probability density function $\varphi(x)$ of normal distribution

that the differences $n_i - n_i\Delta\Phi(x_i)$ are not substantial, and that the normal distribution pattern may fit the observed data well.

i	x_i	n_i	$n_i\Delta\Phi(x_i)$	$[n_i - n\ \Delta\Phi(x_i)]^2$	n_i	$n_i\Delta\Phi(x_i)$	$[n_i - n\ \Delta\Phi(x_i)]^2$
1	17	1	1.044	0.002			
2	19	3	3.645	0.416			
3	21	12	9.093	8.450	16	13.770	4.973
4	23	15	16.209	1.461	15	16.209	1.461
5	25	20	20.646	0.417	20	20.646	0.417
6	27	18	18.792	0.627	18	18.792	0.627
7	29	11	12.222	1.494	11	12.222	1.494
8	31	8	5.680	5.380	10	7.560	5.954
9	33	2	1.887	0.013			
Σ				18.260		89.198	14.925

In order to comply with the recommendation to combine classes when the theoretical frequency $n\ \Delta\Phi(x)$ is less than 5, the first three and last two classes are combined. Then the test value χ^2 follows from Eq. (10.8) as

$$\chi^2 = 14.925/89.198 = 0.167$$

The critical value for the confidence level $\alpha = 0.05$ and $\nu = k-c-1 = 6-2-1 = 3$ degress of freedom is $\chi_p^2 = \chi_{1-\alpha}^2 = \chi_{0.95}^2 = 7.815$ and the hypothesis that normal distribution fits the observed data well is confirmed.

A visual inspection of Fig. 10.4 confirms the test conclusion that the normal distribution, having the mean $\mu = 25.44$ MPa and standard deviation $\sigma = 3.45$, fit the available data of concrete strength well.

Example 10.9. A sample of 100 observations is split into 8 classes, taking into account conditions concerning the theoretical class frequency $n\ \Delta\Phi(x) \geq 5$. Then the test value is determined using Eq. (10.8) as $\chi^2_{0.95} = 1.55$. The critical value χ^2_p is determined for significance level $\alpha = 0.05$ (upper fractile 0.95), and $\nu = 8 - 2 - 1 = 5$ degrees of freedom ($c = 2$ for normal distribution). The critical value is then $\chi^2_{0.95} = 11.07$. Obviously, normal distribution is a suitable theoretical model.

Figure 10.4 shows the histogram of the data together with the probability density function of the normal distribution having the mean $\mu = 25.44$ MPa and standard deviation $\sigma = 3.45$ MPa.

10.7 Tests of Outliers

If the minimum or maximum observation of a sample is at a markedly greater distance from the remaining observations, then it could be an outlier (caused by an error in the observation). It should be verified by a statistical test to determine whether the deviation of such an observation is just random or significant, and then it may be discounted from any further evaluation.

Two tests of outliers are commonly used: Grubbs test and Dixons test. In both cases it is assumed that the sample has been taken from a population having the normal distribution. Sample data are ordered upward into a raw:

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \ldots \leq x_{(n_{-1})} \leq x_{(n)} \tag{10.9}$$

The extreme values $x_{(1)}$ and $x_{(1)}$ are to be tested.

10.7.1 Grubbs Test

The Grubbs test was developed by F.E. Grubbs. The procedure is dependend on whether the standard deviation of the population σ is known or not. If σ is known, then the tested variable is given as

$$\tau_0 = \frac{m - x_{(1)}}{\sigma} \quad \text{or} \quad \tau_0 = \frac{x_{(n)} - m}{\sigma} \tag{10.10}$$

If σ is unknown, then the sample standard deviation s should be used instead of σ, and the tested variable is then given as

$$\tau'_0 = \frac{m - x_{(1)}}{s} \quad \text{or} \quad \tau'_0 = \frac{x_{(n)} - m}{s} \tag{10.11}$$

The critical values τ_p and τ_p' are given in special tables that are available on the internet and copied below for the usual significance levels $\alpha = 1-p = 0.05$ and 0.01. If

$$\tau_0 \leq \tau_p \quad or \quad \tau'_0 \leq \tau'_p \tag{10.12}$$

then the deviations of the extreme values are considered as random. In the opposite case the deviations are considered to be significant and the relevant observations should be discounted from any further evaluation.

Example 10.10. The relative compaction of sand and gravel are recorded using 12 randomly chosen specimens:

0.83; 0.88; 0.84; 0.78; 0.82; 0.82; 0.86; 0.81; 0.98; 0.83; 0.85; 0.80

Sample data are ordered upwards into a row as follows:

$0.78 \leq 0.80 \leq 0.81 \leq 0.82 \leq 0.82 \leq 0.83 \leq 0.83 \leq 0.84 \leq 0.85 \leq 0.86 \leq 0.88 \leq 0.98$

The value 0.98 seems to be dubious. The mean and standard deviation of the above sample are $m = 0.842$ and $s = 0.049$. The test value follows from Eq. (10.11) as

$$\tau'_0 = \frac{0.98 - 0.842}{0.049} = 2.816$$

The critical value for the significance level $\alpha = 0.01$ is $\tau'_{0.99} = 2.663$. Thus the observed value 0.98 is really an outlier and should be deleted from further analysis.

Critical values τ_p and τ_p' for the Grubbs test

n	α: 0.05	0.01	n	α: 0.05	0.01
3	1.412	1.414	15	2.493	2.800
4	1.689	1.723	16	2.523	2.837
5	1.869	1.955	17	2.551	2.871
6	1.996	2.130	18	2.577	2.903
7	2.093	2.265	19	2.600	2.932
8	2.172	2.374	20	2.623	2.959
9	2.237	2.464	21	2.644	2.984
10	2.294	2.540	22	2.664	3.008
11	2.343	2.606	23	2.683	3.030
12	2.387	2.663	24	2.701	3.051
13	2.426	2.714	25	2.717	3.071
14	2.461	2.759			

10.7.2 Dixon Test

The Dixon test, developed by W.J. Dixon, is based directly on the sample data Eq. (10.9) and does not need sample characteristics (the mean and standard deviation). If the minimum observation $x_{(1)}$ is dubious, then the tested variable is defined as

$$\omega_0 = \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}} \tag{10.13}$$

or, if $x_{(n)}$ is also dubious then the tested variable is

$$\omega'_0 = \frac{x_{(2)} - x_{(1)}}{x_{(n-1)} - x_{(1)}} \tag{10.14}$$

Similarly, if the maximum value $x_{(n)}$ is dubious then the tested variable is

$$\omega = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}} \tag{10.15}$$

or, if $x_{(1)}$ is also dubious then the tested variable

$$\omega'_0 = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(2)}} \tag{10.16}$$

The critical values ω_p and ω'_p are given in special tables that are available on the internet and copied below for the common significance level $\alpha = 1-p = 0.05$ and 0.01. If

$$\omega_0 \leq \omega_p \quad or \quad \omega'_0 \leq \omega'_p \tag{10.17}$$

then the deviations of the extreme values are considered as random. In the opposite case the deviations are considered to be significant and the relevant observations should be discounted from any further evaluation.

Critical values ω_p and ω'_p for the Dixon test

n	α: 0.05	0.01	n	α: 0.05	0.01
3	0.941	0.988	17	0.320	0.416
4	0.765	0.889	18	0.313	0.407
5	0.642	0.780	19	0.306	0.398
6	0.560	0.698	20	0.300	0.391
7	0.507	0.637	21	0.295	0.384
8	0.468	0.590	22	0.290	0.378
9	0.437	0.555	23	0.285	0.372

(continued)

10	0.412	0.527	24	0.281	0.367
11	0.392	0.502	25	0.277	0.362
12	0.376	0.482	26	0.273	0.357
13	0.361	0.465	27	0.269	0.353
14	0.349	0.450	28	0.266	0.349
15	0.338	0.438	29	0.263	0.345
16	0.329	0.426	30	0.260	0.341

Example 10.11. Let us apply the Dixon test for the data describe in Example 10.10. As the maximum observation $x_{(n)} = 0.98$ is dubious, the following observations are needed

$$x_{(1)} = 0.78; \quad x_{(n-1)} = 0.88; \quad x_{(n)} = 0.98$$

The test value follows from Eq. (10.15) as

$$\omega'_0 = \frac{0.98 - 0.88}{0.98 - 0.78} = 0.5$$

The critical value for $n = 12$ and the significance level $\alpha = 0.01$ is $\omega_p = 0.482$ (obtained from tables on internet). Thus, the maximum value $x_{(n)} = 0.98$ is really an outlier and should be deleted from further analysis.

References

1. Ang, A.H.-S., Tang, W.H.: Probabilistic Concepts in Engineering. Emphasis on Applications to Civil Environmental Engineering. Wiley, New York (2007)
2. Devore, J., Farnum, N.: Applied Statistics for Engineers and Scientists. Thomson, London (2005)
3. ISO 12491: Statistical Methods for Quality Control of Building Materials and Components. ISO, Geneve (1997)
4. ISO 3534–1: Statistics – Vocabulary and Symbols – Part 1: Probability and General Statistical Terms. ISO, Geneve (1993)
5. ISO 3534–2: Statistics – Vocabulary and Symbols – Part 2: Statistical Quality Control. ISO, Geneve (1993)
6. Barnett, V., Lewis, T.: Outliers in Statistical Data, 2nd edn. Wiley, New York (1985)

Chapter 11
Correlation and Regression

Two dimensional random variables are frequently investigated in engineering and scientific applications. Normal distribution is commonly assumed as a theoretical model for the population of two dimensional random variables. The mutual linear dependence of the two variables is described by the coefficient of correlation. Regression lines are used to analyse the dependence of one random variable, on one hand as the dependent variable, on the other as the independent variable. While the correlation is a symmetrical property with respect to the two random variables, regression is influenced by the choice of a dependent or independent variable. The estimate of the coefficients of correlation and regression from sample data is an extremely important task, as only limited samples for two dimensional random variables are commonly available. For the same reason testing concerning the coefficient of correlation and regression is an essential step in many engineering and scientific applications.

11.1 Two-Dimensional Random Variables

Very often in engineering or scientific practice a relationship between two or more variables has to be investigated. It is then desirable to express this relationship in mathematical form. It should be emphasised that correlation theory investigates the mutual dependence of two or more variables, while regression analysis studies the dependence of one random variable (as a dependent variable) on the other variable that is considered independent [1, 2].

When only two variables are involved, then we talk about simple correlation or regression. When more than two variables are involved, then multiple correlation or regression is applied [3, 4]. This chapter will consider only simply correlation and regression. A detailed discussion concerning the multivariate random variables, probabilistic models, parameters of population and sample characteristics can be found in specialist literature [1, 2].

M. Holický, *Introduction to Probability and Statistics for Engineers*,
DOI 10.1007/978-3-642-38300-7_11, © Springer-Verlag Berlin Heidelberg 2013

If two variables (two characteristics) X and Y are studied for each item (entity), every time a set of conditions π (see Sect. 2.1) is realised, i.e. a certain random event is realised, and given that the variable X takes on the very value x, and the variable Y takes on the very value y, then the variables X and Y form a pair of joint random variables. An example is the force X and the weight Y studied when a concrete cube fails when loaded under given conditions into a test machine. It is certainly possible to study more than two characteristics, e.g. the force, weight and moisture content.

In the following, only two-dimensional random variables, having two normally distributed components (two joint random variables), X and Y, are analysed. The realisations of each component are denoted by the small letters x and y. The summary of all possible realizations, x and y, of a pair of joint random variables, X and Y, is called the two-dimensional population. Similarly, as in the case of the one-dimensional random variable, the two-dimensional random variable is described by the distribution of probabilities, i.e. by a function which determines the probability that the random variables X and Y make up part of some given sets (for continuous random variables), or take on some given values (for discrete random variables). The two-dimensional distribution function $\Phi(x, y)$ (sometimes denoted $\Phi_{XY}(x, y)$) gives, for every pair of values x, y, the probability that the random variable X is less than, or equal to, x, and the random variable Y is less than, or equal to, y

$$\Phi(x, y) = \mathrm{P}(X \leq x; \quad Y \leq y) \tag{11.1}$$

The probability density function of a continuous random variable $\varphi(x)$ is the derivative (if it exists) of the distribution function

$$\varphi(x, y) = \frac{\partial^2 \Phi(x, y)}{\partial x \partial y} \tag{11.2}$$

The marginal distribution function of the variable X, $\Phi_X(x)$, is a special case of the distribution function $\Phi(x, y)$ without any constraint on the variable Y, i.e. for all realizations $Y < \infty$

$$\Phi(x, \infty) = \mathrm{P}(X \leq x; \quad Y \leq \infty) = \Phi_X(x) \tag{11.3}$$

The marginal distribution function of the variable Y, $\Phi_Y(y)$, is defined in a similar way. It is a special feature of the distribution function $\Phi(x, y)$, without any constraint on the variable X, i.e. for the sum of all possible realizations of the variable $X < \infty$

$$\Phi(\infty, y) = \mathrm{P}(X \leq \infty; \quad Y \leq y) = \Phi_Y(y) \tag{11.4}$$

We can say that the random variables X and Y are independent if it holds that

$$\Phi(x, y) = \Phi(x, \infty) \quad \Phi(\infty; y) = \Phi_X(x)\Phi_Y(y) \tag{11.5}$$

Then it holds for the probability density function that

$$\varphi(x, y) = \varphi_X(x)\, \varphi_Y(y) \tag{11.6}$$

where $\varphi_X(x)$ and $\varphi_Y(y)$ are the marginal probability density functions of the variables X and Y.

The two-dimensional random variable is described by moment parameters and various types of distribution (usually by the normal), in a similar way to one-dimensional variables. Besides the one-dimensional moments which lead to the definition of averages μ_X, μ_Y, and the standard deviations σ_X, σ_Y, the joint moments of both variables X and Y are also applied. The most important one is the joint central moment of the first order σ_{XY}, which is called the covariance

$$\sigma_{XY} = \int \phi(x, y)(x - \mu_X)(y - \mu_Y)dxdy \tag{11.7}$$

The covariance provides the basis for the definition of the correlation coefficient ρ_{XY}

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{11.8}$$

It always holds for the value of the correlation coefficient that $-1 \leq \rho_{XY} \leq +1$. If the variables X and Y are independent, then $\rho_{XY} = 0$. An inverse proposition holds only in the case of the two-dimensional normal distribution (which is commonly applied and is described below). In the case of multivariate random variables $\boldsymbol{X}$ [X_1, X_2,... X_n], the covariance σ_{ij} and the correlation coefficients ρ_{ij} between the individual components X_1, X_2,... X_n form matrices. The matrix of covariances is applied in the transformation of the vector of dependent variables to the vector of independent random variables, which are used in reliability analysis of more complex problems (see the software product STRUREL).

11.2 Two-Dimensional Normal Distribution

A two-dimensional normal distribution of two continuous random variables X and Y, having the parameters μ_x, μ_y, σ_x, σ_y, and a correlation coefficient $\rho_{xy} = \rho$, is given by the following equation

$$\varphi(x,y) = \frac{1}{2\pi\,\sigma_X\,\sigma_Y\,\sqrt{1-\rho^2}}$$
$$\exp\left(-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)\right) \quad (11.9)$$

The marginal distributions $\varphi_X(x)$ and $\varphi_Y(y)$ are also normal and have parameters μ_X, σ_X and μ_Y, σ_Y similar to the conditional distributions for given $y = y_0$ and $x = x_0$, which have parameters $\mu_x + \rho(y_0-\mu_Y)\sigma_X/\sigma_Y$, $\sigma_x(1-\rho^2)^{1/2}$ and $\mu_Y + \rho(x_0 - \mu_X)\sigma_Y/\sigma_X$, $\sigma_Y(1-\rho^2)^{1/2}$ [1]. The conditional distributions may come in useful for (very frequent) indirect experimental verification of properties of one of the joint random variables X and Y by means of the other.

Similarly, as in the case of the one-dimensional random variable through transformations, the standardized random variables U and V are given as

$$U = \frac{X-\mu_X}{\sigma_X}, \quad V = \frac{Y-\mu_Y}{\sigma_Y} \quad (11.10)$$

The standardized two-dimensional normal distribution can then be written in the form

$$\varphi(u,v) = \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left(u^2 - 2\rho uv + v^2\right)\right) \quad (11.11)$$

The bivariate normal distribution can be generalized [1, 2] to a distribution of multivariate random variables described by the vector $\boldsymbol{X}$ $[X_1, X_2, \ldots X_n]$, where the covariance's and correlation coefficients between the individual components X_1, $X_2, \ldots X_n$, form matrices.

11.3 Two-Dimensional Samples

As a result of the sampling procedure, paired observations x_1,y_1; x_2,y_2; …; x_n,y_n, may be obtained and analysed in a similar way to one dimensional random samples. Thus the sample means m_X and m_Y, and sample standard deviations s_X and s_Y, may be obtained using the formulae given in Chap. 3. However, in addition to these moment characteristics, so-called product-moments can be generally defined. In the following, only the first order product moment, called covariance, is considered.

Using a sample of paired observations x_1,y_1; x_2,y_2; …; x_n,y_n, the sample covariance is given as

$$s_{XY} = \frac{1}{n}\sum_i (x_i - m_x)(y_i - m_y) \quad (11.12)$$

Note that an unbiased estimate of the population's covariance should have the denominator $n-1$ similar to the case of the one dimensional random variable discussed in Chap. 8. Analogically to (11.8), the sample correlation coefficient follows as

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{\sum_i (x_i - m_X)(y_i - m_Y)}{\sqrt{\sum_i (x_i - m_X)^2 \sum_i (y_i - m_Y)^2}} \tag{11.13}$$

The sample correlation coefficient r_{XY} is often used for the numerical expression of the mutual linear dependence between X and Y in a number of paired observations. The value of r_{XY} lies between -1 and $+1$. If it equals one of these limits, it means that the dependence between X and Y in a number of paired observations is exactly linear. When possible, a scatter diagram showing the observed set should be used to verify the linearity, and possibly to reduce the domain so that the assumption of linearity is justified.

Usually, the coefficient of correlation is used to classify verbally the degree of mutual dependence of random variables X and Y. The following scale is sometimes used:

$|r_{XY}| \leq 0.3$ low degree of dependence
$0.3 < |r_{XY}| \leq 0.5$ some degree of dependence
$0.5 < |r_{XY}| \leq 0.7$ significant degree of dependence
$0.7 < |r_{XY}| \leq 0.9$ high degree of dependence
$0.9 < |r_{XY}|$ very high degree of dependence

The above scale provides only an indicative marking of mutual dependence that is not based on any objective criteria.

Example 11.1. Measurements of ten components yield as a rule positive deviations from the nominal width and high shown in the following table.

i	1	2	3	4	5	6	7	8	9	10
x_i	3	4	4	5	6	7	7	8	8	9
y_j	4	6	5	6	7	4	8	7	6	9

The following characteristics can be found

$$m_X = 6.1;\ m_Y = 6.2;\ s_X = 1.92;\ s_Y = 1.54;\ s_{XY} = 1.88;\ r_{XY} = 0.64$$

The coefficient of correlation $r_{XY} = 0.64$ indicates a significant degree of mutual dependence of deviations from width and height. The point graph of observed values is shown in Fig. 11.1.

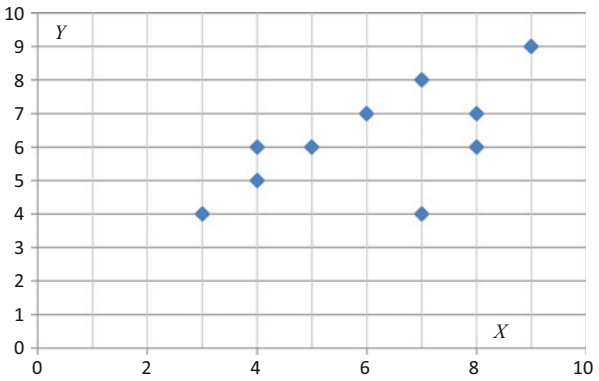


Fig. 11.1 Point diagram of data in the Example 11.1

11.4 Regression Lines

Regression lines describe a linear dependence of one of the variables Y or X on the other variable considered as an independent variable. The regression line of the dependent variable Y on the independent variable X is written as

$$Y = a_0 + a_1 X \tag{11.14}$$

The constants a_0 and a_1 denote the coefficients of regression of Y on X. Similarly, the regression line of dependent X on independent Y is written as

$$X = b_0 + b_1 Y \tag{11.15}$$

The constants b_0 and b_1 are the coefficients of regression of X on Y. Figure 11.1 shows regression lines (11.4) and (11.15) for the data of Example 11.1.

The regression coefficients a_0 and a_1 (and similarly b_0 and b_1) are commonly derived using the well known method of the least squares minimising the residuum

$$\sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 \tag{11.16}$$

Using this method it can be shown that the regression coefficients are

$$a_0 = m_Y - r_{XY} m_X s_Y / s_X \tag{11.17}$$

$$a_1 = r_{XY} s_Y / s_X \tag{11.18}$$

Similarly the regression coefficients b_0 and b_1 are given as

$$b_0 = m_X - r_{XY} m_X s_X / s_Y \tag{11.19}$$

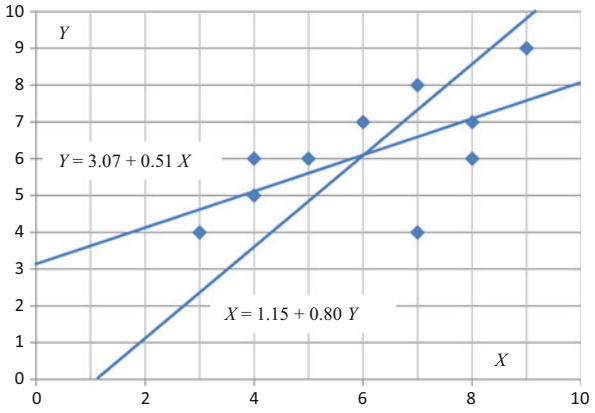


Fig. 11.2 Regression lines of Example 11.2

$$b_1 = r_{XY} s_X / s_Y \tag{11.20}$$

Example 11.2. Measurements of ten components described in Example 11.1 indicate the following regression coefficients (Fig. 11.2)

$a_0 = 6.2 - 0.64 \times 6.1 \times 1.54/1.92 = 3.07$; $a_1 = 0.64 \times 1.54/1.92 = 0.51$

$b_0 = 6.1 - 0.64 \times 6.2 \times 1.92/1.54 = 1.15$; $b_1 = 0.64 \times 1.92/1.54 = 0.80$

So, the regression lines of Y on X and X on Y are

$Y = 3.07 + 0.51\ X$ and $X = 1.15 + 0.80\ Y$

Note that the regression lines intersect at the mean point ($m_X = 6.1$; $m_Y = 6.2$).

11.5 Estimation of the Coefficient of Correlation

The sampling distribution of the coefficients of correlation $r = r_{XY}$ can be approximated by normal distribution, using the transformation (a simplified notation for r is used for r_{XY} in the following)

$$z = \frac{1}{2} ln \frac{1+r}{1-r} \tag{11.21}$$

The mean and standard deviation of the transformed variable z are

$$\mu_z = \frac{1}{2} ln \frac{1+\rho}{1-\rho} \tag{11.22}$$

$$\sigma_z = \frac{1}{\sqrt{n-3}} \tag{11.23}$$

Here ρ denotes the coefficient of the correlation of the population, and n is again size of the sample that has been used to determine the coefficient of correlation.

The point estimate of the population coefficient of correlation ρ is equal to the sample coefficient of correlation r. The interval estimate of ρ can be made using the transformation (11.21). The two sided interval estimate can be expressed as

$$z + u_p\, \sigma_z < \mu_z < z + u_{1-p}\, \sigma_z \tag{11.24}$$

This interval covers the unknown mean μ_z with the confidence level (probability coefficient) $1-2p$; u_p and u_{1-p} are the standardized normal variables corresponding to the probabilities indicated in the subscripts. Taking into account the standard deviation σ_z given by Eq. (11.23), the interval estimate given by Eq. (11.24) can be written as

$$z + \frac{u_p}{\sqrt{n-3}} < \mu_z < z + \frac{u_{1-p}}{\sqrt{n-3}} \tag{11.25}$$

The whole estimation procedure for the population coefficient of correlation ρ can be summarised by the following main steps.

- Transformation of r into the variable z.
- Determination of values u_p and u_{1-p} for specified confidence level $1-2p$.
- Determination of the interval estimate of μ_z using Eq. (11.25).
- Recalculation of the population using transformation (11.22).

Example 11.3. Consider a sample of $n = 16$ observations having the sample coefficient of correlation $r = 0.73$. The interval estimate for the population coefficient of correlation should be determined by the confidence level $1-2p = 0.95$. The above mentioned estimate procedure yields

- For $r = 0.73$ the transformed variable z $= 0.9287$
- For $p = 0.025$ the standardised variables are $u_p = -1.96$ and $u_{1-p} = 1.96$
- The interval estimate for the variable z

$$0.9287 - \frac{1.96}{\sqrt{16-3}} < \mu_z < 0.9287 + \frac{1.96}{\sqrt{16-3}}$$

$$0.3851 < \mu_z < 1.4723$$

- The interval estimate of the population correlation coefficient ρ

$$0.37 < \rho < 0.90$$

11.6 Estimation of the Coefficients of Regression

The population coefficients of regression $\alpha_0, \alpha_1, \beta_0, \beta_1$, corresponding to the sample coefficients a_0, a_1, b_0, b_1, are to be estimated. It is assumed that a sample of n observations is described by the characteristics m_X, m_Y, s_X, s_Y, and the sample coefficients of regression a_0, a_1, b_0, b_1.

It can be shown that the sampling distributions of regression coefficients are normal, with the mean and standard deviation of the coefficient a_0 as

$$\mu_{a_0} = \alpha_0; \quad \sigma_{a_0} = \frac{\sigma_Y}{\sqrt{n}} \sqrt{1 + \frac{m_X^2}{s_X^2}} \tag{11.26}$$

The mean and standard deviation of the coefficient a_1 are

$$\mu_{a_1} = \alpha_1; \quad \sigma_{a_1} = \frac{\sigma_Y}{s_X \sqrt{n}} \tag{11.27}$$

Similar expressions hold for the regression coefficients b_0 and b_1.

Taking into account Eqs. (11.26) and (11.27), it follows that the point estimates of the population regression coefficients α_0, α_1, β_0, β_1 are equal to the sample coefficients a_0, a_1, b_0, b_1.

The interval estimates covering the unknown population coefficients with the confidence level (probability) $1-2p$ are given by similar expressions to those given in Chap. 8. For the regression coefficient α_0 it holds that

$$a_0 + u_p Df_{a_0} < \alpha_0 \quad < a_0 + u_{1-p} Df_{a_0} \tag{11.28}$$

Substituting the standard deviation Df_{a_0} with the expression given in Eq. (11.26) it follows that

$$a_0 + u_p \frac{Df_Y}{\sqrt{n}} \sqrt{1 + \frac{m_X^2}{s_X^2}} < \alpha_0 \quad < a_0 + u_{1-p} \frac{Df_Y}{\sqrt{n}} \sqrt{1 + \frac{m_X^2}{s_X^2}} \tag{11.29}$$

The population standard deviation σ_Y is generally unknown, and must be assessed by using the estimate of residual variance $s_{Y|X}^2$ about the regression line with the expression

$$s_{Y|X}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 \tag{11.30}$$

The residual variance $s^2_{Y|X}$ given by Eq. (11.30) represents a point estimate of the variance as a sum of residuum squares divided by the degree of freedom to $n-2$. Then the interval estimate of the population coefficient α_0 (11.28) may be written as

$$a_0 + t_p \frac{Df_{Y|X}}{\sqrt{n}} \sqrt{1 + \frac{m_X^2}{s_X^2}} < \alpha_0 \quad < a_0 + t_{1-p} \frac{Df_{Y|X}}{\sqrt{n}} \sqrt{1 + \frac{m_X^2}{s_X^2}} \tag{11.31}$$

Due to the assessment of the population standard deviation by residual standard deviation, t random variable is used instead of the standardised normal u variable.

Similarly the interval estimate of the population coefficient α_1 is given as

$$a_1 + \frac{u_p}{\sqrt{n}} \frac{Df_Y}{s_X} < \alpha_1 < a_1 + \frac{u_{1-p}}{\sqrt{n}} \frac{Df_Y}{s_X} \tag{11.32}$$

Again, the assessment of population variance using the residual variance Eq. (11.31) can be written as

$$a_1 + \frac{t_p}{\sqrt{n}} \frac{Df_{Y|X}}{s_X} < \alpha_1 < a_1 + \frac{t_{1-p}}{\sqrt{n}} \frac{Df_{Y|X}}{s_X} \tag{11.33}$$

The resulting inequalities Eqs. (11.31) and (11.33) can also be written for the population coefficients of regression β_0 and β_1.

Example 11.4. A sample of $n = 16$ measurements of cubic strength Y and cylindrical strength X of concrete yields the regression coefficients $a_0 = 0.39$ and $a_1 = 1.03$ that are considered as the point estimates of the population coefficients α_0 and α_1. To determine the interval estimates the following quantities are evaluated:

$$m_X = 14.3 \text{ MPa}; \quad s_X = 2.96 \text{ MPa}; \quad s_{Y|X} = 1.92 \text{ MPa}$$

For specified confidence level 0.95 ($p = 0.025$), and the degree of freedom to $n-2 = 14$, the value $t_p = -2.145$ and $t_{1-p} = 2.145$. The interval estimates are

$$0.39 - 2.145 \frac{1.92}{\sqrt{16}} \sqrt{1 + \frac{14.3^2}{2.96^2}} < \alpha_0 < 0.39 + 2.145 \frac{1.92}{\sqrt{16}} \sqrt{1 + \frac{14.3^2}{2.96^2}}$$

$$1.03 - \frac{2.145}{\sqrt{16}} \frac{1.92}{2.96} < \alpha_1 < 1.03 - \frac{2.145}{\sqrt{16}} \frac{1.92}{2.96}$$

Thus the resulting interval estimates are

$$-4.69 < \alpha_0 < 5.47$$

$$0.68 < \alpha_0 < 1.38$$

11.7 Boundaries of the Regression Line

It is interesting to know that the region where the population regression line $Y = \alpha_0 + \alpha_1 X$ may occur with a given confidence level (probability) $1-2p$. It can be derived from the above estimates of the population regression coefficients α_0 and α_1 that the location of the regression line is limited by the lower and upper limits given by the following inequality

$$a_0 + a_1 X + t_p \frac{Df_{Y|X}}{\sqrt{n}} \sqrt{1 + \frac{(X - m_X)^2}{s_X^2}} < a_0 + \\ < a_0 + a_1 X + t_{1-p} \frac{Df_{Y|X}}{\sqrt{n}} \sqrt{1 + \frac{(X - m_X)^2}{s_X^2}} \tag{11.34}$$

The variable t is again taken from t distribution for $n-2$ degree of freedom.

Example 11.5. Considering Example 11.4 Eq. (11.34) may be written as

$$0.39 + 1.03_1 X - 2.145 \frac{1.92}{\sqrt{16}} \sqrt{1 + \frac{(X - 14.3)^2}{2.96^2}} < \alpha_0 + \alpha_1 X < 0.39 + 1.03_1 X \\ + 2.145 \frac{1.92}{\sqrt{16}} \sqrt{1 + \frac{(X - 14.3)^2}{2.96^2}}$$

The graphical representation of this inequality is shown in Fig. 11.3. The curves denoted as the lower and upper limits represent the limits of the area where the population regression line is expected with a confidence level (probability) of 0.95. There is only a small probability 0.025 that the regression line will be above or below the designated area. Obviously, with increasing confidence level, the expected region of the population regression line is getting larger.

11.8 Tests of Correlation Coefficient

There are two frequently used tests concerning the sample coefficient of correlation r to verify the hypothesis that:

- A sample is taken from a population having the coefficient of correlation ρ,
- Two samples are taken from populations having the same coefficient of correlation ρ.

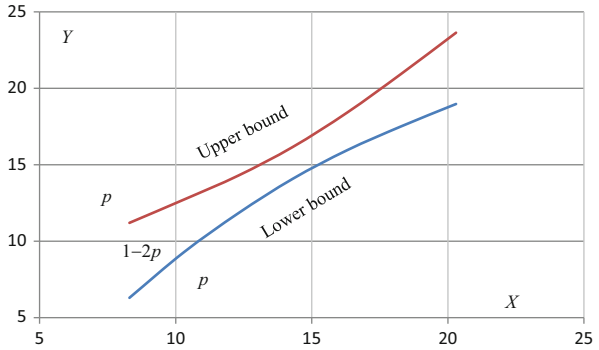


Fig. 11.3 Boundaries of the population correlation line

In the first case the difference between the sample coefficient r and the population coefficient of correction ρ is assessed, considering a given significance level α (0.01 or 0.05). The tested variable is defined as

$$u = (z - \mu_z)\sqrt{n-3} \tag{11.35}$$

Here, the variable z is given by the transformation (11.21) of the sample coefficient r, and μ_z.by the same transformation, but for the population coefficient ρ. The test value u_0 is compared with the critical values taken from the standardised normal variable $u_{p1} = u_{\alpha/2} = -1.96$ and $u_{p2} = u_{1-\alpha/2} = 1.96$ for the significance level $\alpha = 0.05$ or $u_{p1} = u_{\alpha/2} = -2.58$ and $u_{p2} = u_{1-\alpha/2} = 2.58$ for the significance level $\alpha = 0.01$.

Example 11.6. The coefficient of the correlation 0.34 of the wind speed at 2 and 10 m aboveground level is evaluated from a sample of $n = 1{,}187$ observation. The hypothesis that the sample is taken from a population having the coefficient of correlation $\rho = 0.4$ is to be tested. The test value follows from Eq. (11.35) as

$$u_0 = (0.3541 - 0.4236)\sqrt{1187 - 3} = 2.391$$

The critical value $u_{p2} = u_{1-\alpha/2} = 2.58$ is greater than the test value and the hypothesis is accepted.

In the second case the tested variable is defined as

$$u = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \tag{11.36}$$

Here again the transformation (11.21) is used for sample coefficients of correlations r_1 and r_2 to get the variables z_1 and z_2. The critical values are the

same as in the first case: $u_{p1} = u_{\alpha/2} = -1.96$ and $u_{p2} = u_{1-\alpha/2} = 1.96$ for the significance level $\alpha = 0.05$ or $u_{p1} = u_{\alpha/2} = -2.58$ and $u_{p2} = u_{1-\alpha/2} = 2.58$ for the significance level $\alpha = 0.01$.

Example 11.7. Two samples similar to that in Example 11.6 are available:

- $n_1 = 1{,}187$, $r_1 = 0.34$
- $n_2 = 956$, $r_2 = 0.40$

The test value follows from Eq. (11.36) as

$$u_0 = \frac{0.3541 - 0.4236}{\sqrt{\frac{1}{1187-3} + \frac{1}{956-3}}} = -1.597$$

Assuming the significance level $\alpha = 0.05$, the critical value $u_{p1} = u_{\alpha/2} = -1.96$ and the difference of the two coefficients of correlation is insignificant.

11.9 Tests of Regression Coefficients

As mentioned previously, in Sect. 11.8, there are two types of tests concerning the coefficients of regression a_0 and a_1 and the population coefficients α_0 and α_1 (regression of Y on X, see Sect. 11.4). The following hypotheses are to be tested:

- A sample having the regression coefficients a_0 and a_1 is taken from a population with the coefficients α_0 and α_1
- Two samples are taken form populations with the same coefficients of regression.

Analogous test are applied for the coefficients b_0 and b_1 and β_0 and β_1 for the regression of variable X on Y. The following procedures concern only regression of Y on X and the first type of the tests.

Assuming that the population standard deviation σ_Y is known, the tested variable of the first type of the tests is given as

$$u_0^0 = \frac{(a_0 - \alpha_0)\sqrt{n}}{\sigma_Y \sqrt{1 + \frac{m_X^2}{s_x^2}}} \tag{11.37}$$

$$u_0^1 = \frac{(a_1 - \alpha_1)s_X\sqrt{n}}{\sigma_Y} \tag{11.38}$$

The critical values are taken from the standardised normal variable $u_{p1} = u_{\alpha/2} = -1.96$ and $u_{p2} = u_{1-\alpha/2} = 1.96$ for the significance level $\alpha = 0.05$ or $u_{p1} = u_{\alpha/2} = -2.58$ and $u_{p2} = u_{1\ -\ \alpha/2} = 2.58$ for the significance level $\alpha = 0.01$.

If the standard deviation σ_Y is unknown, then it must be substituted by the residual standard deviation $s_{Y|X,}$ and the tested variable of the first type of tests becomes

$$t_0^0 = \frac{(a_0 - \alpha_0)\sqrt{n}}{s_{Y|X}\sqrt{1 + \frac{m_X^2}{s_X^2}}} \tag{11.39}$$

$$t_0^1 = \frac{(a_1 - \alpha_1)s_X\sqrt{n}}{s_{Y|X}} \tag{11.40}$$

In this case, the critical values are taken from the t distribution $t_{p1} = t_{\alpha/2}$ and $t_{p2} = t_{1-\alpha/2}$ for the significance level $\alpha = 0.05$ or 0.01 and $\nu = n-2$ for the degree of freedom.

Example 11.8. Consider again the sample mentioned in Example 11.6: the sample size is $n = 1{,}187$, the coefficient of regression $a_0 = 5.4\ \mathrm{ms}^{-1}$ and $a_1 = 1.85$, the residual standard deviation $s_{Y|X.} = 3.66\ \mathrm{ms}^{-1}$. The characteristics of the independent random variable X are $m_X = 13.2\ \mathrm{ms}^{-1}$ and $s_X = 2.12\ \mathrm{ms}^{-1}$. The hypothesis that the population coefficients of regression are $\alpha_0 = 4\ \mathrm{ms}^{-1}$ and $\alpha_1 = 2$ is to be tested.

The test value follows from Eqs. (11.39) and (11.40) as

$$t_0^0 = \frac{(5.40 - 4.0)\sqrt{n}}{3.66\sqrt{1 + \frac{13.2^2}{2.12^2}}} = 2.09$$

$$t_0^1 = \frac{(1.85 - 2.00)12\sqrt{1187}}{3.66} = -2.99$$

Considering the significance level $\alpha = 0.05$ and the number of the degree of freedom $\nu = n - 2 = 1{,}185$, the critical values may be taken from a standardised normal variable as $u_{p1} = u_{\alpha/2} = -1.96$, and $u_{p2} = u_{1-\alpha/2} = 1.96$. The difference of the sample coefficients from the population coefficients is significant, and the hypothesis rejected.

References

1. Ang, A.H.-S., Tang, W.H.: Probabilistic Concepts in Engineering. Emphasis on Applications to Civil Environmental Engineering. Wiley, New York (2007)
2. Devore, J., Farnum, N.: Applied Statistics for Engineers and Scientists. Thomson, London (2005)
3. ISO 12491: Statistical Methods for Quality Control of Building Materials and Components. ISO, Geneve (1997)
4. Holicky, M.: Reliability Analysis for Structural Design. SUNN MeDIA, Stellenbosch (2009)

Chapter 12
Random Functions

Random functions and random fields are applied in current engineering and scientific tasks more and more frequently. Random functions are actually random variables that are functions of deterministic arguments, for example of time, planar or spatial coordinates. A brief introduction to random functions includes a definition of the basic parameters such as the mean, variance and autocorrelation function. The definition of the stationary and ergodic functions is supplemented by a description of the spectral representation of stationary random functions. The fundamental properties of random functions and operations with commonly used random functions are illustrated by practical examples.

12.1 Basic Concepts

A random function or a random field $X(t)$ of a deterministic argument t (spatial or planar coordinates, time) is a function delivering for a given argument t a random variable $X = X(t)$ [1–4]. In agreement with previously used symbols random functions will be denoted by capitals, such as $X(t)$, $Y(t)$, $Z(t)\ldots$, their individual realizations by the lower case letters $x(t)$, $y(t)$, $z(t)\ldots$. If the fixed values of the argument are denoted by t_1, t_2, $t_3\ldots$ then the corresponding random variables represent a system of different random variables

$$X(t_1), X(t_2), X(t_3) \ \ldots \tag{12.1}$$

If the number of arguments t_i, $i = 1,2,..,m$ increases then the system of random variables (12.1) describes sufficiently well the random function $X(t)$. Thus, the random function is a generalisation of the system of random variables.

Example 12.1. The inside temperature of a structure depends on a number of circumstances difficult to describe completely. Consequently the temperature is considered as a random function $X(t)$ of the deterministic argument t representing

M. Holický, *Introduction to Probability and Statistics for Engineers*,
DOI 10.1007/978-3-642-38300-7_12, © Springer-Verlag Berlin Heidelberg 2013

the spatial coordinates and time. Similarly, wind speed may be considered as a random function of the coordinates and time.

12.2 Parameters of Random Function

The mean $m_X(t)$ of a random function $X(t)$ is defined analogously to the mean of random variables described in Sect. 4.5. Taking into account the fact that the probability density $\varphi_X(x,t)$ of a random function $X(t)$ is generally dependent on two variables, on a point x of variable $X = X(t)$ and on the deterministic argument t, then the mean $\mu_X(t)$ is a generalisation of the definition (4.11)

$$\mu_X(t) = \int_{-\infty}^{\infty} x\varphi_X(x,t)\mathrm{d}x \tag{12.2}$$

The mean $\mu_X(t)$ is a deterministic function of t that represents the basic component of the random function $X(t)$ describing its location (or central tendency). The difference $X(t) - \mu_X(t)$ is called the fluctuation component of the function $X(t)$. This fluctuation component is used to define the variance $\sigma_X^2(t)$ of the random function $X(t)$ in a similar way to defining the variance σ_X^2 of a random variable X given by Eq. (4.14). Thus the variance $\sigma_X^2(t)$ is defined as

$$\sigma_X^2(t) = \int_{-\infty}^{\infty} [X(t) - \mu_X(t)]^2 \varphi_X(x,t)\mathrm{d}x \tag{12.3}$$

The variance $\sigma_X^2(t)$ is a measure of dispersion of the random function $X(t)$ around the mean $\mu_X(t)$ as indicated in Fig. 12.1.

Instead of the variance $\sigma_X^2(t)$ the standard deviation $\sigma_X(t)$ is often used in technical applications. Following relationship (4.16) the standard deviation is defined as the square root of the variance

$$\sigma_X(t) = \sqrt{\sigma_X^2(t)} \tag{12.4}$$

It is interesting to note that although the realization $x(t)$ and the mean $\mu_X(t)$ of the random function $X(t)$ in Fig. 12.1 have a similar shape, the character of the random function $Y(t)$ in Fig. 12.2 is different. Nevertheless, both random functions $X(t)$ and $Y(t)$ have the same variance. If, for example, for the given argument t_1 a realisation $x(t_1)$ of the function $X(t)$ is above the mean $\mu_X(t_1)$, then it is very likely that also the value $x(t_2)$ of the same realisation at a nearby point t_2 will also be above the mean $\mu_X(t_2)$. This cannot be said about the random function $Y(t)$ shown in Fig. 12.2. So, it can be intuitively stated that values of the random function $X(t)$ have a greater degree of mutual correlation than the values of the random function $Y(t)$.

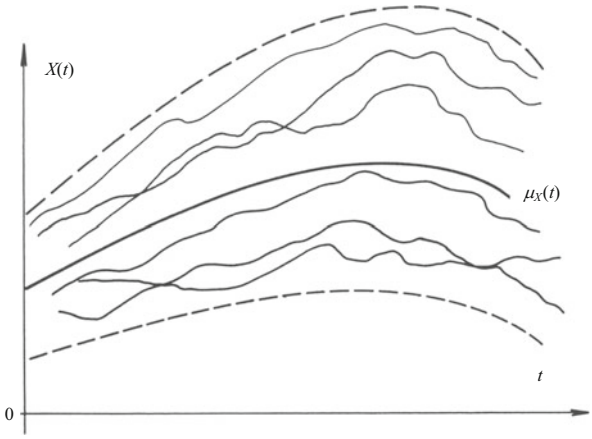


Fig. 12.1 Random function $X(t)$

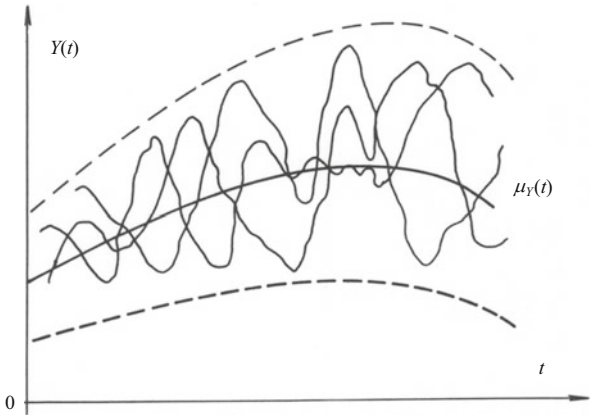


Fig. 12.2 Random function $Y(t)$

Note that the subscripts X,Y indicating the random functions $X(t)$ and $Y(t)$ may be omitted in unambiguous cases.

12.3 Correlation Function

The mutual dependence of two random variables $X(t_1)$ and $X(t_2)$ can be expressed by the coefficient of correlation discussed in detail in Chap. 11. However, the random variables $X(t_1)$ and $X(t_2)$ depend on deterministic arguments t_1 and t_2 Consequently, the covariance (11.7) becomes a function of t_1 and t_2 and is commonly called the correlation function, sometimes the auto-correlation function (even though strictly speaking it should be called the covariance function, this being the name used in literature). Following the general principles of Chap. 11, the correlation function $K_X(t_1,t_2)$ is defined as

$$K_X(t_1,t_2)=\int\int_{-\infty}^{\infty}\Big(x_1-\mu_X(t_1)\Big)\Big(x_2-\mu_X(t_2)\Big)\varphi_2(x_1,x_2,t_1,t_2)\mathrm{d}x_1\mathrm{d}x_2 \tag{12.5}$$

Here, $\varphi_2(x_1,x_2,t_1,t_2)$ denotes two dimensional probability density function, as used previously in Chap. 11.

The correlation function $K_X(t_1,t_2)$ defined by Eq. (12.5) has several important properties that are essential from a practical point of view. Firstly, it follows directly from the definition (12.5) that for an identical argument $t_1 = t_2 = t$ the correlation function $K_X(t_1,t_2)$ becomes the variance $\sigma_X^2(t)$

$$K_X(t,t)=\sigma_X^2(t) \tag{12.6}$$

So, the correlation function $K_X(t_1,t_2)$ automatically includes information on the variance $\sigma_X^2(t)$ that does not need to be specified separately.

There are another three basic properties of the correlation function $K_X(t_1,t_2)$ that can be derived from the definition (12.5) indicated in Fig. 12.3:

1. The correlation function $K_X(t_1,t_2)$ is symmetrical with respect to t_1 and t_2

$$K_X(t_1,t_2)=K_X(t_2,t_1) \tag{12.7}$$

2. It follows from the definition (12.5) that

$$K_X(t_1,t_2)\le\sigma_X(t_1)\sigma_X(t_2) \tag{12.8}$$

3. The correlation function of a random function $X(t)$ and the sum $Y(t)$ of function $X(t)$ and a deterministic function $\xi(t)$, $Y(t) = X(t) + \xi(t)$ is the same

$$K_X(t_1,t_2)=K_Y(t_1,t_2) \tag{12.9}$$

Instead of the correlation function $K_X(t_1,t_2)$ so-called normalised correlation function (corresponding to the coefficient of correlation defined for two random variables (11.8)) is used

$$R_X(t_1,t_2)=\frac{K_X(t_1,t_2)}{\sigma_X(t_1)\sigma_X(t_1)} \tag{12.10}$$

It follows from the property (12.8) that the normalised correlation function (12.10) is a function of two variables, and that its values are within the interval $<-1, 1>$ similar to the values of the coefficient of correlation (12.8).

When two random functions, say $X(t)$ and $Y(s)$, are involved together then mutual correlation function may be needed

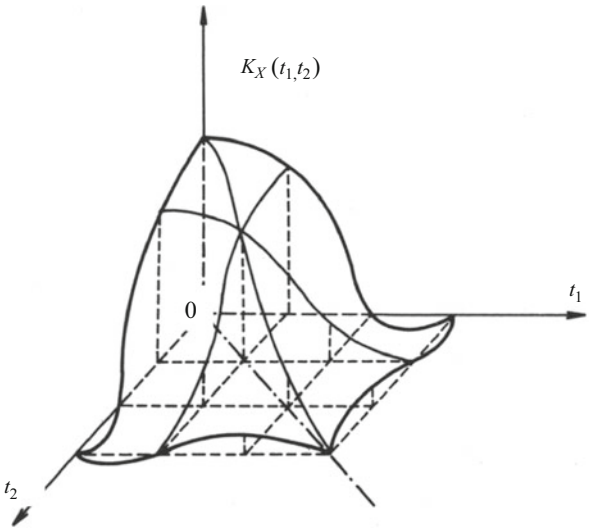


Fig. 12.3 Correlation function $K_X(t_1,t_2)$

$$K_{XY}(t,s) = \int\int_{-\infty}^{\infty} \Big(x - \mu_X(t)\Big)\Big(y - \mu_Y(s)\Big)\varphi_2(x,y,t,s)\mathrm{d}x\mathrm{d}y \tag{12.11}$$

The random functions $X(t)$ and $Y(s)$ are said to be mutually dependent when the correlation function $K_{XY}(t,s)$ is not exactly equal zero. Together with the correlation function $K_{XY}(t,s)$ the normalised correlation function $R_{XY}(t,s)$ is used

$$R_{XY}(t,s) = \frac{K_{XY}(t,\mathrm{s})}{\sigma_X(t)\sigma_Y(\mathrm{s})} \tag{12.12}$$

Example 12.2. Consider a random function describing harmonic vibration

$$X(t) = A\cos(\omega t + \varepsilon)$$

Here $A \geqq 0$ and ε are random variables, $\omega \geqq 0$ is assumed to be a constant. If the phase shift ε is independent of A and ε has a uniform distribution within the interval $<0, 2\pi>$ then the random function is fully described by the joint distribution

$$\varphi(A,\varepsilon) = \frac{1}{2\pi}\varphi(A); 0 \leq \varepsilon \leq 2\pi$$

The mean of the random function follows from Eq. (12.2) as

$$\mu_X(t) = \int\int_{-\infty,0}^{\infty,2\pi} \frac{1}{2\pi}\varphi(A)A\cos(\omega t + \varepsilon)\mathrm{d}A\mathrm{d}\varepsilon = 0$$

The correlation function may be obtained from the definition (12.5) in the form

$$K_X(t_1, t_2) = \frac{1}{2}\mu'_{2,A} \cos[\omega\,(t_1 - t_2)]$$

The variance is obtained from Eq. (12.6)

$$K_X(t, t) = \sigma_X^2(t) = \frac{1}{2}\mu'_{2,A}$$

Finally, the normalised correlation function follows Eq. (12.10)

$$R_X(t_1, t_2) = \frac{K_X(t_1, t_2)}{\sigma_X(t_1)\sigma_X(t_2)} = \cos[\omega\,(t_1 - t_2)]$$

Note that with increasing $|t_1 - t_2| \to \infty$ the correlation function does not converge at zero but always fluctuates within the interval $<-1, 1>$.

12.4 Stationary Random Functions

The stationary random functions represent an important group of random functions by which many mathematical operations may be substantially simplified. That is why non-stationary functions are often transformed into stationary functions.

A random function is stationary if its mean $\mu_X(t)$ is constant (and may be denoted simply as μ_X), and if the correlation function $K_X(t_1,t_2)$ depends on the difference $t_1 - t_2 = \tau$ thus

$$\mu_X(t) = \mu_X; \quad K_X(t_1, t_2) = K_X(t_1 - t_2) = K_X(\tau) \tag{12.13}$$

The correlation function $K_X(\tau)$ depends now on one variable τ only. Consequently it follows from Eq. (12.6) that the variance is constant

$$K_X(0) = \sigma_X^2 \tag{12.14}$$

Equations (12.7) and (12.8) can now be stationary functions simplified as follows:

– The correlation function $K_X(\tau)$ is even function as

$$K_X(-\tau) = K_X(\tau) \tag{12.15}$$

– It follows from the definition (12.8) that

$$K_X(\tau) \leq \sigma_X^2 \tag{12.16}$$

The normalised correlation function $R(\tau)$ of a stationary random function $X(t)$ follows from Eq. (12.10) as

$$R_X(\tau) = \frac{K_X(\tau)}{\sigma_X^2} \tag{12.17}$$

Example 12.3. Consider a random function with a simple correlation function (subscript indicating the random function is omitted)

$$K(\tau) = \sigma^2 \exp(-c|\tau|)$$

The correlation function $K(\tau) = \sigma^2 \exp(-c|\tau|)$ is shown in Fig. 12.4 for selected values of the constant c.

Note that for c approaching 0, the correlation function becomes constant (as indicated in Fig. 12.4 by the dashed line). In that case, the values of the random functions at two different points are completely dependent and the function is reduced to the random variable. With an increasing constant c, the correlation function $K(\tau) = \sigma^2 \exp(-c|\tau|)$ approaches the horizontal axes τ (except $\tau = 0$ where $K(\tau) = \sigma^2$). This is the case when the values of the random function at two different points are mutually independent (their correlation is zero).

Example 12.4. Another commonly used correlation function (the subscript indicating the random function is omitted) of stationary random function is

$$K(\tau) = \sigma^2 \exp(-c_0|\tau|) \cos(c_1\tau)$$

Here $c_0 \geqq 0$ and $c_1 \geqq 0$ are constants determining the shape of the correlation function. The function is shown in Fig. 12.5. While the correlation function described in the previous Example 12.2 attains only positive values, the above correlation function fluctuates periodically from positive to negative values. For $c_1 = 0$ both the correlation functions are identical.

12.5 Ergodic Random Functions

Most of the stationary random functions comply with another important property in that – they are ergodic. A random function is said to be ergodic if all required properties of the function may be deduced from one realisation in a sufficiently large interval of the argument t. In mathematical formulation (subscript X is

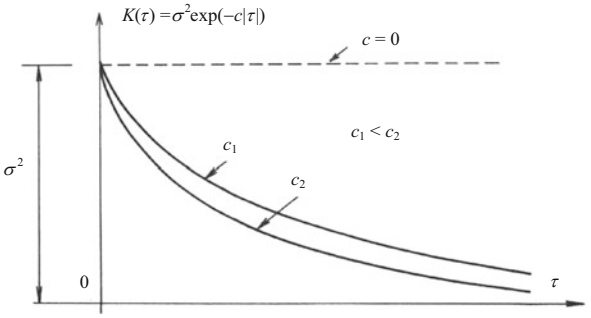


Fig. 12.4 Correlation function $K(\tau) = \sigma^2 \exp(-c|\tau|)$ of a stationary random function

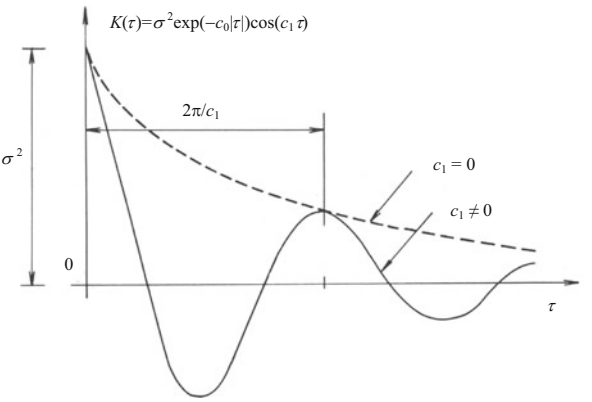


Fig. 12.5 Correlation function $K(\tau) = \sigma^2 \exp(-c_0|\tau|)\cos(c_1\tau)$ of a stationary random function

omitted) a function is said to be ergodic if the following relationships concerning the mean μ and correlation function $K(\tau)$ are valid.

$$\mu_X = \int_{-\infty}^{\infty} x\varphi(x,t)\,\mathrm{d}t = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)\mathrm{d}t \tag{12.18}$$

$$\begin{aligned} K_X(\tau) &= \int\int_{-\infty}^{\infty} (x_1-\mu)(x_2-\mu)\varphi_2(x_1,x_2,t_1,t_2)\mathrm{d}x_1\mathrm{d}x_2 \\ &= \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} [x(t)-\mu][x(t+\tau)-\mu]\mathrm{d}t \end{aligned} \tag{12.19}$$

It is well to note that there are stationary random functions which are not ergodic. Figure 12.6 shows such an example.

Example 12.5. A random function $Y(t)$ is given by a sum of the stationary ergodic function $X(t)$ and the random variable Z (mutually independent).

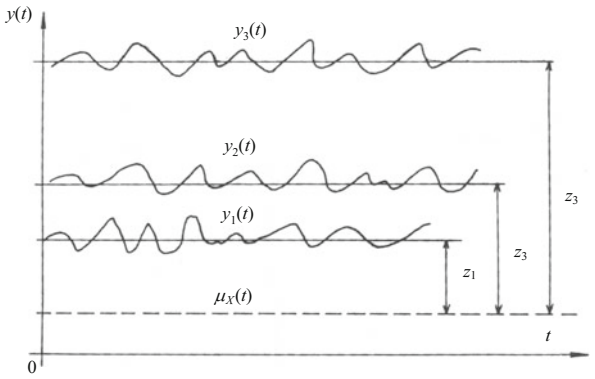


Fig. 12.6 An example of a stationary random function which is not ergodic

$$Y(t) = X(t) + Z$$

Three realizations of $Y(t)$ are shown in Fig. 12.6. The mean and correlation function of $Y(t)$ are

$$\mu_Y(t) = \mu_X(t) + \mu_z$$

$$K_Y(\tau) = K_X(\tau) + \sigma_Z^2$$

The mean $\mu_Y(t)$ depends on a particular realization of Z (there are three realizations z_1, z_2 and z_3 as shown in Fig. 12.5) and cannot be determined from one realization $y_i(t)$ of $Y(t)$. Thus $Y(t)$ is an example of stationary but not ergodic random function.

It should be mentioned that the definition of ergodic random function is not unified. The ergodic properties are sometimes required only with respect to the mean or the correlation function (not with respect to both).

12.6 Spectral Representation of Random Functions

The spectral representation of random function is commonly understood as its expression in the form of a sum of harmonic functions having different amplitudes and frequencies. It's a special case of canonic decomposition of random functions when the coordinate functions are

$$\cos\ \omega_k t, \sin\ \omega_k t,\ k = 0,\ 1,\ 2, \ldots \tag{12.20}$$

Here ω_k denotes the frequencies of the coordinate functions (12.20).

Consider a stationary random function $X(t)$ in the definite interval $<-T, T>$. The corresponding correlation function $K_X(\tau)$ has the definition domain of the argument

$\tau = t_1 - t_2$ the interval $<-2T, 2T>$. It is known that any even function can be expressed using the Fourier sequence with cosine function only:

$$K_X(\tau) = \sum_{k=0}^{\infty} D_k \cos \omega_k \tau, \quad \omega_k = k\frac{\pi}{2T} \tag{12.21}$$

$$D_0 = \frac{1}{2T}\int_0^{2T} K_X(\tau)d\tau \tag{12.22}$$

$$D_k = \frac{1}{T}\int_0^{2T} K_X(\tau)\cos(\omega_k \tau)\mathrm{d}\tau \tag{12.23}$$

This is a canonic representation of the correlation function from which the canonic representation of the of stationary random function follows as

$$X(\tau) = \mu_X + \sum_{k=0}^{\infty} (A_k \cos \omega_k \tau + B_k \cos \omega_k \tau) \tag{12.24}$$

The coefficients A_k and B_k are mutually independent random variables of the zero means and the variance D_k. The expression (12.24) is called the spectral decomposition of the random function $X(t)$.

Substituting $\tau = 0$ into Eq. (12.21) the variance of the random function $X(t)$ follows as

$$\sigma_X^2 = \sum_{k=0}^{\infty} D_k \tag{12.25}$$

Thus the variance of the function is equal to the sum of variance of all the harmonic functions of the spectral decomposition. Equation (12.25) indicates that the variance σ_X^2 can be decomposed into partial variances D_k corresponding to frequencies ω_k. Distribution of the partial variances D_k, with regard to the frequencies ω_k, shown in Fig. 12.7, is called spectrum of the stationary random function.

The above procedure can be extended to the infinite interval $<-\infty, \infty>$ in which the summation is replaced by integrals. In this way the partial variances D_k (12.23) are replaced by the spectral density function $S(\omega)$ of the stationary random function given by the integral

$$S(\omega) = \frac{1}{\pi}\int_{-\infty}^{\infty} K_X(\tau)\cos(\omega\tau)\mathrm{d}\tau \tag{12.26}$$

The spectral density $S(\omega)$ is a continuous even function that generalizes the partial variances D_k given in Eq. (12.23). The correlation function $K_X(\tau)$ can then be expressed on the basis of the spectral density $S(\omega)$ as

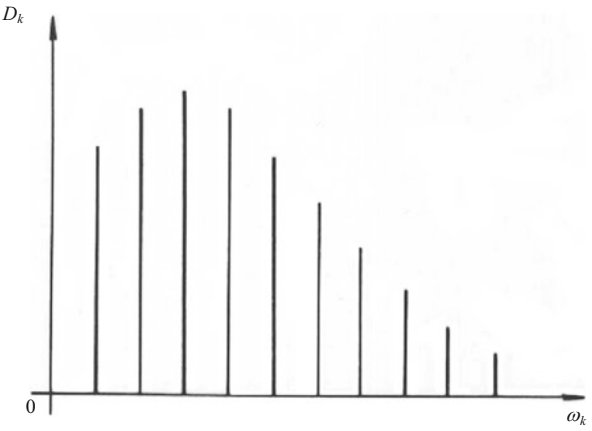


Fig. 12.7 Spectrum of a stationary random function

$$K_X(\tau) = \frac{1}{2}\int_{-\infty}^{\infty} S(\omega)\cos(\omega\tau)\mathrm{d}\omega \tag{12.27}$$

Equations (12.26) and (12.27) represent formulae of the Fourier transformation that are commonly written in a complex form (in the following formulae $\mathrm{i} = \sqrt{-1}$) as

$$S(\omega) = \frac{1}{\pi}\int_{-\infty}^{\infty} K_X(\tau)\exp(-i\omega\tau)\mathrm{d}\tau \tag{12.28}$$

$$K_X(\tau) = \frac{1}{2}\int_{-\infty}^{\infty} S(\omega)\exp(-i\omega\tau)\mathrm{d}\omega \tag{12.29}$$

The variance (12.25) can be now given in an integral form

$$\sigma_X^2 = K_X(0) = \int_0^{\infty} S(\omega)\mathrm{d}\omega \tag{12.30}$$

Thus the integral (12.28) of the spectral density $S(\omega)$ is equal to the variance of the stationary random function. This finding is frequently used in practical assessment of the variance of stationary random function.

Example 12.6. A random function has the correlation function (subscript indicating the random function is omitted)

$$K(\tau) = \sigma^2(1 - |\tau|),\ |\tau| \le 1$$

The spectral density S(ω) follows from Eq. (12.26)

$$S(\omega) = \frac{\sigma^2}{\pi}\int_{-\infty}^{\infty}(1 - |\tau|)\ \cos(\omega\tau)\ \mathrm{d}\tau = \frac{4\sigma^2}{\pi\omega^2}\sin^2\left(\frac{\omega}{2}\right)$$

References

1. Gurskij, E.I.: Theory of Probability and Elements of Mathematical Statistics. Higher School, Moscow (1971) (in Russian)
2. Krishnana, V.: Probability and Random Processes. Wiley, New York (2005)
3. Sveshnikov, A.A.: Applied Methods of the Random Functions. Sudpromgiz, Leningrad (1961) (in Russian)
4. Sveshnikov, A.A.: Collection of Problems on the Theory of Probability, Mathematical Statistics and the Theory of Random Functions. Nauka, Moscow (1965) (in Russian)

Appendix 1: Sample Characteristics and Population Parameters

M. Holický, *Introduction to Probability and Statistics for Engineers*,
DOI 10.1007/978-3-642-38300-7, © Springer-Verlag Berlin Heidelberg 2013

Characteristics, parameters	Sample characteristics m, s^2, a	Population estimates and their standard deviations $\hat{m}$, $\hat{s}^2$, $\hat{a}$	Population parameters μ, σ^2, α
Mean	$m = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\mu_m = \hat{m} = m = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\mu = \int_{-\infty}^{\infty} x\varphi(x)\mathrm{d}x$
		$\sigma_{\hat{m}} = \frac{\sigma}{\sqrt{n}}$	
Variance	$s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - m)^2$	$\mu_{s^2} = \hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - m)^2 = \frac{n}{n-1}s^2$	$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2\varphi(x)\mathrm{d}x$
		$\sigma_{\hat{s}^2} = \frac{\sigma^2}{n}\sqrt{2(n-1)}$	
Skewness	$a = \frac{1}{ns^3}\sum_{i=1}^{n} (x_i - m)^3$	$\mu_a = \hat{a} = \frac{n}{(n-1)(n-2)\hat{s}^3}\sum_{i=1}^{n} (x_i - m)^3 = \frac{\sqrt{n(n-1)}}{(n-2)}a$	$\alpha = \frac{1}{\sigma^3}\int_{-\infty}^{\infty} (x-\mu)^3\varphi(x)\mathrm{d}x$
		$\sigma_{\hat{a}} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}} \approx \sqrt{\frac{6}{n}}$ if $\alpha = 0$	

Appendix 2: Theoretical Models of Discrete Random Variables

M. Holický, *Introduction to Probability and Statistics for Engineers*,
DOI 10.1007/978-3-642-38300-7, © Springer-Verlag Berlin Heidelberg 2013

Distribution, notation	Probability function P(x)	Domain of x	Parameters	Mean μ	Standard deviation σ	Coefficient of variation V	Skewness α
Alternative P(x)	$P(x=1)=p,\ P(x=0)=1-p$	0, 1	p	p	$\sqrt{p(1-p)}$	$\sqrt{\frac{1-p}{p}}$	–
Binomial P(x, n, p)	$P(x,\ n,\ p)=\binom{n}{x}p^x(1-p)^{n-x}$	0,1,2,3 … n	n, p	np	$\sqrt{np(1-p)}$	$\sqrt{\frac{1-p}{np}}$	$\frac{1-2p}{\sqrt{np(1-p)}}$
Hyper-geometric P(x;n;X;N)	$P(x;n;X;N)=\frac{\binom{X}{x}\binom{N-X}{n-x}}{\binom{N}{n}}$	$\max\{0,n+X-N\} < x < \min\{X,n\}$	n, p, N $p = X/N$	np	$\sqrt{np(1-p)\frac{N-n}{N-1}}$	$\sqrt{\frac{1-p}{np}\ \frac{N-n}{N-1}}$	$\frac{(1-2p)\sqrt{(N-1)}(N-2n)}{\sqrt{np(1-p)(N-n)}(N-2)}$
Poisson P(x,λ)	$P(x,\lambda)=\frac{\lambda^x}{x!}e^{-\lambda}$	0,1,2,3 …	λ	λ	$\sqrt{\lambda}$	$\frac{1}{\sqrt{\lambda}}$	$\frac{1}{\sqrt{\lambda}}$
Geometric P(n, p)	$P(n,\ p)=p\ (1-p)^{n-1}$	n	p	$1/p$	$\sqrt{(1-p)}/p$	$\sqrt{(1-p)}$	$\frac{2-p}{\sqrt{1-p}}$

Appendix 3: Theoretical Models of Continuous Random Variables

M. Holický, *Introduction to Probability and Statistics for Engineers*,
DOI 10.1007/978-3-642-38300-7, © Springer-Verlag Berlin Heidelberg 2013

Distribution, notation	Probability density function $\varphi(x)$	Domain of X	Parameters	Mean μ	Standard deviation σ	Skewness α
Rectangular R(a,b)	$1/(b-a)$	$a \le x \le b$	a $b > a$	(a + b)/2	$(b-a)/\sqrt{12}$	0
Normal N(μ,σ)	$\frac{1}{\sigma\sqrt{2\pi}}\ \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$	$-\infty \le x \le \infty$	μ,σ	μ	σ	0
Lognormal, general	$\frac{1}{\lvert x-x_0\rvert\sqrt{\ln(1+c^2)}\sqrt{2\pi}}\exp\left(-\left(\ln\frac{\lvert x-x_0\rvert\lvert c\rvert\sqrt{1+c^2}}{\sigma}\right)/\right.$	$x_0 \le x < \infty$, $\alpha>0$	$x_0 = \mu-\sigma/c$	$x_0 + c\sigma$	σ	$3c + c^3$
LN(μ,σ,α), LN(μ,σ,x_0)	$\left.(2\ln(1+c^2))\right)$	$-\infty < x \le x_0$, $\alpha < 0$	σ, c			
Lognormal, zero lower bound	$\frac{1}{x\sqrt{\ln(1+V^2)}\sqrt{2\pi}}\exp\left(-\left(\ln\frac{x\sqrt{1+V^2}}{\mu}\right)^2/\right.$	$0 \le x < \infty$	μ $V = \sigma/\mu$	μ	$V\,\mu$	$3\,V + V^3$
LN(μ,σ)	$\left.(2\ln(1+V^2))\right)$					
Gamma Gam(μ,σ)	$\lambda^k\,x^{k-1}\exp(-\lambda x)/\Gamma(k)$	$0 \le x < \infty$	$\lambda = \mu/\sigma^2$ $k = (\mu/\sigma)^2$	k/λ	$\sqrt{k}/\lambda$	$2/\sqrt{k}$
Beta, general Beta(μ,σ,α,b) Beta(μ,σ,a,b)	$\frac{(x-a)^{c-1}(b-x)^{d-1}}{B(c,d)(b-a)^{c+d-1}}$	$a \le x \le b$	a, $b > a$, $c \ge 1$, $d \ge 1$	$a + \frac{(b-a)\,c}{c+d}$	$\frac{(b-a)}{cg+dg}$, $g = \sqrt{\frac{c+d+1}{cd}}$	$\frac{2g(d-c)}{c+d+2}$, $g = \sqrt{\frac{c+d+1}{cd}}$
Beta, zero lower bound Beta(μ,σ,α) Beta(μ,σ,b)	$\frac{(x)^{c-1}(b-x)^{d-1}}{B(c,d)\ b^{c+d-1}}$	$0 \le x \le b$	$b > 0$ $c \ge 1$, $d \ge 1$	$\frac{b\ c}{c+d}$	$\frac{b}{cg+dg}$, $g = \sqrt{\frac{c+d+1}{cd}}$	$\frac{2g(d-c)}{c+d+2}$, $g = \sqrt{\frac{c+d+1}{cd}}$
Gumbel, maximum Gum(μ,σ)	$\exp(-\exp(-c(x - x_{\text{mod}})))$	$-\infty \le x < \infty$	$x_{\text{mod}} = \mu-0.577\sigma\sqrt{6}/\pi$ $c = \pi/(\sigma\sqrt{6})$	x_{mod} + 0.577/c	$\pi/(c\sqrt{6})$	1.14

Appendix 4: Functions of Independent Random Variables

M. Holický, *Introduction to Probability and Statistics for Engineers*,
DOI 10.1007/978-3-642-38300-7, © Springer-Verlag Berlin Heidelberg 2013

Function Z	The mean μ_Z	Standard deviation σ_Z	Skewness α_Z
$aX + b$	$a\mu_X + b$	$\lvert a\rvert\sigma_X$	α_X for $a > 0$, $-\alpha_X$ for $a < 0$
X^2 [a]	$\mu_X^2 + \sigma_X^2$	$2\sigma_X\left(\mu_X^2 + \mu_X\sigma_X\alpha_X\right)^{1/2}$	$\frac{8\,\mu_X^3\,\sigma_X^3(\,\alpha_X+3\,V_X)}{\sigma_Z^3}$
$\sqrt{X}$[a]	$\sqrt{\mu_X}$	$\sqrt{\frac{\sigma_X^2}{4\mu_X}}$	$\frac{\alpha_X-1{,}5V_X}{1-0{,}5V_X\alpha_X}$
$\frac{1}{X}$ [a]	$\frac{1+V_X^2-V_X^3\,\alpha_X}{\mu_X}$	$\frac{\left(V_X^2-2V_X^3\alpha_X\right)^{1/2}}{\mu_X}$	$\frac{6\,V_X^4-V_X^3\,\alpha_X}{\mu_X^3\,\sigma_Z^3}$
$aX + bY + c$	$a\mu_X + b\mu_Y + c$	$\left(a^2\sigma_X^2 + b^2\sigma_Y^2\right)^{1/2}$	$\frac{a^3\,\sigma_X^3\,\alpha_X+b^3\,\sigma_Y^3\,\alpha_Y}{\sigma_Z^3}$
$X + Y$	$\mu_X + \mu_Y$	$\left(\sigma_X^2 + \sigma_Y^2\right)^{1/2}$	$\frac{\sigma_X^3\,\alpha_X+\sigma_Y^3\,\alpha_Y}{\sigma_Z^3}$
$X-Y$	$\mu_X-\mu_Y$	$\left(\sigma_X^2 + \sigma_Y^2\right)^{1/2}$	$\frac{\sigma_X^3\,\alpha_X-\sigma_Y^3\,\alpha_Y}{\sigma_Z^3}$
$X\,Y^a$	$\mu_X\,\mu_Y$	$\mu_X\,\mu_Y\,\left(V_X^2 + V_Y^2 + V_X^2\,V_Y^2\right)^{1/2}$	$\frac{\mu_X^3\,\mu_Y^3\,\left(V_X^3\,\alpha_X+V_Y^3\,\alpha_Y+6\,V_X^2\,V_Y^2\right)}{\sigma_Z^3}$
$\frac{X}{Y}$ [a]	$\frac{\mu_X\left(1+V_Y^2-V_Y^3\,\alpha_Y+3V_Y^4+1.5V_Y^4\alpha_Y^2\right)}{\mu_Y}$	$\frac{\mu_X\left(V_X^2+V_Y^2-2\,V_Y^3\,\alpha_Y+8V_Y^4+3V_X^2V_Y^2+4{,}5V_Y^4\alpha_Y^2\right)^{1/2}}{\mu_Y}$	$\frac{\mu_X^3\,\left(V_X^3\,\alpha_X-V_Y^3\,\alpha_Y+6\,V_X^4+6\,V_X^2\,V_Y^2+4{,}5V_Y^4\alpha_Y^2\right)}{\mu_Y^3\,\sigma_Z^3}$

[a]Expressions for parameters of marked functions are approximations only

Appendix 5: Fractiles of Random Variables

M. Holický, *Introduction to Probability and Statistics for Engineers*,
DOI 10.1007/978-3-642-38300-7,

Distribution, notation	Domain of X	Fractile x_p of the theoretical model $P(X \le x_p) = \Phi(x_p) = p$	Estimates by coverage method		Estimates by prediction method	
			σ known	σ unknown	σ known	σ unknown
Rectangular R(a,b)	$a \le x \le b$	$a + p\ (b-a)$	–	–	–	–
Normal N(μ,σ)	$-\infty \le x \le \infty$	$\mu + u_p\,\sigma = \mu\ (1 + u_p\,V)$ u_p from Table 9.1	$m - \kappa_p\,\sigma$ κ_p Table 9.3	$m - k_p\,s$ k_p Table 9.4	$m + u_p(1/n + 1)^{1/2}\,\sigma$ u_p from Table 9.1	$m + t_p(1/n + 1)^{1/2}\,s$ $t_p(1/n + 1)^{1/2}$ from Table 9.4
Lognormal, general	$x_0 \le x < \infty$ pro $\alpha > 0$, $-\infty < x \le x_0$	$\mu - \frac{\sigma}{c}\left(1 - \frac{1}{\sqrt{1+c^2}}\exp\left(\text{sign}(\alpha)u_p\sqrt{\ln(1+c^2)}\right)\right) =$ $= x_0 + \frac{\mu + x_0}{\sqrt{1+c^2}}\exp\left(\text{sign}(\alpha)u_p\sqrt{\ln(1+c^2)}\right)$	$m - \kappa_p\,\sigma$ κ_p not given	$m - k_p\,s$ k_p from Tables 9.5 and 9.6	$m + u_p(1/n + 1)^{1/2}\,\sigma$ u_p from Table 9.2	$m + t_p(1/n + 1)^{1/2}\,s$ t_p from Table 9.7
LN(μ,σ,α)	pro $\alpha < 0$	u_p for normal distribution or $\mu + u_p\,\sigma = \mu\ (1 + u_p\,V)$				
LN(μ,σ,x_0)		u_p for lognormal distribution from Table 9.2				
Lognormal, zero lower bound	$0 \le x < \infty$	$\frac{\mu}{\sqrt{1+V^2}}\exp\left(u_p\sqrt{\ln(1+V^2)}\right) \cong$ $\cong \mu\ \exp\left(u_p \times V\right)$ for $V < 0.2$ u_p for normal distribution or $\mu + u_p\,\sigma = \mu\ (1 + u_p\,V)$	$m - \kappa_p\,\sigma$ κ_p not given	$m - k_p\,s$ k_p from Tables 9.5 and 9.6	$m + u_p(1/n + 1)^{1/2}\,\sigma$ u_p from Table 9.2	$m + t_p(1/n + 1)^{1/2}\,s$ t_p from Table 9.7
LN(μ,σ)		And u_p for lognormal distribution from Table 9.2				
Gumbel Maximum Gum(μ,σ)	$-\infty \le x < \infty$	$x_{\text{mod}} - \frac{1}{c}\ln(-\ln(p)) \cong$ $\cong \mu - (0,45 + 0,78\ln(-\ln(p)))\,\sigma$	Fractile can be estimated using general lognormal distribution as an approximation $m + t_p(1/n + 1)^{1/2}\,s$ t_p estimated using Table 9.7 when σ is unknown			

Appendix 6: Conventional Probabilistic Models

Introduction

Probabilistic models of basic variables used in different reliability studies often deviate from each other. Obviously, the reliability studies based on different probabilistic models may lead to different results, to a greater or lesser degree, and to undesirable discrepancies in recommendations concerning the partial safety factors, combination factors and other elements of reliability. It is the aim of this Appendix to propose conventional models in order to enable an efficient comparison of reliability studies of various structural members made of different materials (steel, concrete, composite). It is intended that this Appendix be used independently of the main text, hence it has been written as a self-contained document with its own references and figures.

The probabilistic models of basic variables presented in this Appendix are intended to be used primarily for calibration procedures expected in the near future in connection with the incorporation of Eurocodes [1, 2, 3, 4] and ISO standard [5] into the national systems of codes. Proposed models are specified considering middle values of action variances, common structural conditions and normal quality control of material properties. Recent documents of JCSS [6, 7], CIB reports [8–11], SAKO report [13] and other references [14–18] are taken into account.

Conventional Models

The following conventional models of basic variables are primarily intended to be used in time-invariant reliability analyses (using Turkstra's combination rule) of simple reinforced concrete and steel members. However, the annual maximum value distribution supplemented by appropriate parameters describing time-variant properties can also be applied in time-variant reliability analysis.

Table 1 includes three fundamental categories of basic variables (actions, material strengths and geometric data), supplemented by uncertainty factors for action

M. Holický, *Introduction to Probability and Statistics for Engineers*,
DOI 10.1007/978-3-642-38300-7, © Springer-Verlag Berlin Heidelberg 2013

Table 1 Conventional models of basic variables for time-invariant reliability analyses

No.	Category of variabl.	Name of basic variables	Sym. X	Dimension	Distrib.	Mean μ_X	St. dev. σ_X	Prob. $\Phi_X(X_k)$	References
1	Actions	Permanent$^+$	G	kN/m^2	N	G_k	0.03–0.10μ_X	0.5	[6, 8]
2		Imposed–5 years	Q	kN/m^2	GU	0.2Q_k	1.1μ_X	0.995	[6, 9]
3		Imposed–50 years	Q	kN/m^2	GU	0.6Q_k	0.35μ_X	0.953	[6, 9]
4		Wind – 1 year	W	kN/m^2	GU	0.3W_k	0.5μ_X	0.999	[6, 10]
5		Wind – 50 years	W	kN/m^2	GU	0.7W_k	0.35μ_X	0.890	[6, 10]
6		Snow – 1 year	S	kN/m^2	GU	0.35 S_k	0.70 μ_X	0.998	[6, 11]
7		Snow –.50 years	S	kN/m^2	GU	1.1 S_k	0.30 μ_X	0.437	[6, 11]
8	Material	Steel yield point	f_y	MPa	LN	$f_{yk} + k\sigma$*	0.07–0.10μ_X	0.02	[6, 13–16]
9		Steel strength	f_u	MPa	LN	$f_{ck} + k\sigma$*	0.05μ_X	–	[6, 13–16]
10	Strengths	Concrete	f_c	MPa	LN	$f_{ck} + k\sigma$*	0.10–0.18μ_X	0.02	[6, 13–16]
11		Reinforcement	f_y	MPa	LN	$f_{yk} + k\sigma$*	30 MPa	0.02	[6, 13–16]
12	Geometry	IPE profiles	A,W,I	m^{2,3,4}	N	0.99X_{nom}	0.01–0.04μ_X	≅0.73	[6, 16]
13	steel sect.	L-section, rods	A,W,I	m^{2,3,4}	N	1.02X_{nom}	0.01–0.02μ_X	≅0.16	[6, 16]
14	Geometry	Cross-section	b, h	m	N	b_k,h_k	0.005–0.01	0.5	[6]
15	concrete	Cover of reinf.	a	m	BET	a_k	0.005–0.015	0.5	[6]
16	cross-sect.	Additional ecc.	e	m	N	0	0.003–0.01	–	[6]
17	Model un-	Load effect factor	θ_E	–	N	1	0.05–0.10	–	[6, 7]
18	certainties	Resistance factor$^+$	θ_R	–	N	1–1.25	0.05–0.20	–	[6, 7]

*If the characteristic value is defined as 5 % fractile then the theoretical value of the factor k is 1.645. However, due to common procedures of quality control, the factor k is usually greater, $k \approx 2$

effects and structural resistance. Note that the data indicated in summary Table 1 represents only reasonable conventional models, which may not be adequate in some specific cases (for example, for the wind load of high-rise buildings).

For the purpose of comparative and calibration studies, the mean values μ_X of all the variables X are related to the characteristic value X_k used in the design calculation. The last column of Table 1 shows the occurrence probability of value X as being smaller than the characteristic value X_k

$$P\{X < X_k\} = \Phi_X(X_k) \tag{1}$$

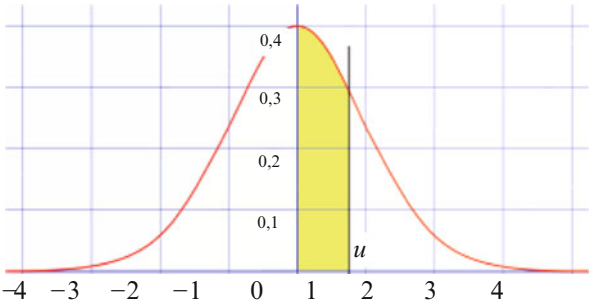
Here Φ_X denotes the distribution function of the basic variable X. Note that due to several reasons (historical development of codified values, quality control of materials) these probabilities in general differ from those recommended for specifications of the characteristic values X_k in the Eurocodes (for example, the actual probability of the material strengths X being less than X_k is only 0.02, rather than the recommended value 0.05, as given in EN 1990 [1]).

References

1. EN 1990: Basis of Structural Design. (Transformation ENV 1991-1, 1994) Brussels (2002)
2. EN 1991-1: Action on Structures. (Transformation ENV 1991-2-1, 1994) Brussels (2002)
3. EN 1991-1-4: Eurocode 1: Actions on Structures – Wind Loads. CEN, Brussels (2004)
4. EN 1991-1-3: Eurocode 1: Actions on Structures – Snow Loads. CEN, Brussels (2004)
5. ISO 2394: General Principles on Reliability for Structures. ISO, Zurich (1998)
6. JCSS, Joint Committee for Structural Reliability: Probabilistic model codes. Working materials. http://www.jcss.ethz.ch/ (2001).
7. Vrouwenvelder, T.: JCSS Probabilistic Model Code. Proc.: Safety, Risk, and Reliability, pp. 65–70. IABSE, Malta (2001)
8. CIB Report: Publication 115, Action on Structures, Self-Weight loads. CIB (1989)
9. CIB Report: Publication 116, Action on Structures, Live Loads in Buildings. CIB (1989)
10. CIB Report: Publication 141, Action on Structures, Snow Load. CIB (1995)
11. CIB Report: Draft of CIB W81 Publication, Action on Structures, Wind Load. CIB (1995)
12. SAKO: Joint Committee of NKB and INSTA-B: Basis of Design of Structures. Proposal for Modification of Partial Safety Factors in Eurocodes (1999)
13. Sorensen, J.D., Hansen, S.O., Nielsen, T.A.: Partial Safety Factors and Target Reliability Level in Danish Codes. Proc.: Safety, Risk, and Reliability, pp. 179–184. IABSE, Malta (2001)
14. Holický, M., Marková, J.: Verification of load factors for concrete components by reliability and optimization analysis: background documents for implementing Eurocodes. Prog. Struct. Eng. Mater. **2**(4), 502–507 (2000)
15. Caramelli, S., Croce, P., Salvatore, W., Sanpaolesi, L.: Partial Safety Factors for Resistance of Steel Elements. University of Pisa (1997)
16. Fajkus, M., Holický, M., Rozlívka, L., Vorlíček, M.: Random Properties of Steel Elements Produced in Czech Republic, Eurosteel'99, Praha 1999, pp. 657–660 (1997)

17. Melcher, J., Kala, Z., Holický, M., Fajkus, M., Rozlívka, L.: Design characteristics of structural steels based on statistical analysis of metallurgical products. J. Constr. Steel Res. **60**(1), 795–808 (2004)
18. Holický, M.: Reliability Analysis for Structural Design. SUNN MeDIA, Stellenbosch (2009)

Appendix 7: Standardized Normal Distribution



$$\text{Area} = \Phi_U(u) - 0.5 = \int_0^u \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \mathrm{d}u$$

	u									
Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389

(continued)

M. Holický, *Introduction to Probability and Statistics for Engineers*,
DOI 10.1007/978-3-642-38300-7, © Springer-Verlag Berlin Heidelberg 2013

	u									
Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Brief Biographical Notes

Prof. Ing. Milan Holický, Ph.D., DrSc.
Klokner Institute, CTU in Prague

Prof. Dr. Milan Holický obtained his civil engineering degree at the Czech Technical University in Prague and his doctor degree at the University of Waterloo, Canada. He is involved in the research of structural reliability and risk assessment. He is the author or co-author of more than 300 scientific and technical publications, including textbooks and five monographs.

Since 1965, he has been employed at the Klokner Institute of CTU in Prague, and also lectures at CTU. Since 1991, he has represented the Czech Republic on the European Committee for Standardisation (CEN) as a member of the Technical committee TC 250 "Structural Eurocodes". In 2010, he became Extraordinary Professor at the University of Stellenbosch, South Africa; in 2011 he was awarded the Honorary Doctor of Science and Engineering degree from Moscow State University of Civil Engineering.

M. Holický, *Introduction to Probability and Statistics for Engineers*,
DOI 10.1007/978-3-642-38300-7,