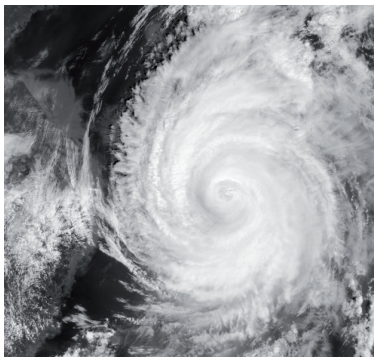
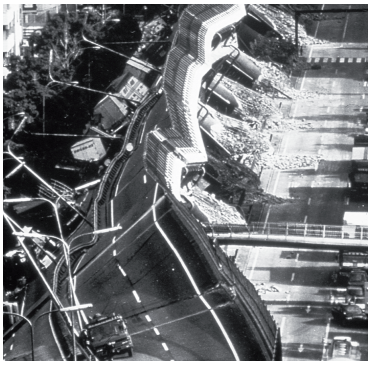
Statistics and Probability Theory






Lecture Notes

Prof. Dr. M. H. Faber

SS 2007

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

PREAMBLE

Introduction

The present script serves as study guidance for the students taking the course on Statistics and Probability Theory at the summer semester at ETH. The present script provides information concerning the:

- Aim of the course.
- Structure and organisation of the course.
- Educational support material for the course.
- Mode of tests and exam.
- Lecture notes for each of the 13 lectures with bibliography and index.

Information about the contents of the course and the organization of the course is also available on http://www.ibk.ethz.ch/fa/.

Aim of the course

The aim of the present course is to provide to the students the basic skills and tools of statistics and probability. Emphasis is directed on the application and the reasoning behind the application of these skills and tools for the purpose of enhancing engineering decision making.

It is expected that the students have only little or no prior knowledge on the subject of statistics and probability. The purpose of the present course is thus to ensure that the students will acquire during the course the required theoretical basis and technical skills such as to feel comfortable with the theory of basic statistics and probability. Moreover, in the present course as opposed to many standard courses on the same subject, the perspective is to focus on the use of the theory for the purpose of engineering model building and decision making.

The course is subdivided into the following seven modules, each consisting of one or more lectures (see also Figure 1):

- *Module A - Engineering decisions under uncertainty*
  Module B - Basic probability theory
  Module C - Descriptive statistics
  Module D - Uncertainty modelling
  Module E – Estimation and model building
  Module F – Methods of structural reliability
  Module G – Bayesian decision analysis

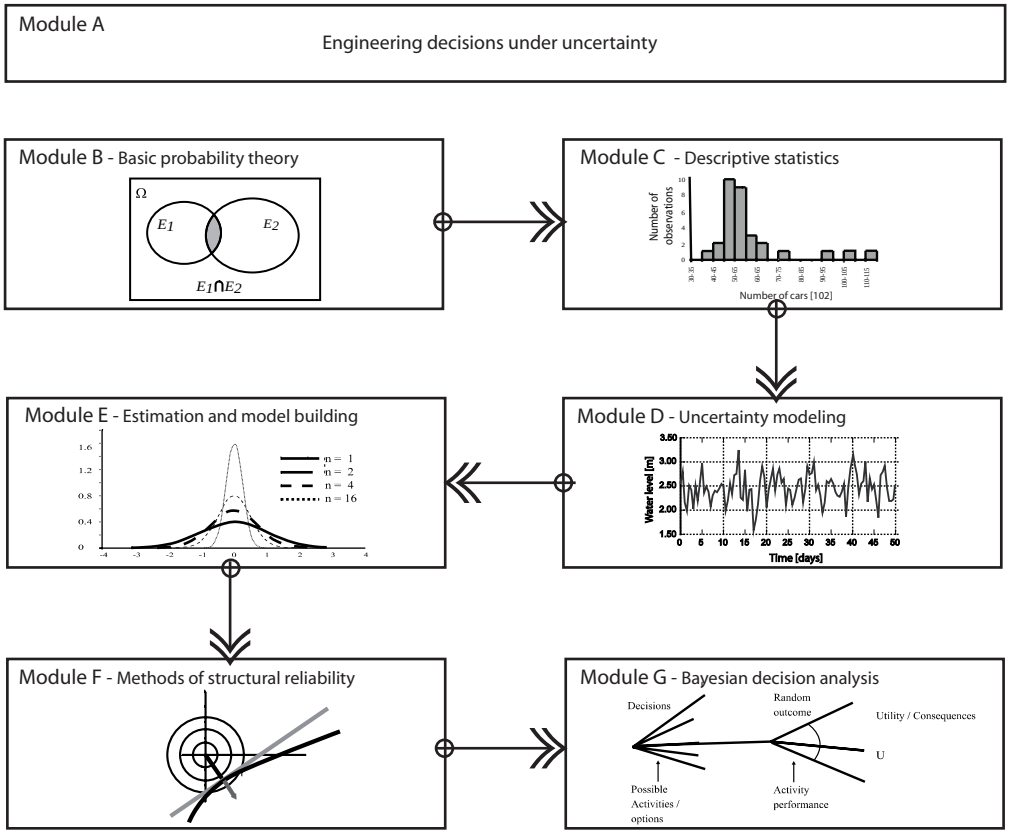


Figure 1: Illustration of the modules of the course and their didactical roles.

The didactical logic behind the presentation of the material in the course is first to provide a motivation for the application of statistics and probability as a basis for developing engineering models and for risk based decision making (Module A, see Figure 1). Thereafter, a basic introduction on the theory of probability (Module B) is provided. This is meant as a "brush up" of the knowledge already acquired by the students during high-school. Subsequently in Module C a selection of tools is provided, which enables engineers to assess and communicate data in a condensed form, namely the descriptive statistics. In Module D an introduction to uncertainty is provided together with a description of the various building stones required to represent uncertainties in engineering modelling in terms of random variables and processes. In Module E the main focus is directed on the aspects of postulating models, assessing model parameters and verifying models. Subsequently in Module F it is shown how, on the basis of formulated probabilistic models of uncertain variables, probabilities of events of significance for engineering decision making may be assessed. Finally in Module G it is shown how the engineering models of uncertainties and their probabilistic descriptions can be utilized in a systematic framework for engineering decision making.

It is believed that students having completed the present course will be able to:

- assess data based on observations and/or experiment results and present these in a standardized and unambiguous form,

- formulate and validate simple engineering models with due consideration of the associated uncertainties due to lack of knowledge and data as well as natural inherent variability,
- perform simple probability assessments such as to evaluate the probability of the appropriate performance of engineering activities,
- formulate and solve simple risk based decision problems.

Structure and organization of the course

With the aim of supporting the students in learning the specified curriculum, the course is built up by four main components, namely lectures, tutorials, assessments and self study:

- 13 weekly lectures of each two sessions of 45 minutes
- 11 weekly exercise tutorials of each two sessions of 45 minutes
- 2 assessments of each 90 minutes
- Self study estimated to 4 times by 45 minutes per week

This scheme is in accordance with the presently considered best ETH practice which assumes that the total efforts required to complete the course correspond to; lectures + exercise tutorials+assessment = 50%, self study = 50%.

Lectures:

The lectures are targeted at providing the students with the most important aspects of the theoretical and methodical material which may also be found in the present lecture notes. However, the lectures will also focus on the philosophical background for the development and use of the theoretical background and are thus to be understood as partly complementary to the material of the lecture notes. It is assumed and strongly suggested that the students study the lecture notes and become familiar with this.

Exercise Tutorials:

The exercise tutorials serve as a means of learning how to apply the theories and methodologies presented in the lectures and in the script. It is expected that the student will actively engage themselves in the tutorial work. During the tutorials, provided that there is sufficient time, the students may also use the possibility to ask about any problems related to the solutions of the solved exercises which are provided on the home page.

The students, furthermore, have the possibility to consult the teaching staff at defined office hours in regard to the contents of the lectures and the exercise tutorials. In order to enable the complete clarification of problems and questions regarding the teaching material, it is advisable to contact personally the teaching assistants, rather than making requests through E-mails.

For the purpose of supporting the students in their self evaluation, solved exercises, including previous exam exercises are made available on the home page http://www.ibk.ethz.ch/fa/.

In addition, in the lecture notes at the end of each chapter, there is a set of small principal exercises which the students can use to check and practice their knowledge and skills.

During the first exercise tutorial the students will be sub-divided into groups which will have to present the solution of a representative exercise during the course.

Each individual exercise tutorial includes the following activities:

- Presentation of 2 or more new exercises (corresponding to the subjects presented in the last lecture) in steps that will enable their solution by the students.
- Presentation by the teaching assistants of the solutions of 1 or more exercises of those presented in the previous exercise tutorial.
- Presentation by one group of students of the solution of one of the exercises presented in the previous exercise tutorial. Student colleagues and teaching assistants may ask questions for clarification during the presentation.

Educational support material for the course

The course is supported by the present script which provides the theory being taught during the lectures. All material for the course will be made available partly prior to the start of the course and partly during the course on the home page http://www.ibk.ethz.ch/fa/.

The course material contains besides the lecture notes also the Power Point presentations used for the lectures as well the solved exercises for each exercise tutorial. The lecture notes will be made available on the home page prior to the start of the course. The Power Point files of both lecture and exercise tutorials will be uploaded on the cours'es web page the latest one day before the respective class.

All solutions to the exercises, except the exercises for which the solutions will be presented by the students during the exercise tutorials will be made available on the home page prior to the start of the course. The solutions of the latter mentioned exercises will appear immediately after the presentations.

The Power Point presentations are only meant as a support for the lectures and can be used only as a support for the learning and preparation before each lecture. It is important to note that it is expected that the students read all the material contained in the lecture notes from lecture to lecture. Reading the Power Point presentations is not a substitute for reading the lecture notes which in many cases provide more detailed information.

Mode of Assessment (Tests and Exams)

The course performance comprises of a joint evaluation of:

- The results of the two assessments, one midterm (03.05.07) and the other one towards the end of the course (14.06.07)
- The final exam which will take place in the autumn (close or within October) as a part of the "Basisprüfung" (to be announced by the ETH Prüfungsplan).

The two assessments during the course have equal weight and must be attended by all students. In case documentation from a medical doctor or a military superior confirming illness or military duties, is presented to Prof. Faber before or within one week after a test which is not attended, arrangements for a substitute examination can be made. In case that an assessment is not attended by a student and no documentation is presented, the assessment will automatically be marked with 1 (1 out of 6, in the formal ETH scale with 6 being the best mark).

At each of the two assessments during the course a maximum mark of 6 can be obtained (in the formal ETH scale with 6 being the best mark). .

The mark achieved at the "Basisprüfung" will be based on the mark obtained at the final exam (autumn) weighted by 2/3 and the mark obtained from the two assessments weighted by 1/3.

The above mode of assessment applies also for PhD students attending the course with the aim to achieve the provided 5 credit points.

REPETITION OF THE BASIC EXAM

In the case that the mark at the "Basisprüfung" does not lead to a successful pass of the course the students have the opportunity to repeat the basic exam according to the ETH rules. In such a case the same conditions as above apply, meaning that the final mark will be based on the mark obtained at the basic exam weighted by 2/3 and the mark obtained from the two assessments during the course weighted by 1/3. The mark of the two assessments counts for whenever a student decides to repeat the basic exam. However the opportunity is provided to the students to follow again the course the next time that it takes place and repeat the two assessments. In that case the mark obtained at the repetition will count weighed by 1/3. Students who will decide to take that opportunity should keep in mind that they need to subscribe themselves for this course before the beginning of the respective semester.

Students who have already attended the course before the summer semester 2007 and need repeat the basic exam they have two opportunities: either to go directly to the basic examination or subscribe again for the course. In the first case the mark obtained at the basic exam will count for 3/3 of the total mark. In the second case the students will take part in the two assessments, the mark of which will count for 1/3 of the final mark and the other 2/3 will come from the mark obtained at the basic exam.

We will try to do our best for achieving the aims of this course and we are looking forward to a good cooperation with all of you.

TABLE OF CONTENTS

Annex T – Tables

MODULE A – ENGINEERING DECISIONS UNDER UNCERTAINTY

1st Lecture

Aim of the present lecture

The aim of the present lecture is to introduce the problem context of societal decision making and to outline how the concept of risk may provide a means for rational decisions in engineering. Focus is directed on the understanding of role of the engineer for the development and maintenance of societal functions.

On the basis of the lecture it is expected that the students should acquire knowledge on the following issues:

- What is sustainability?
- What is the role of the engineering in society?
- How may aspects of sustainability be related to life safety and cost optimal decision making?
- Which are the main different types of consequences to be considered in risk assessment?
- Why are there possible conflicts between economy, safety and environment?
- Why is engineering decision making influenced by uncertainties?
- What is the role of probability and consequence in decision making?
- What is the definition of risk?
- Which are the main phases to be considered in life cycle risk assessments in engineering decision making?

A.1 Introduction

During the last two decades, there is growing awareness that our world only has limited non-renewable natural resources such as energy and materials but also limited renewable resources like drinking water, clean air etc.. This led the so-called Brundtland Commission (1987) to the conclusion that a *sustainable* development is defined as a development "that meets the needs of the present without compromising the ability of future generations to meet their own needs". Sustainable *decision making* is thus presently understood as based on a joint consideration of society, economy and environment. In regard to environmental impacts the immediate implications for the planning, design and operation of civil engineering infrastructures are clear: Save energy, save non-renewable resources and find out about re-cycling of building materials, do not pollute the air, water or soil with toxic substances, save or even regain arable land and much more.

For civil engineering infrastructures and facilities in general, but not only for those, also the financial aspect is of crucial importance. Civil engineering infrastructures are financed by the public via taxes, public charges or other. In the end it is the individuals of society who pay and, of course, also enjoy the benefits derived from their existence. However, seen in the light of the conclusions of the Brundtland report (Brundtland, 1987), the intergenerational equity must be accounted for. Our generation must not leave the burden of maintenance or replacement of too short-lived structures to future generations and it must not use more of the financial resources than those really available. In this sense, civil engineering facilities should be optimal not only from a technological point of view but also from a sustainability point of view.

It is in general a concern how society may maintain and even improve the quality of life. All activities in society should thus aim at improving the life expectancy and increasing the gross domestic product (GDP); resulting in the conclusion that investments into life saving activities must be in balance with the resulting increase in life expectancy. For the present it is just stated that this problem constitutes a decision problem that can be analyzed using *cost benefit analysis*.

At present approximately 10 to 20% of the GDP for developed countries is being re-invested into life saving activities, such as public health, risk reduction and safety. Furthermore, the economical burden of degradation of infrastructure amounts for example for the USA to about 10% of the GDP in 1997 (see Alsalam et al., 1998). From these numbers it becomes apparent that the issue of safety and well being of the individuals in society as well as the durability of infrastructure facilities has a high importance for the performance of society and the quality of life of the individuals of society.

The present course attempts to provide the basic tools for supporting decision making in the context of planning, design and maintenance of civil engineering activities and structures.

Engineering facilities such as bridges, power plants, dams and offshore platforms are all intended to benefit, some way or another, the quality of life of the individuals of society. Therefore, whenever such facilities are planned, it is a prerequisite that the benefit of the facility can be proven considering all phases of the life of the facility, i.e. including design,

manufacturing, construction, operation and eventually decommissioning. If this is not the case, clearly the facility should not be established.

A.2 Societal Decision Making and Risk

On a societal level a beneficial engineering facility is normally understood as:

- being economically efficient in serving a specific purpose
- fulfilling given requirements in regard to the safety of the personnel directly involved with or indirectly exposed to the facility
- fulfilling given requirements to limit the adverse effects of the facility on the environment.

Taking basis in these requirements it is realised that the ultimate task of the engineer is to make decisions or to provide the decision basis for others such that it may be ensured that engineering facilities are established in such a way that they provide the largest possible benefit; if they cannot be proven to benefit they are not realized at all.

Example A.1 – Feasibility of hydraulic power plant

Consider as an example the decision problem of exploitation of hydraulic power. A hydraulic power plant project involving the construction of a water reservoir in a mountain valley is planned. The benefit of the hydraulic power plant is for simplicity assumed associated only with the monetary income from selling electricity to consumers. The decision problem thus simplifies to comparing the costs of establishing, operating and eventually decommissioning the hydraulic power plant with the incomes to be expected during the service life of the plant. In addition it must of course be ensured that the safety of the personnel involved in the construction and operation of the plant and the safety of third persons, i.e. the individuals of the society in general, is satisfactorily high.

Different solutions for establishing the power plant may be considered and their efficiency can be measured in terms of the expected income relative to the costs of establishing the power plant. However, a number of factors are important for the evaluation of the income and the costs of establishing the power plant. These are e.g. the period of time where the plant will be operating and produce electricity and the capacity of the power plant in terms of kWh. Moreover, the future income from selling electricity will depend on the availability of water, which depends on the future snow and rainfall. But also the market situation may change and competing energy recourses such as thermal and solar power may cause a reduction of the market price on electricity in general.

In addition the different possible solutions for establishing the power plant will have different costs and different implications on the safety to personnel. Obviously, the more capacity of the power plant, i.e. the higher the dam the larger the construction costs will be, but also the potential flooding (consequence of dam failure) will be larger in case of dam failure and more people would be injured or die, see Figure A.1.

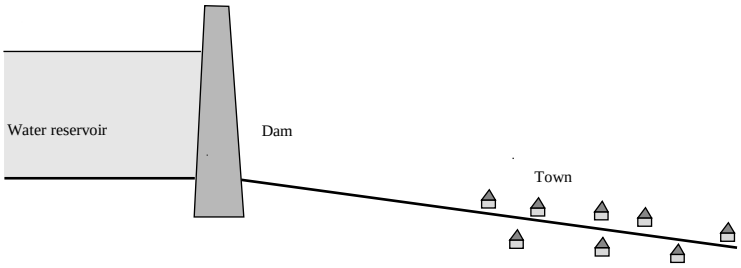


Figure A.1: Water reservoir/dam for exploitation of hydraulic power.

The safety of the people in a town downstream of the reservoir will also be influenced by the load carrying capacity of the dam structure relative to the pressure loading due to the water level in the reservoir. The strength of the dam structure depends in turn on the material characteristics of the dam structure and the properties of the soil and rock on which it is founded. As these properties are subject to uncertainty of various sources as shall be seen later, the load carrying capacity relative to the loading may be expressed in terms of the *probability* that the loading will exceed the load carrying capacity or equivalently the probability of dam failure.

Finally, the environmental impact of the power plant will depend on the water level in the reservoir, the higher the water level the more land will be flooded upstream of the dam structure and various habitats for animals and birds will be destroyed. On the other hand the water reservoir itself will provide a living basis for new species of fish and birds and may provide a range of recreational possibilities for people such as sailing and fishing which would not be possible without the reservoir.

In order to evaluate whether or not the power plant is feasible it is useful to make a list of the various factors influencing the benefit and their effects. As the problem may be recognized to be rather complex only the interrelation of the water level in the reservoir will be considered, the load carrying capacity of the dam structure, the costs of constructing the dam structure and the implications on the safety of the people living in a town down-stream the power plant.

Reservoir water level	Load carrying capacity of dam structure	Income	Costs	Consequence of dam failure	Probability of dam failure
Low	Low **Medium** High	Small	**Low** Medium High	Small	High **Medium** Low
Medium	Low **Medium** High	Medium	Low **Medium** High	Medium	High **Medium** Low
High	Low **Medium** High	Large	Low Medium **High**	Large	High **Medium** Low

Table A.1: Interrelation of benefits, costs and safety for the reservoir.

From Table A.1, which is clearly a simplified summary of the complex interrelations of the various factors influencing the benefit of realizing the power plant, it is seen that the various factors have different influences and that the different attributes such as income, costs and safety are conflicting. In the table it is assumed that the medium load carrying capacity of the dam structure corresponds to a medium probability of dam failure but of course other

combinations are also possible. Consider the case with a high water level in the reservoir. In this case the potential income is large but the costs of constructing the dam structure will also be high. Furthermore, the potential *consequences* in case of dam failure will be large as well. Table A.1 clearly points to the true character of the *decision problem*, namely that the optimal decision depends on the consequences should something go wrong and moreover the probability that something goes wrong. The product of these two factors is denoted the *risk*, a measure that will be considered in much more detail in the chapters to follow. Furthermore, not only the load carrying capacity of the dam structure is associated with *uncertainty* but in fact as indicated previously also the income expected from the power plant, due to uncertainties in the future market situation. In a similar way the costs of constructing the power plant are uncertain as also various difficulties encountered during the construction, such as unexpected rock formations, delay in construction works due to problems with material supplies, etc. may lead to additional costs.

When deciding on whether or not to establish the hydraulic power plant it is thus necessary to be able to assess consequences and probabilities; two key factors for the decision problem.

Both consequences and probabilities vary through the life of the power plant and this must be taken into account as well. In the planning phase it is necessary to consider the risk contributions from all subsequent phases of its life-cycle including decommissioning, see Figure A.2.

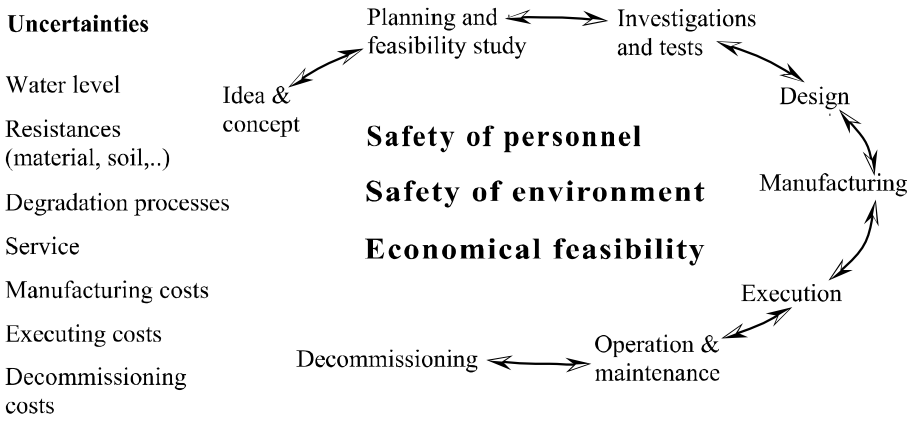


Figure A.2: Risk contributions from different service life phases to be considered at the planning stage.

It is important to recognize that different things may go wrong during the different phases of the service life including events such as mistakes and errors during design and failures and accidents during construction, operation and decommissioning. The potential causes of errors, mistakes, failures and accidents may be numerous, including human errors, failures of structural components, extreme load situations and not least natural hazards. Careful planning during the very first phase of a project is the only way to control the risks associated with such events.

As an illustration the dam structures must be designed such that the safety of the dam is ensured in all phases of the service life, taking into account yet another factor of uncertainty, namely the future deterioration, but also taking into account the quality of workmanship, the degree of quality control implemented during construction and not least the foreseen

strategies for the inspection and maintenance of the structures and mechanical equipment during the operation of the power plant. As a final aspect concerning the structures these should at the end of the service life be in such a condition that the work to be performed during the decommissioning of the power plant can be performed safely for both the persons involved and the environment.

A final fundamental problem arises in regard to the question – what are the *acceptable risks*? - what is one prepared to invest and/or pay for the purpose of getting a potential benefit? The decision problem of whether or not to establish the hydraulic power plant is thus seen to be a decision problem involving a significant element of uncertainty.

The mathematical basis for the treatment of such decision problems is the *decision theory*. Important aspects of decision theory are the assessment of consequences and probabilities and in a very simplified manner one can say that risk and reliability analysis in civil engineering is concerned with the problem of decision making subject to uncertainty.

A.3 Definition of Risk

In daily conversation *risk* is a rather common notion used interchangeably with words like *chance, likelihood* and *probability* to indicate that people are uncertain about the state of the activity, item or issue under discussion. For example the risk of getting cancer due to cigarette smoking is discussed, the chance of succeeding to develop a vaccine against the HIV virus in 2007, the likelihood of getting a "Royal Flush" in a Poker game and the probability of a major earthquake occurring in the Bay area of San Francisco within the next decade.

Even though it may be understandable from the context of the discussion what is meant by the different words it is necessary in the context of engineering decision making that those involved are precise in our understanding of risk. Risk is to be understood as the expected consequences associated with a given activity, the activity being e.g. the construction, operation and decommissioning of a power plant.

Considering an activity with only one event with potential consequences C the risk R is the probability that this event will occur P multiplied with the consequences given the event occurs i.e.:

$$R = P\ C \tag{A.1}$$

If e.g. n events with consequences C_i and occurrence probabilities P_i may result from the activity the total risk associated with the activity is simply assessed through the sum of the risks from the individual events, i.e.:

$$R = \sum_{i=1}^{n} P_i\ C_i \tag{A.2}$$

This definition of risk is consistent with the interpretation of risk used e.g. in the insurance industry and risk may e.g. be given in terms of Euros, Dollars or the number of human

fatalities. Even though most risk assessments have some focus on the possible negative consequences of events the definitions in Equations (A.1)-(A.2) are also valid in the case where benefits are taken into account. In fact and as will be elaborated in Module F, in this case the definitions in Equations (A.1)-(A.2) are more general and consistent with *expected utility* utilized as basis for *decision analysis*.

Self Assessment Questions/Exercises

A.1 What is meant by the term "sustainable development" and why is it important for engineering decision making?

A.2 How a beneficial engineering facility is normally understood on a societal level?

A.3 How may the risk of an event be defined and how may be expressed analytically?

A.4 What is meant by the term "acceptance risks"?

A.5 Considering an activity with only one event with potential consequences, the risk is that probability that this event will occur multiplied with the consequences given the event occurs.

Which of the following events is associated with the highest risk?

Event	1	2	3
Event probability	10%	1%	20%
Consequences	100 SFr	500 SFr	100 SFr
Risk			

MODULE B – BASIC PROBABILITY THEORY

2nd Lecture

Aim of the present lecture

The aim of the present lecture is to introduce the basics of set and probability theory. Different interpretations of the important concept of probability are provided and it is outlined that the Bayesian probability interpretation facilitates for an integration of the other interpretations. The very simple and few axioms of probability theory are given and the important results regarding conditional probabilities and the associated Bayes' rule are outlined.

On the basis of the lecture it is expected that the students should acquire knowledge and skills in regard to:

- Which are the different interpretations of probability?
- What is a sample space and how may events be illustrated?
- What is an event and what is a complementary event?
- How are intersections and unions of sets defined?
- How may operations involving intersections and unions of events be performed?
- Which are the axioms of probability theory?
- Which are the implications of mutual exclusivity between events?
- What is a conditional probability and how may it be evaluated?
- Which are the implications of independence?
- What is Bayes' rule, and how can it be interpreted?
- How can Bayes' rule be applied for probability updating?

B.1 Introduction

Probability theory forms the basis of the assessment of probabilities of occurrence of uncertain events and thus constitutes a cornerstone in risk and decision analysis. Only when a consistent basis has been established for the treatment of the uncertainties influencing the probability that events with possible adverse consequences may occur, it is possible to assess the risks associated with a given activity and thus to establish a rational basis for decision making.

The level of uncertainty associated with a considered activity or phenomenon may be expressed by means of purely qualitative statements such as “the chance is good” or “the likelihood is low” but may also be quantified in terms of numbers or percentages. However, the different words in fact all have the meaning of probability and in the following section this notion will be investigated and especially the theoretical framework for its quantification in more detail.

B.2 Definition of Probability

The purpose of the theory of probability is to enable the quantitative assessment of probabilities but the real meaning and interpretation of probabilities and probabilistic calculations as such is not a part of the theory. Consequently two people may have completely different interpretations of the probability concept, but still use the same calculus. In the following, three different interpretations of probability are introduced and discussed based on simple cases. A formal presentation of the axioms of probability theory is provided in Section B.4.

Frequentistic Definition

The *frequentistic* definition of probability is the typical interpretation of probability of the *experimentalist*. In this interpretation the probability $P(A)$ is simply the *relative frequency* of occurrence of the *event* A as observed in an experiment with n trials, i.e. the probability of an event A is defined as the number of times that the event A occurs divided by the number of experiments that are carried out:

$$P(A)=\lim \frac{N_A}{n_{\text{exp}}} \qquad \text{for} \qquad n_{\text{exp}} \to \infty \tag{B.1}$$

N_A = number of experiments where A occurred

n_{exp} = total number of experiments.

If a frequentist is asked what the probability is for achieving a “head” when flipping a coin she would principally not know what to answer until she would have performed a large number of experiments. If say after 1000 experiments (flips with the coin) it is observed that “head” has occurred 563 times the answer would be that the probability for “head” is 0.563.

However, as the number of experiments is increased the probability would converge towards 0.5. In the mind of a frequentist, probability is a characteristic of nature.

Classical Definition

The *classical probability* definition originates from the days when the probability calculus was founded by Pascal and Fermat[1]. The inspiration for this theory was found in the games of cards and dice. The classical definition of the probability of the event A can be formulated as:

$$P(A) = \frac{n_A}{n_{tot}} \tag{B.2}$$

n_A = number of equally likely ways by which an experiment may lead to A

n_{tot} = total number of equally likely ways in the experiment.

According to the classical definition of probability, the probability of achieving a "head" when flipping a coin would be 0.5 as there is only one possible way to achieve a "head" and there are two equally likely outcomes of the experiment.

In fact there is no real contradiction to the frequentistic definition, but the following differences may be observed:

- The experiment does not need to be carried out as the answer is known in advance.
- The classical theory gives no solution unless all equally possible ways can be derived analytically.

Bayesian Definition

In the *Bayesian interpretation* the probability $P(A)$ of the event A is formulated as a *degree of belief* that A will occur:

$$P(A) = \text{degree of belief that } A \text{ will occur} \tag{B.3}$$

Coming back to the coin-flipping problem the Bayesian would argue that there are two possibilities, and as she has no preferences as to "head" or "tail" she would judge the probability of achieving a "head" to be 0.5.

The degree of belief is a reflection of the state of mind of the individual person in terms of experience, expertise and preferences. In this respect the Bayesian interpretation of probability is *subjective* or more precisely person-dependent. This opens up the possibility that two different persons may assign different probabilities to a given event and thereby contradicts the frequentist interpretation that probabilities are a characteristic of nature.

[1] Pierre de Fermat, mathematician, 1601-1665; Blaise Pascal, mathematician, 1623-1662.

The Bayesian statistical interpretation of probability includes the frequentistic and the classical interpretation in the sense that the subjectively assigned probabilities may be based on experience from previous experiments (frequentistic) as well as considerations of e.g. symmetry (classical).

The degree of belief is also referred to as a *prior belief* or *prior probability*, i.e. the belief, which may be assigned prior to obtaining any further knowledge. It is interesting to note that Immanuel Kant[2] developed the purely philosophical basis for the treatment of subjectivity at the same time as Thomas Bayes[3] developed the mathematical framework later known as the *Bayesian statistics*.

Modern structural reliability and risk analysis is based on the Bayesian interpretation of probability. However, the degree of freedom in the assignment of probabilities is in reality not as large as indicated in the above. In a formal Bayesian framework the subjective element should be formulated before the relevant data are observed. Arguments of objective symmetrical reasoning and physical constraints, of course, should be taken into account.

Practical Implications of the Different Interpretations of Probability

In some cases probabilities may adequately be assessed by means of *frequentistic information*. This is e.g. the case when the probability of failure of mass produced components are considered, such as pumps, light bulbs and valves. However, in order to utilise reported failures for the assessment of probability of failure for such components it is a prerequisite that the components are in principle identical, that they have been subject to the same operational and/or loading conditions and that the failures can be assumed to be independent.

In other cases when the considered components are e.g. bridges, high-rise buildings, ship structures or unique configurations of pipelines and pressure vessels, these conditions are not fulfilled. In these cases the number of identical structures may be very small (or even just one) and the conditions in terms of operational and loading conditions are normally significantly different from structure to structure. In such cases the Bayesian interpretation of probability is far more appropriate.

The basic idea behind the Bayesian statistics is that lack of knowledge should be treated by probabilistic reasoning, similarly to other types of uncertainty. In reality, decisions have to be made despite the lack of knowledge and probabilistic tools are a great help in that process.

[2] Immanuel Kant, philosopher, 1724-1804

[3] Thomas Bayes, mathematician, 1702-1761

B.3 Sample Space and Events

Considering e.g. the compressive strength of concrete this material characteristic may be estimated by performing laboratory experiments on standardized test specimens (cylinders or cubes). The test results will, however, probably be different from one another and the concrete compressive strength shall be assumed to be an uncertain quantity or a random quantity. The set of all possible outcomes of the concrete compressive strength experiments is called the *sample space* (denoted Ω) for the random quantity – the concrete compressive strength. In this example the sample space is the open interval $\Omega =]0;\infty[$, i.e. the set of all positive real numbers. In this case the sample space is furthermore continuous but in other cases (e.g. when considering the outcome of throwing a dice) the sample space can also be discrete and countable.

An *event* is defined as a subset of a sample space and thus a set of sample points. If the subset is empty (i.e. contains no sample points) it is said to be impossible. An event is said to be certain if it contains all sample points in the sample space (i.e. the event is identical to the sample space).

Consider the events E_1 and E_2 shown in Figure B.1. The subset of sample points belonging to the event E_1 or the event E_2 is denoted the *union* of the events E_1 and E_2 written as $E_1 \cup E_2$.

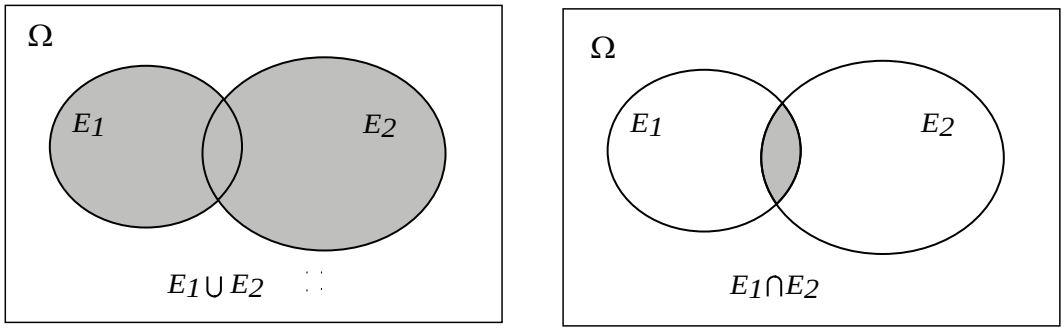


Figure B.1: ***Venn diagrams*** **illustrating the union of events (left) and the intersection of events (right).**

The subset of sample points belonging to E_1 and E_2 is denoted the *intersection* of E_1 and E_2 and is written as $E_1 \cap E_2$. The intersection of these two events is illustrated in the right portion of Figure B.1.

The two events are said to be *mutually exclusive* if they are *disjoint* (i.e. if they have no common sample points). In this case the intersection of E_1 and E_2 is empty (i.e. $E_1 \cap E_2 = \varnothing$), where $\varnothing$ denotes the empty set.

Consider the event E in the sample space Ω. The event containing all sample points in Ω, which are not included in E is called the complementary event to E and denoted $\overline{E}$. It then follows directly that $E \cup \overline{E} = \Omega$ and that $E \cap \overline{E} = \varnothing$.

It can be shown that the intersection and union operations obey the following *commutative, associative and distributive laws*:

$$
\begin{aligned}
&E_1 \cap E_2 = E_2 \cap E_1 \\
&E_1 \cap (E_2 \cap E_3) = (E_1 \cap E_2) \cap E_3 \\
&E_1 \cup (E_2 \cup E_3) = (E_1 \cup E_2) \cup E_3 \\
&E_1 \cap (E_2 \cup E_3) = (E_1 \cap E_2) \cup (E_1 \cap E_3) \\
&E_1 \cup (E_2 \cap E_3) = (E_1 \cup E_2) \cap (E_1 \cup E_3)
\end{aligned}
\tag{B.4}
$$

From which the following laws (denoted *De Morgan's laws*) may be derived:

$$
\begin{aligned}
&E_1 \cap E_2 = \overline{\overline{E}_1 \cup \overline{E}_2} \\
&E_1 \cup E_2 = \overline{\overline{E}_1 \cap \overline{E}_2}
\end{aligned}
\tag{B.5}
$$

B.4 The three Axioms of Probability Theory

Mathematically the probability theory is built up using only the following three axioms:

Axiom 1:

$$0 \le P(E) \le 1 \qquad \text{for any given } E \subset \Omega \tag{B.6}$$

where P is the probability measure.

Axiom 2:

$$P(\Omega) = 1 \tag{B.7}$$

where Ω is the sample space.

Axiom 3:

Given that $E_1, E_2, ..E_n$ are mutually exclusive events then:

$$P(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i) \tag{B.8}$$

These three *axioms of probability theory* form the sole basis of the theory of probability.

B.5 Conditional Probability and Bayes' Rule

Conditional probabilities are of special interest in risk and reliability analysis as they form the basis of the updating of probability estimates based on new information, knowledge and evidence.

The conditional probability of the event E_1 given that the event E_2 has occurred is written as:

$$P(E_1 \mid E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} \tag{B.9}$$

It is seen that the conditional probability is not defined if the conditioning event is the empty set, i.e. when $P(E_2) = 0$.

The event E_1 is said to be probabilistically independent of the event E_2 if :

$$P(E_1 \mid E_2) = P(E_1) \tag{B.10}$$

implying that the occurrence of the event E_2 does not affect the probability of E_1.

From Equation (B.9) the probability of the event $E_1 \cap E_2$ may be given as:

$$P(E_1 \cap E_2) = P(E_1 \mid E_2) P(E_2) \tag{B.11}$$

and it follows immediately that if the events E_1 and E_2 are independent, then:

$$P(E_1 \cap E_2) = P(E_1) P(E_2) \tag{B.12}$$

Based on the above findings, the important *Bayes' rule* can be derived.

Consider the sample space Ω divided into n mutually exclusive events $E_1, E_2, .. E_n$ (see also Figure B.2, where the case of $n = 8$ is considered).

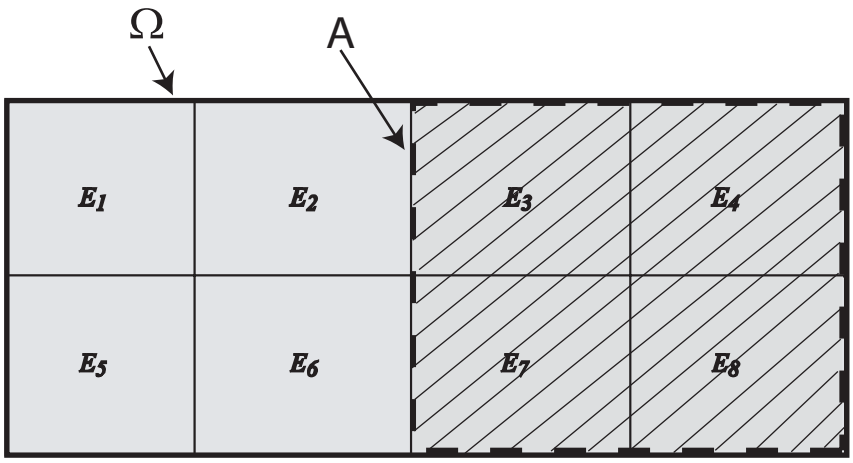


Figure B.2: Illustration of the rule of Bayes.

Furthermore let the event A be an event in the sample space Ω. Then the probability of the event A, i.e. $P(A)$, can be written as:

$$\begin{aligned} P(A) &= P(A \cap E_1) + P(A \cap E_2) + ... + P(A \cap E_n) \\ &= P(A|E_1)P(E_1) + P(A|E_2)P(E_2) + ... + P(A|E_n)P(E_n) \\ &= \sum_{i=1}^{n} P(A|E_i)P(E_i) \end{aligned} \tag{B.13}$$

this is also referred to as the *total probability theorem*.

From Equation (B.9) there is $P(A|E_i)P(E_i) = P(E_i|A)P(A)$ implying that:

$$P(E_i|A) = \frac{P(A|E_i)P(E_i)}{P(A)} \tag{B.14}$$

Now by inserting Equation (B.13) into Equation (B.14), the *Bayes' rule*results:

$$P(E_i|A) = \frac{P(A|E_i)P(E_i)}{\sum_{j=1}^{n} P(A|E_j)P(E_j)} \tag{B.15}$$

In Equation (B.15) $P(E_i|A)$ is denoted the *posterior probability* of E_i, the conditional term $P(A|E_i)$ is often referred to as the *likelihood* (i.e. the probability of observing a certain state given the true state). The term $P(E_i)$ is the *prior probability* of the event E_i (i.e. prior to the knowledge about the event A).

As mentioned previously, the rule due to Bayes' is extremely important, and in order to facilitate the appreciation of this a few illustrative applications of Bayes' rule will be given in the following.

Example B.1 – Using Bayes' rule for concrete assessment

A reinforced concrete beam is considered. From experience it is known that the probability that corrosion of the reinforcement has initiated (the event *CI*) is $P(CI) = 0.01$. However, in order to know the condition more precisely an inspection method (non-destructive) has been developed.

The quality of the inspection method may be characterised by the probability that the inspection method will indicate (I) initiated corrosion given that corrosion has initiated $P(I|CI)$ (the probability of detection or equivalently the likelihood of an indication *I* given corrosion initiation *CI*) and the probability that the inspection method will indicate initiated

corrosion given that no corrosion has initiated $P(I|\overline{CI})$ (the probability of erroneous findings or the likelihood of an indication given no corrosion initiation).

For the inspection method at hand the following characteristics have been established:

$P(I|CI) = 0.8$

$P(I|\overline{CI}) = 0.1$

An inspection of the concrete beam is conducted with the result that the inspection method indicates that corrosion has initiated. Based on the findings from the inspection, what is the probability that corrosion of the reinforcement has initiated?

The answer is readily found by application of Bayes' rule:

$$P(CI|I) = \frac{P(I|CI)P(CI)}{P(I|CI)P(CI) + P(I|\overline{CI})P(\overline{CI})} = \frac{P(I \cap CI)}{P(I)} \tag{B.16}$$

With $P(I)$, the probability of obtaining an indication of corrosion at the inspection:

$$P(I) = P(I|CI)P(CI) + P(I|\overline{CI})P(\overline{CI}) = 0.8 \cdot 0.01 + 0.1 \cdot (1 - 0.01) = 0.107$$

and $P(I \cap CI)$ the probability of receiving an indication of initiated corrosion and at the same time to have initiated corrosion:

$$P(I \cap CI) = P(I|CI)P(CI) = 0.8 \cdot 0.01 = 0.008$$

Thus, the probability that corrosion of the reinforcement has initiated given an indication of initiated corrosion by the inspection method is:

$$P(CI|I) = \frac{0.008}{0.107} = 0.075$$

The probability of initiated corrosion, given an indication of initiated corrosion, is surprisingly low. This is due to the high probability of an erroneous indication of initiated corrosion by the inspection method relative to the small probability of initiated corrosion (i.e. the inspection method is not sufficiently accurate for the considered application).

□

Example B.2 – Using Bayes' rule for bridge upgrading

An old reinforced concrete bridge is reassessed in connection with an upgrading of the allowable traffic (see also Schneider, 1994). The concrete compressive strength class is unknown but concrete cylinder samples may be taken from the bridge and tested in the laboratory.

The following classification of the concrete is assumed:

B1: $0 \le \sigma_c < 30$

B2: $30 \le \sigma_c < 40$

B3: $40 \le \sigma_c$

Even though the concrete class is unknown, experience with similar bridges suggests that the probability of the concrete of the bridge belonging to class B_1, B_2 and B_3 is 0.65, 0.24 and 0.11, respectively. This information comprises the prior information – prior to any experiment result.

The test method is not perfect in the sense that even though the test indicates a value of the concrete compressive strength belonging to a certain class, there is a certain probability that the concrete belongs to another class. The likelihoods for the considered test method are given in Table B.1.

It is assumed that one test is performed and it is found that the concrete compressive strength is equal to 36.2 MPa, i.e. in the interval of class *B2*.

Using Bayes' rule, the probability that the concrete belongs to one of the different classes may now be updated. The posterior probability that the concrete belongs to class B_2 is given by:

$$P(B_2 \,|\, I = B_2) = \frac{0.61 \cdot 0.24}{0.61 \cdot 0.24 + 0.28 \cdot 0.65 + 0.32 \cdot 0.11} = 0.40$$

The posterior probabilities for the other classes may be calculated in a similar manner, the results are given in Table B.1.

Concrete Class	Prior Probability	Likelihood $P(I \,\|\, B_i)$			Posterior probabilities
		$I = B_1$	$I = B_2$	$I = B_3$	
B_1	0.65	0.71	0.28	0.01	0.50
B_2	0.24	0.18	0.61	0.21	0.40
B_3	0.11	0.02	0.32	0.66	0.10

Table B.1: Summary of prior probabilities, likelihoods of experiment outcomes and posterior probabilities given one test result in the interval of class B_2.

Self Assessment Questions/ Exercises

B.1 A person is asked what is the probability for achieving a "head" when flipping a coin. The person after 1000 experiments (flips with the coin) observes that "head" has occurred 333 times and hence answers that the probability for "head" is 0.333. On which interpretation of probability is this estimation based on?

B.2 How may the conditional probability of an event E_1 ,given that the event E_2 has occurred, be written?

B.3 In probability theory the probability, $P(A)$, of an event A can take any value within the following boundaries:

$0 \le P(A) \le 1$ ☐

$-1 \le P(A) \le 1$ ☐

$-\infty < P(A) < \infty$ ☐

B.4 If the intersection of two events, A and B corresponds to the empty set $\varnothing$, i.e. $A \cap B = \varnothing$, the two events are:

Mutually exclusive. ☐

Independent. ☐

Empty events. ☐

B.5 Which one(s) of the following expressions is(are) correct?

The probability of the union of two events A and B is equal to the sum of the probability of event A and the probability of event B, given that the two events are mutually exclusive. ☐

The probability of the union of two events A and B is equal to the probability of the sum of event A and event B, given that the two events are mutually exclusive. ☐

The probability of the intersection of two events A and B is equal to the product of the probability of event A and the probability of event B, given that the two events are mutually exclusive. ☐

The probability of the intersection of two events A and B is equal to the product of the probability of event A and the probability of event B, given that the two events are independent. ☐

B.6 The probability of the intersection of two mutually exclusive events is equal to:

The product of the probabilities of the individual events. ☐

The sum of the probabilities of the individual events. ☐

The difference between the probabilities of the individual events. ☐

One (1). ☐

Zero (0). ☐

B.7 Which one of the following statements is correct?

An event A is defined as a subset of a sample space Ω. ☐

A sample space Ω is defined as a subset of an event A. ☐

B.8 The probability of the union of two not mutually exclusive events A and B is given as: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. It is provided that the probability of event A is equal to 0.1, the probability of event B is 0.1 and the probability of event B given event A, i.e. $P(B|A)$ is 0.8. Which result is correct?

$P(A \cup B) = -0.6$ ☐

$P(A \cup B) = 0.12$ ☐

$P(A \cup B) = 0.04$ ☐

B.9 For an event A in the sample space Ω, event $\overline{A}$ represents the complementary event of event A. Which one(s) of the following hold?

$A \cup \overline{A} = \Omega$ ☐

$A \cap \overline{A} = \Omega$ ☐

$A \cup \overline{A} = \varnothing$ ☐

B.10 The commutative, associative and distributive laws describe how to:

Operate with intersections of sets. ☐

Operate with unions of sets. ☐

None of the above. ☐

B.11 Research in ETH is often funded by the Swiss National Foundation of research (SNF). The normal procedure is that a Professor submits a proposal for a new project. Experts working for SNF read the proposal and they may come to one of the following decisions:

D_1: the proposal is accepted and the project will be funded.

D_2: the proposal should be revised by the Professor and resubmitted to SNF.

D_3 : the proposal is not accepted and hence no funding is provided.

Professor Muster works at ETH. During the past few years he has submitted many proposals to SNF. Based on experience, over many years, Professor Muster in general assesses that when he submits a proposal the probabilities associated with the possible final decisions of SNF are as follows:

$P(D_1) = 0.45$, $P(D_2) = 0.35$, $P(D_3) = 0.2$.

By coincidence, just at the time when Professor Muster considers to submit a new proposal to SNF, he meets Dr. Beispiel. Dr. Beispiel used to work at SNF as one of the experts who review proposals and make the final decisions. Professor Muster kindly asks Dr. Beispiel to have a look at the new proposal before submitting it to SNF with the purpose of assessing the probabilities that the proposal would be accepted as it is. Of course Dr. Beispiel cannot say with certainty what will be the final SNF decision. However, his assessment can be considered as an indication, I , of the final decision of SNF. Based on experience from previous assessments and final decisions the conditional probabilities, $P(I = D_j \,|\, D_i)$, of the indications I of Dr. Beispiel given the final decisions D_i of SNF are as summarized in the following table.

SNF final decision D_i	**Dr. Beispiel's indicative assessment,** I_j		
	$I = D_1$	$I = D_2$	$I = D_3$
D_1	**0.86**	**0.1**	
D_2	**0.2**		**0.06**
D_3		**0.1**	**0.9**

a. Complete the above table.

b. Having read the new proposal Dr. Beispiel explains to Professor Muster that if he would still have been working with SNF he would have asked for revisions and resubmission. Based on this new information - what is the probability that the final decision of SNF is the same as the assessment of Dr. Beispiel?

MODULE C – DESCRIPTIVE STATISTICS

3rd Lecture

Aim of the present lecture

The aim of the present lecture is to introduce the descriptive statistics in terms of numerical summaries and graphical representations. It is outlined how data may be represented in a standardized manner in terms of different numerical measures as well as in the form of different graphs.

On the basis of the lecture it is expected that the students should acquire knowledge and skills in regard to:

- What is the purpose of descriptive statistics?
- In what principally different ways may data be assessed and communicated?
- What are the assumptions underlying descriptive statistics?
- Which are the different “central numerical measures” and what do they describe?
- What is a measure of dispersion and which such measures are available?
- What does peakedness and skewness refer to?
- What is the significance of correlation and how may it be calculated?
- Which are typical graphical representations of data sets?
- What is the difference between a sample histogram and a frequency distribution?
- What information is contained in a Quantile-Quantile plot?
- What are the main components of a Tukey-Box plot?
- In what way may numerical summaries be related to graphical representations?

C.1 Introduction

In order to assess the characteristics and the level of uncertainty of a given quantity of interest, one of the first steps is to investigate the data available, such as *observations* and *test results*. For this purpose, the use of *descriptive statistics* is useful. Descriptive statistics do not assume anything in terms of the degree or nature of the randomness underlying the data analysed, but are merely a convenient tool to reduce the data to a manageable form suitable for further analysis, as well as for communication of the data in a standardized format to other professionals.

In the following the so-called *numerical summaries* will first be introduced. These can be considered to be numerical characteristics of the observed data containing important information about the data and the nature of uncertainty associated with these. These are also referred to as *sample characteristics* in the following. Thereafter *graphical representations* are introduced as means of visual characterisation and as a useful tool for data analysis. Descriptive statistics play an important role in engineering risk analysis as this forms a standardized basis for assessing and documenting data obtained for the purpose of understanding and representing uncertainties in risk assessment.

C.2 Numerical Summaries

Central Measures

One of the most useful numerical summaries is the *sample mean*. If the data set is collected in the vector $\mathbf{x} = (x_1, x_2, .., x_n)^T$ the sample mean $\bar{x}$ is simply given as:

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

The sample mean may be interpreted as a central value of the data set. If, on the basis of the data set, one should give only one value characterising the data, one would normally use the sample mean. Another central measure is the *mode* of the data set i.e. the most frequently occurring value in the data set. When data samples are real values, the mode in general cannot be assessed numerically, but may be assessed from graphical representations of the data as will be illustrated in Section C.3.

As will be seen repeatedly in the present lecture notes it is often convenient to work with an ordered data set which is readily established by rearranging the original data set $\mathbf{x} = (x_1, x_2, .., x_n)^T$ such that the data are arranged in increasing order as $x_1^O \le x_2^O \le .. \le x_i^O .. \le x_{n-1}^O \le x_n^O$. In the subsequent the $i-th$ value of an ordered data set is denoted by x_i^O.

The *median* of the data set is defined as the middle value in the ordered list of data if *n* is odd. If *n* is even the median is taken as the average value of the two middle values (see also the examples below).

Example C.1 - Concrete Compressive Strength Data

Consider the data set given in Table C.1 corresponding to concrete cube compressive strength measurements. In the table the data are listed both unordered, e.g. in the order they were observed and ordered according to increasing values.

The sample mean for the data set is readily evaluated using Equation (C.1) and found to be equal to 32.67 MPa. All the observed values are different and therefore the mode cannot be determined without dividing the observations into intervals as will be shown in Section C.3. However, the median is readily determined as being equal to 33.05 MPa.

□

	Unordered	Ordered
i	x_i	x_i^o
1	35.8	24.4
2	39.2	27.6
3	34.6	27.8
4	27.6	27.9
5	37.1	28.5
6	33.3	30.1
7	32.8	30.3
8	34.1	31.7
9	27.9	32.2
10	24.4	32.8
11	27.8	33.3
12	33.5	33.5
13	35.9	34.1
14	39.7	34.6
15	28.5	35.8
16	30.3	35.9
17	31.7	36.8
18	32.2	37.1
19	36.8	39.2
20	30.1	39.7

Table C.1: Concrete cube compressive strength experiment results in MPa.

Example C.2 - Traffic Flow Data

Consider the data shown in Table C.2. The data correspond to the daily traffic flow in both directions through the Gotthard tunnel for the month of January 1997 obtained within a project carried out by the Swiss Federal Highways Office (ASTRA).

For this dataset the sample mean values of the traffic flow in direction 1 and direction 2 may be calculated from either the unordered or ordered data sets to be equal to 4697.39 and 5660.77 respectively. The corresponding median values can be read from the ordered data sets as 4419 and 5100 respectively (there are in total 31 observations in the data sets, so the median corresponds to observation 16).

□

	Direction 1			Direction 2		
	Unordered		Ordered	Unordered		Ordered
i	Date	x_i	x_i^O	Date	x_i	x_i^O
1	01.01	3087	3087	01.01	3677	3677
2	02.01	4664	3578	02.01	7357	4453
3	03.01	4164	3710	03.01	9323	4480
4	04.01	3710	3737	04.01	11748	4560
5	05.01	4029	3906	05.01	10256	4635
6	06.01	4323	4029	06.01	4453	4648
7	07.01	4041	4041	07.01	4815	4672
8	08.01	3737	4085	08.01	4757	4757
9	09.01	4103	4103	09.01	4672	4791
10	10.01	5457	4164	10.01	5401	4815
11	11.01	4563	4323	11.01	5688	4880
12	12.01	3906	4359	12.01	6308	4928
13	13.01	4419	4366	13.01	4946	4946
14	14.01	4359	4368	14.01	4635	5005
15	15.01	4667	4371	15.01	5100	5013
16	16.01	5098	4419	16.01	4791	5100
17	17.01	6551	4563	17.01	5235	5220
18	18.01	4371	4588	18.01	4560	5235
19	19.01	3578	4664	19.01	5729	5281
20	20.01	4366	4667	20.01	5005	5318
21	21.01	4368	4727	21.01	4480	5398
22	22.01	4588	4739	22.01	4880	5401
23	23.01	5001	4741	23.01	4928	5679
24	24.01	7118	5001	24.01	5398	5688
25	25.01	4727	5098	25.01	4648	5729
26	26.01	4085	5193	26.01	6183	6183
27	27.01	4741	5457	27.01	5220	6308
28	28.01	4739	5892	28.01	5013	7357
29	29.01	5193	6551	29.01	5281	9323
30	30.01	5892	7118	30.01	5318	10256
31	31.01	7974	7974	31.01	5679	11748

Table C.2: Daily traffic flow through the Gotthard tunnel, January 1997.

Dispersion Measures

The variability or the *dispersion* of the data set around the sample mean is also an important characteristic of the data set. This dispersion may be characterised by the *sample variance* s^2 given by:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{C.2}$$

and the *sample standard deviation* s is defined as the square root of the sample variance. From Equation (C.2) it is seen that the sample standard deviation s is assessed in terms of the variability of the observations around the sample mean value $\bar{x}$.

Thus, the sample variance is the mean of the squared deviations from the sample mean and is in this way analogous to the moment of inertia as used in e.g. structural engineering.

As a means of comparison of the dispersions of different data sets, the dimensionless *sample coefficient of variation* ν is convenient. The sample coefficient of variation ν is defined as the ratio of the sample standard deviation to the sample mean, i.e. given by:

$$\nu = \frac{s}{\bar{x}} \tag{C.3}$$

The *sample variance* for the concrete cube compressive strengths of Table C.1 may be evaluated using Equation (C.3) and is found to be 16.36 MPa^2. The sample standard deviation is thus 4.04 MPa. For the considered concrete cube compressive strength data the sample coefficient of variation is equal to 0.12. In the same manner the sample coefficient of variation for the traffic flow data in Table C.2 is equal to 0.21 and 0.30 for direction 1 and direction 2 respectively. It is seen that the coefficient of variation for direction 2 is higher than for direction 1. That indicates that the data observed in direction 2 are more dispersed than in direction 1.

Other Measures

Whereas the sample mean, mode and median are central measures of a data set, and the sample variance is a measure of the dispersion around the sample mean it is also useful to have some characteristic indicating the degree of symmetry of the data set. To this end the sample coefficient of *skewness*, which is a simple logical extension of the sample variance is suitable. The sample coefficient of skewness η is defined as:

$$\eta = \frac{1}{n} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{s^3} \tag{C.4}$$

This coefficient is positive if the mode of the data set is less than its mean value (skewed to the right) and negative if the mode is larger than the mean value (skewed to the left). For the concrete cube compressive strengths (Table C.1) the sample coefficient of skewness is – 0.12. For the traffic flow data (Table C.2) the observations in direction 1 and 2 have a skewness coefficient of 1.54 and 2.25 respectively. The coefficients are positive and that show that both distributions are skewed to the right.

In a similar way the sample coefficient of *kurtosis* κ is defined as:

$$\kappa = \frac{1}{n} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{s^4} \tag{C.5}$$

which is a measure of how closely the data are distributed around the mode (peakedness). Typically one would compare the sample coefficient of kurtosis to that of a normal distribution (introduced in Module D), which is equal to 3.0. The kurtosis for the concrete cube compressive strength (Table C.1) is evaluated as equal to 2.23, i.e. the considered data set is less peaked than the normal distribution. For the traffic flow data (Table C.2) it is equal to 5.48 and 7.44 for direction 1 and 2 respectively.

Measures of Correlation

Observations are often made of two characteristics simultaneously as shown in Figure C.1 where pairs of data observed simultaneously are plotted jointly along the *x*-axis and the *y*-axis (this representation is also called a two-dimensional *scatter diagram* as outlined in Section C.3.).

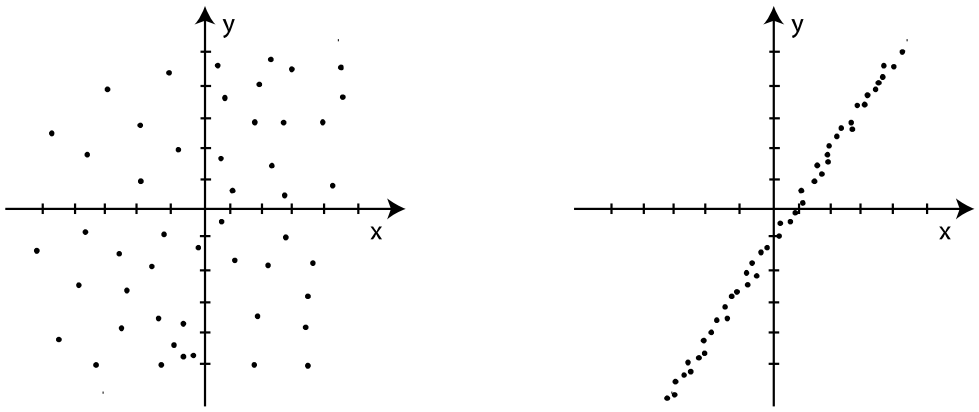


Figure C.1: Two examples of paired data sets.

As a characteristic indicating the tendency toward high-high pairings and low-low pairings, i.e. a measure of the *correlation* between the observed data sets, the *sample covariance* s_{XY} is useful, and is defined as:

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \tag{C.6}$$

The sample covariance has the property that, if there is a tendency in the data set that the values of x_i and y_i are both higher than $\overline{x}$ and $\overline{y}$ at the same time, and the trend is linear, then most of the terms in the sum will be positive and the sample covariance will be positive. The other way around will result in a negative sample covariance. Such behaviours are referred to as correlation.

In the scatter diagram to the left in Figure C.1 there appears to be only little correlation between the observed data pairs whereas the opposite is evident in the example to the right.

The sample covariance may be normalised in respect to the sample standard deviations of the individual data sets s_X and s_Y and the result is called the sample *correlation coefficient* r_{XY} defined as:

$$r_{XY} = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_X s_Y} \tag{C.7}$$

The sample correlation coefficient has the property that it is limited to the interval $-1 \le r_{XY} \le 1$ and the extreme values of the interval are only achieved in case the data pairs are perfectly correlated, implying that the points on the scatter diagram lie on a straight line. For the example shown in Figure C.1 there is almost zero correlation at the left hand side and almost full positive correlation at the right hand side.

Considering the observations of traffic flow data from Table C.2 a two-dimensional scatter plot can be produced as shown in Figure C.2.

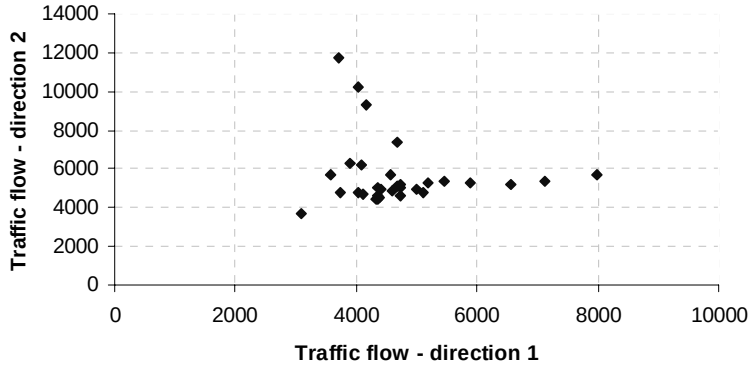


Figure C.2: 2-dimensional scatter plot of traffic flow data in two directions.

From Figure C.2 it is seen that there appears to be no simple relation between the observed traffic flows in the two directions. The correlation coefficient may be calculated using Equation (C.7) to be equal to -0.14 which confirms a very low correlation of the traffic flow in the two directions.

C.3 Graphical Representations

Graphical representations provide a convenient and strong basis for assessing data and to communicate these to other persons. There exist a relatively large number of different possible graphical representations of data, of which some are better suited than others depending on the purpose of the representations. Some are better for representing the characteristics of data sets containing observations of one characteristic, like e.g. the concrete compressive strength and others are better for representing the characteristics of two or more data sets (e.g. the simultaneously observed traffic flows). In the following, the most frequently applied graphical representations are introduced and discussed with the help of examples.

One-Dimensional Scatter Diagrams

The simplest graphical representation is the *scatter diagram* which provides a means to represent observations contained in one or more data sets. The scatter diagram may be constructed by plotting the observed values of the data set along an axis labelled according to the scale of the observations. In a *one-dimensional scatter diagram* the minimum and maximum values of the data set can be readily observed. Furthermore, as long as the number of data is not very large, the central value of the observed data may be observed directly from the plot. In the case where a data set contains a large number of data, some of these may be overlapping and this makes it difficult to distinguish the individual observations. In such cases it may be beneficial to apply another graphical representation such as histograms, as described subsequently.

Consider again the traffic flow data from Table C.2. For each of the two directions a one-dimensional scatter diagram can be produced by plotting the data along one axis. In Figure

C.3 the resulting scatter diagram is shown for direction 1. In the same manner the data for the other direction may be plotted (see also Figure C.4). It can be seen from Figure C.3 that the lowest value of the data lies close to 3000 while the highest lies close to 8000. Moreover, a high concentration of observations is observed in the range 4000 to 5000, indicating that the central value of this data is in that range.

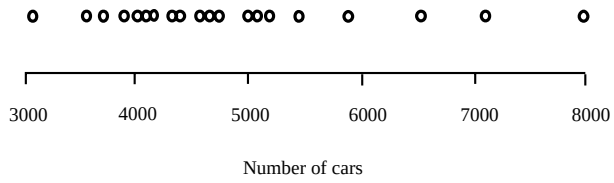


Figure C.3: One-dimensional scatter diagram of the traffic flow in the Gotthard tunnel (direction 1).

The traffic flow for both directions is plotted in a single one-dimensional scatter plot in Figure C.4. It can be seen from Figure C.4 that the lowest value of the traffic flow in direction 2 is larger than in direction 1. It is also observed that there is a significant difference between the largest values for direction 2 and direction 1; while most data for direction 1 concentrate around a value of 4000 cars per day, for direction 2 the corresponding value lies closer to 5000 cars per day.

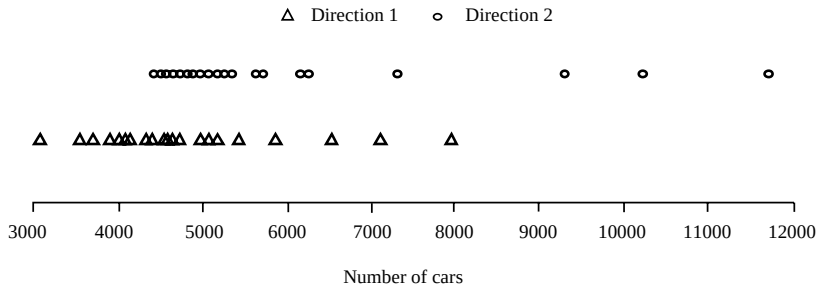


Figure C.4: One-dimensional scatter diagram of the traffic flow in the Gotthard tunnel – comparison of two data sets.

In Figure C.3 and Figure C.4 some of the observations are overlapping and this reduces the clarity of the scatter plot. This problem can be circumvented by projecting the observations vertically on a *y*-axis by allocating an integer random number j, for $j = 1$ to n (where n is the number of measurements), to each observation in the unordered data set and then by plotting j against the observations. To keep the display nearly one-dimensional, the range of the *y*-axis should be kept small compared to the range of the *x*-axis.

An example can be seen in Figure C.5 where the traffic flow observations for direction 1 have been plotted as described above. It can be seen that the observations are easily distinguished and overlaps have been almost eliminated.

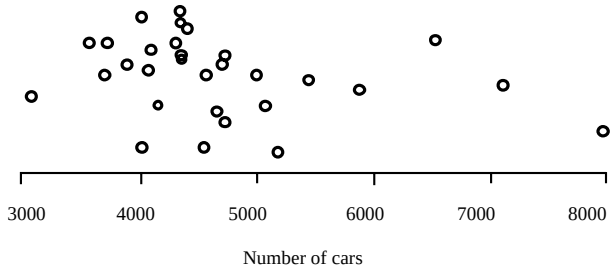


Figure C.5: One-dimensional scatter plot of the traffic flow in the Gotthard tunnel with a vertical projection of the data (direction 1).

Consider now another data set corresponding to the concrete cube compressive strength measurements from Table C.1. The corresponding one-dimensional scatter diagram is given in Figure C.6. It can be seen that the data are more widely distributed and there are not many overlaps.

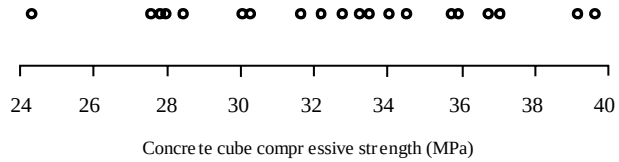


Figure C.6: One-dimensional scatter plot of the concrete cube compressive strength data.

Histograms

A frequently applied graphical representation of data sets is the *histogram.* Consider again as an example the traffic flow data from Table C.2 for direction 2. The data are further processed and the observed number of cars is subdivided into intervals, see Table C.3. For each interval the mid point is determined and the number of observations within each interval is counted. Thereafter the *frequencies* of the measurements within each interval are evaluated as the number of observations within one interval divided by the total number of observations. The *cumulative frequencies* are estimated by summing up the frequencies for each interval in increasing order. This is a common way to estimate the cumulative frequencies especially in cases where the exact observations are not known but instead the frequency of observations within an interval is known. In the following for illustration purposes the cumulative frequencies are estimated from available observations. However when observations are readily available the cumulative frequency plot can be replaced by a plot similar to a quantile plot (see section Quantile plots in the following) but a slightly different representation. Figure C.7 and Figure C.8 show the graphical representation of the processed data of Table C.3.

Interval (Number of cars x 10^2 /day)	Interval Midpoint (Number of cars x 10^2 /day)	Number of observations	Frequency (%)	Cumulative frequency
35-40	37.5	1	3.2258	0.0323
40-45	42.5	2	6.4516	0.0968
45-50	47.5	10	32.2581	0.4194
50-55	52.5	9	29.0323	0.7097
55-60	57.5	3	9.6774	0.8065
60-65	62.5	2	6.4516	0.8710
65-70	67.5	0	0.0000	0.8710
70-75	72.5	1	3.2258	0.9032
75-80	77.5	0	0.0000	0.9032
80-85	82.5	0	0.0000	0.9032
85-90	87.5	0	0.0000	0.9032
90-95	92.5	1	3.2258	0.9355
95-100	97.5	0	0.0000	0.9355
100-105	102.5	1	3.2258	0.9677
105-110	107.5	0	0.0000	0.9677
110-115	112.5	0	0.0000	0.9677
115-120	117.5	1	3.2258	1.0000

Table C.3: Summary of the observed traffic flow in the Gotthard tunnel (direction 2).

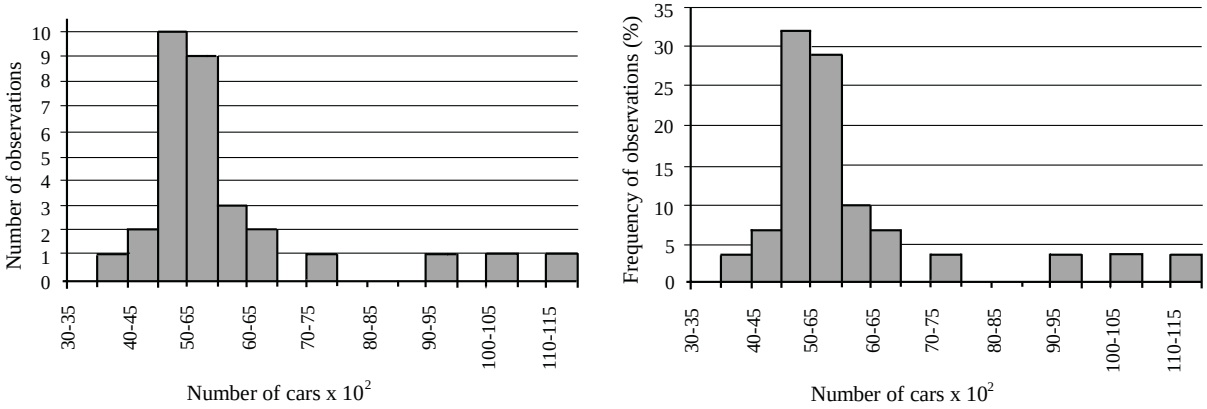


Figure C.7: Histogram and frequency distribution representations of the observed traffic flow in the Gotthard tunnel (direction 2).

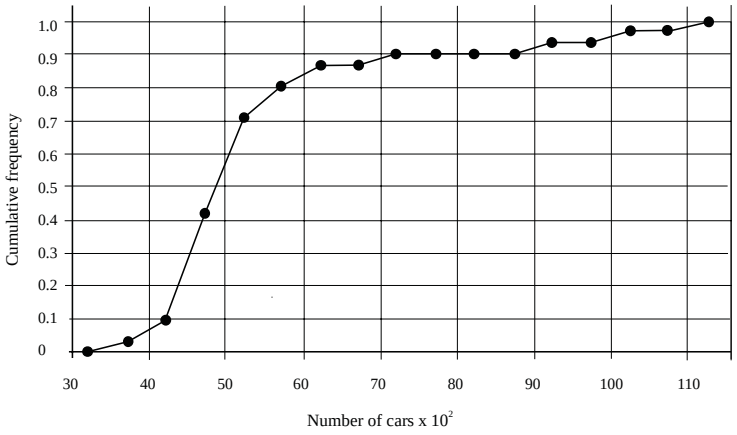


Figure C.8: Cumulative frequency plot of the observed traffic flow in the Gotthard tunnel (direction 2).

It has to be noted that a histogram may reduce the information provided by the data examined. The interval width plays an important role for the resolution of the representation of the observations. However, there are no general guidelines concerning the choice of the interval width. In most applications the goal is to identify an interval which with a sufficient resolution can represent the observations and this may comprise an iterative process where several different subdivisions are applied and the results are evaluated. In Benjamin and Cornell (1971) it is suggested to subdivide the interval between the maximum and minimum value into k intervals where k is given by:

$$k = 1 + 3.3\log(n) \tag{C.8}$$

where n is the number of data points in the data set. Using the above formula for the observations in Table C.2 k equals 5.92. By rounding up, 6 intervals should have been applied for the subdivision of observations while as shown in Table C.3 the number of intervals used is equal to 17. Figure C.9 illustrates the frequency distribution of the traffic flow data using the 6 intervals given in Table C.4.

Interval (Number of cars x 10^2 /day)	Interval Midpoint (Number of cars x 10^2 /day)	Number of observations	Frequency (%)	Cumulative frequency
35-50	42.5	13	41.9355	0.4194
50-65	57.5	14	45.1613	0.8710
65-80	72.5	1	3.2258	0.9032
80-95	87.5	1	3.2258	0.9355
95-110	102.5	1	3.2258	0.9677
110-125	117.5	1	3.2258	1.0000

Table C.4 Summary of the observed traffic flow in the Gotthard tunnel (direction 2) using 6 intervals.

Comparing with the frequency distribution in Figure C.7 it can be seen that using a smaller number of intervals the resolution of the graphical representation is significantly reduced.

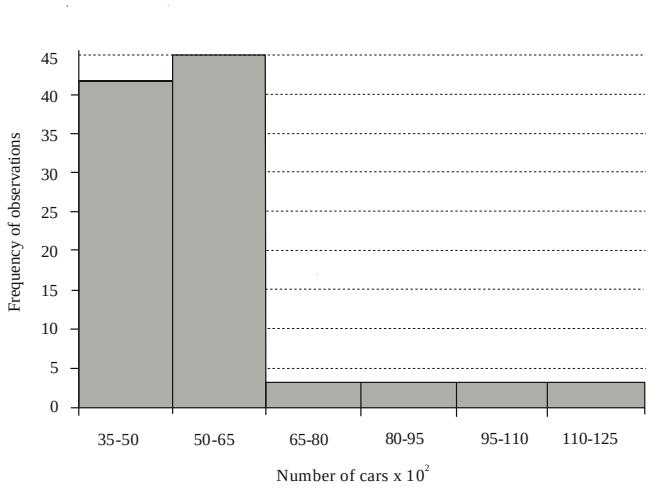


Figure C.9: Frequency distribution representation of the observed traffic flow in the Gotthard tunnel (direction 2) using a small number of intervals.

For the concrete compressive strength observations the application of Equation (C.8) seems to work well. Equation (C.8) in this case gives a value of $k = 5.29$ and by rounding up 6 intervals should be used for this data set. The data tabularised according to the result of Equation (C.8) are given in Table C.5.

Interval	Midpoint	Number of observations	Frequency [%]	Cumulative frequency
23-26	24.5	1	5	0.05
26-29	27.5	4	20	0.25
29-32	30.5	3	15	0.40
32-35	33.5	6	30	0.70
35-38	36.5	4	20	0.90
38-41	39.5	2	10	1.00

Table C.5: Summary of the observed concrete cube compressive strength measurements.

Figure C.10 and Figure C.11 show the graphical representation of the processed data of

Table C.5. It is seen from Figure C.10 that the rule implied from Equation (C.8) works fine and the resulting frequency distribution provides a good resolution of the observations.

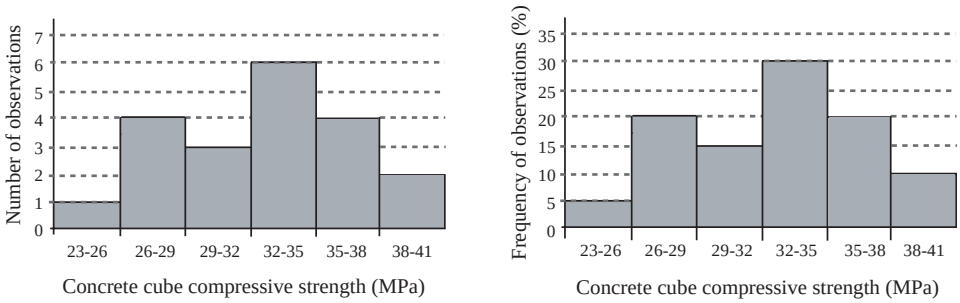


Figure C.10: Histogram and frequency distribution representations of the observed concrete cube compressive strength.

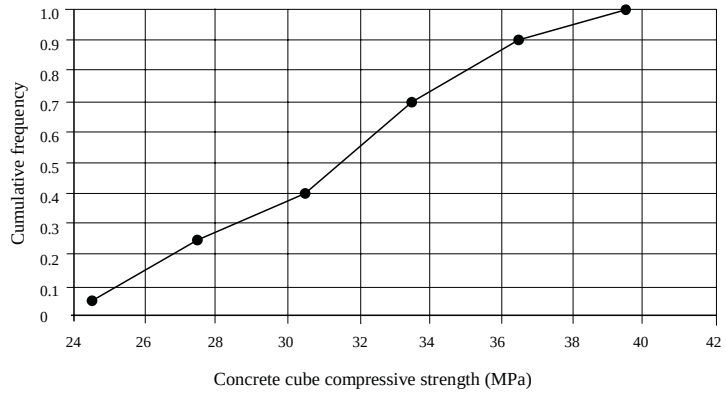


Figure C.11: Cumulative frequency plot of the observed concrete cube compressive strength.

Quantile Plots

Quantile plots are graphical representations containing information that is similar to the cumulative frequency plots introduced above. A quantile is related to a given percentage, and

e.g. the 0.65 quantile of a given data set of observations correspond to the observation for which 65% of all observations in the data set have smaller values. The 0.75 quantile is also denoted the *upper* quartile (see also the Tukey box plots in the next section) while the 0.25 quantile is denoted the *lower quartile*. The median thus equals the 0.5 quantile.

In order to construct a quantile plot the observations in the data set are arranged in ascending order. The quantile Q_i corresponding to a given observation x_i^O in the ordered data set is given by:

$$Q_i = \frac{i}{n+1} \tag{C.9}$$

As an example consider the traffic flow data from Table C.2. In Table C.6 the data are ordered in ascending order and the corresponding quantile values are shown. The median (i.e. the 0.5 quantile) has been highlighted.

i	Direction 1 Ordered x_i^O	Direction 2 Ordered x_i^O	Q_i
1	3087	3677	0.0313
2	3578	4453	0.0625
3	3710	4480	0.0938
4	3737	4560	0.1250
5	3906	4635	0.1563
6	4029	4648	0.1875
7	4041	4672	0.2188
8	4085	4757	0.2500
9	4103	4791	0.2813
10	4164	4815	0.3125
11	4323	4880	0.3438
12	4359	4928	0.3750
13	4366	4946	0.4063
14	4368	5005	0.4375
15	4371	5013	0.4688
16	4419	5100	0.5000
17	4563	5220	0.5313
18	4588	5235	0.5625
19	4664	5281	0.5938
20	4667	5318	0.6250
21	4727	5398	0.6563
22	4739	5401	0.6875
23	4741	5679	0.7188
24	5001	5688	0.7500
25	5098	5729	0.7813
26	5193	6183	0.8125
27	5457	6308	0.8438
28	5892	7357	0.8750
29	6551	9323	0.9063
30	7118	10256	0.9375
31	7974	11748	0.9688

Table C.6: Quantile values of the traffic flow observations in the Gotthard tunnel.

As mentioned in section “Histograms” when the observations are known it is preferable to use their quantiles to represent the cumulative distribution instead of the frequency of

observations within interval, as was the case in section “Histograms”. So for example in the case of the traffic flow data for direction 2 (Table C.4) the cumulative distribution can be plotted the data values and the respective quantiles. Similarly the cumulative distribution of the concrete cube compressive strength data can be plotted using the respective quantiles (Table C.7). Figure C.12 illustrates the two above mentioned cumulative distribution plots. The median of the data set can be directly read from such a representation by finding the value that corresponds to the 0.5 quantile.

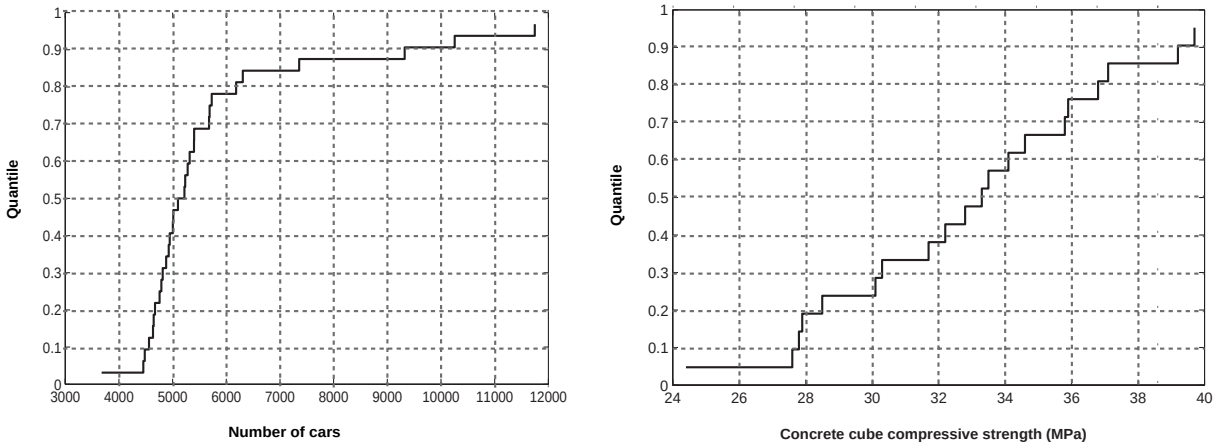


Figure C.12: Cumulative distribution plot of the traffic flow data of direction 2 (left) and the concrete cube compressive strength (right).

Coming back to the quantile plots concept, in Figure C.13 the quantile plots for the traffic flow data for both directions are illustrated. In order to enable the comparison of the data these have been plotted on the same scale. It can be seen that the quantiles for the data in direction 2 are slightly higher than the corresponding ones for direction 1. For direction 1 the median (the 0.5 quantile) is close to 4500 (the real value read from Table C.6 is 4419). The corresponding value in direction 2 is slightly higher than 5000 (5100 read from Table C.6). The approximate values for the *upper and lower quantiles* may also be observed from the quantile plots. Thus for example the lower quartile (0.25 quantile) in direction 1 is approximately equal to 4000 while the upper quartile (0.75 quantile) is close to 5000.

The slope of the quantile plot indicates the concentration of the data; a high slope corresponds to a low concentration and a low slope to a high concentration. The highest local concentration occurs when there are many observations with exactly the same value and this appears on the quantile plot by a horizontal series of points. For direction 1 the slope is quite small up to about the 0.7 quantile. Thereafter the slope increases and thus the concentration of the data is smaller. This matches the information provided by the one-dimensional scatter plot in Figure C.3. The 0.7 quantile corresponds to a value close to 5000. It can be seen from Figure C.3 that for larger observed traffic flow values the concentration of the observations decreases.

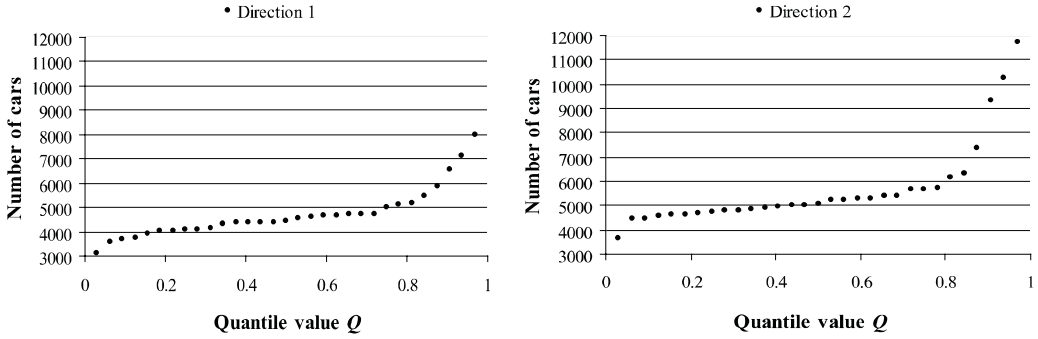


Figure C.12: Quantile plots of the observed traffic flow in the Gotthard tunnel.

Quantile plots may also provide information regarding the symmetry of data. If the observations in the data set are symmetrically dispersed around the median then the shape of the quantile plot in the upper half is a double mirrored image of shape from the lower half. From Figure C.12 it can be seen that for both directions the data are not symmetric and that for direction 2 the asymmetry is more pronounced.

Following the same procedure as described above, the concrete cube compressive strength data are plotted, Figure C.13, against the respective quantile values, see also Table C.7. It can be seen that the quantile plot has an almost constant slope over the whole range of observations.

i	Ordered x_i^O	Q_i
1	24.4	0.048
2	27.6	0.095
3	27.8	0.143
4	27.9	0.190
5	28.5	0.238
6	30.1	0.286
7	30.3	0.333
8	31.7	0.381
9	32.2	0.429
10	32.8	0.476
11	33.3	0.524
12	33.5	0.571
13	34.1	0.619
14	34.6	0.667
15	35.8	0.714
16	35.9	0.762
17	36.8	0.810
18	37.1	0.857
19	39.2	0.905
20	39.7	0.952

Table C.7: Quantile values of the observed concrete cube compressive strength [MPa].

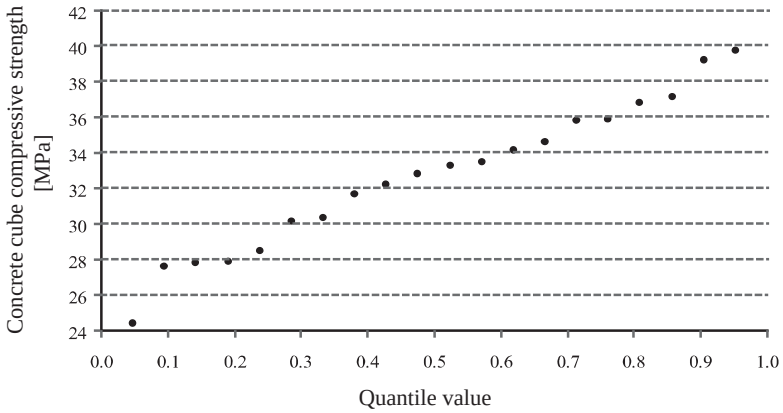


Figure C.13: Quantile plots of the observed concrete cube compressive strength.

From Table C.7 it can be seen that no observation corresponds directly to the median of the data set. In general the evaluation of a quantile which does not correspond to a given observation must be based on an interpolation. This may be performed by first calculating the hypothetical υ-th observation x_υ^O corresponding to a given quantile Q_υ from:

$$\frac{\upsilon}{n+1} = Q_\upsilon \tag{C.10}$$

Solving Equation (C.10) in regard to υ yields:

$$\upsilon = nQ_\upsilon + Q_\upsilon \tag{C.11}$$

If υ is an integer, the υ-th observation x_υ^O exists and corresponds to Q_υ. If υ is not an integer it will have a value consisting of an integer part say k and a fractional part say p. The Q_υ-quantile x_υ^O using interpolation is given as:

$$x_\upsilon^O = (1-p)x_k^O + px_{k+1}^O \tag{C.12}$$

For example in the case of the concrete cube compressive strength data (Table C.7) looking for the upper quartile (0.75 quantile) gives a value of υ equal to:

$\upsilon = 20 \cdot 0.75 + 0.75 = 15.75$

Thus $k = 15$ and $p = 0.75$

Therefore based on Equation (C.12) the 0.75 quantile is

$0.25x_{15}^O + 0.75x_{16}^O = 0.25 \cdot 35.8 + 0.75 \cdot 35.9 = 35.875$.

Tukey Box Plots

Tukey box plots provide information about several sample characteristics of the observations contained in a data set, see Figure C.14.

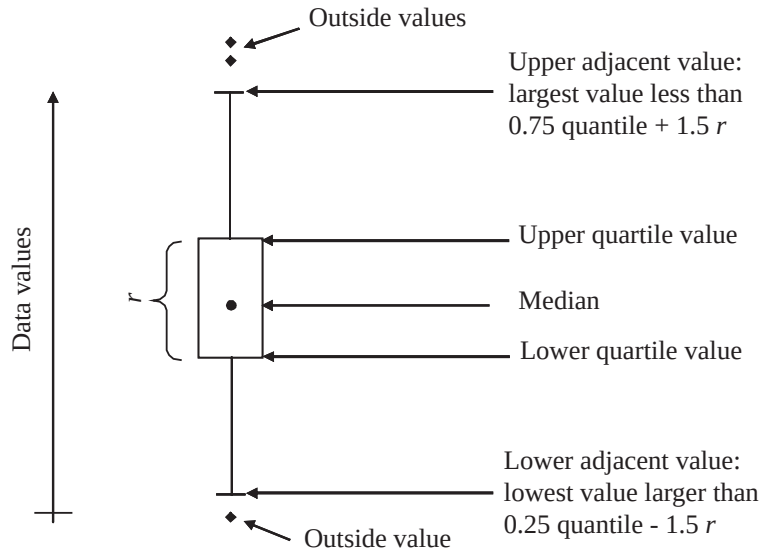


Figure C.14: Tukey box plot with indication of the characteristics of the data set.

The median is typically represented by a circle or a horizontal line within the box. The upper and lower sides of the box indicate the values of the upper and the lower quartiles, respectively. The distance between these quartiles is called the *interquartile range*, r ; 50% of the data are located within this range. A large interquartile range indicates that the observations are widely dispersed around the median and vice versa.

Another feature of the Tukey box plot is the so called *adjacent values*. The *upper adjacent value* is defined as the largest observation less than or equal to the upper quartile plus 1.5 r . The *lower adjacent value* is defined as the smallest observation greater than or equal to the lower quartile minus 1.5 r . If an observation has a value outside the adjacent values, the observation is called an *outside value* and is shown in the box plot by a single point.

In Table C.8 the sample characteristics needed to construct the Tukey box plots (Figure C.15) for the traffic flow data are given.

Statistic	Direction 1	Direction 2
Lower adjacent value	3087	3677
Lower quartile	4085	4757
Median	4419	5100
Upper quartile	5001	5688
Upper adjacent value	5892	6308
Outside values	6551	7357
	7118	9323
	7974	10256
		11748

Table C.8: Sample characteristics for the Tukey box plot for the traffic flow data in the Gotthard tunnel (Table C.2).

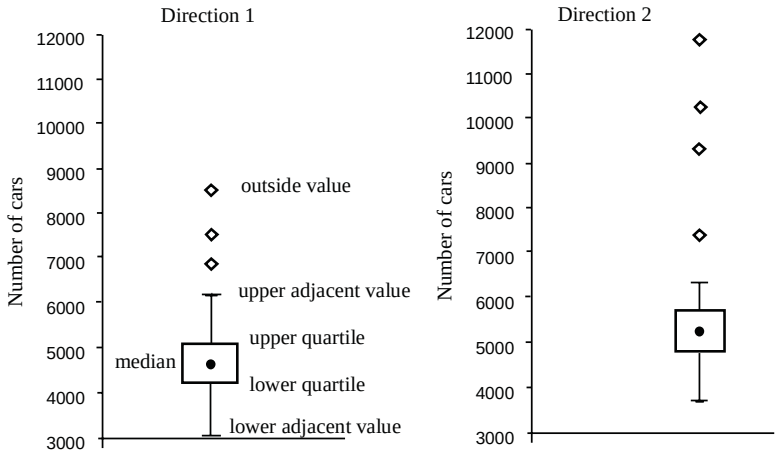


Figure C.15: Tukey box plots of the traffic flow data in the Gotthard tunnel.

The symmetry of the observations represented in a Tukey box plot may be partially assessed. From Figure C.15 can be seen that in direction 1 the observations with values in the lower and the upper range are more symmetric than the ones in direction 2. It is seen that the observed values of traffic flow in direction 2 are systematically larger than for direction 1.

In Figure C.16 the Tukey box plot for the concrete cube compressive strength data is given based on the evaluation of the respective sample statistics, see Table C.9. For this set of data there are no outside values as the upper adjacent value is the maximum value of the data and the lower adjacent value corresponds to the lower value of the data.

Statistic	Value
Lower quartile	29.30
Lower adjacent value	24.40
Median	33.05
Upper adjacent value	39.70
Upper quartile	35.85

Table C.9: Statistics for the Tukey box plot for the concrete cube compressive strength data [MPa] (Table C.1).

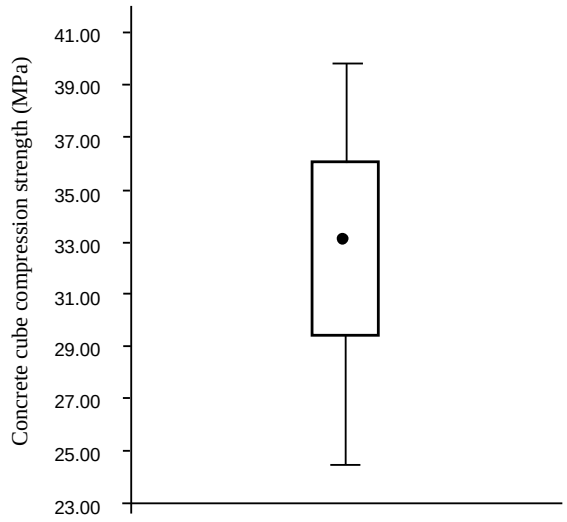


Figure C.16: Tukey box plot of the concrete cube compressive strength data [MPa].

Q-Q Plots and Tukey Mean-Difference Plot

Quantile-quantile plots (or in short Q-Q plots) provide an efficient means of comparing observations from different data sets. For example in the case of the traffic flow data a comparison may be made between the number of cars in one direction and the number of cars in another direction. To do so, the corresponding quantiles may be compared, i.e. the 0.25 quantile of the observations for direction 1 with the 0.25 quantile of the observations for direction 2 etc. For this purpose the corresponding quantiles are plotted against each other in a Q-Q plot.

The data sets as given in Table C.6 contain the same number of observations and therefore their Q-Q plot may be made by just plotting the evaluated quantiles against each other, see Figure C.17.

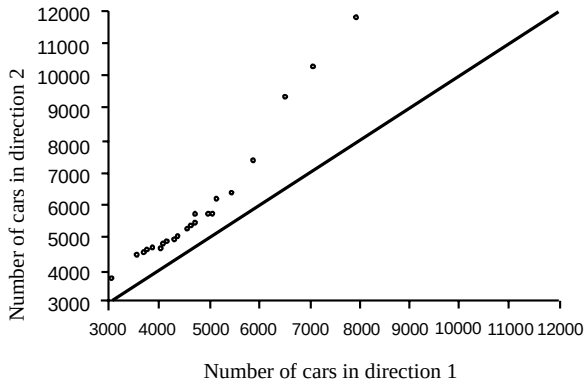


Figure C.17: Q-Q plot of the traffic flow observations in the two directions.

If the data sets compared do not have the same number of observations then the quantiles for the observations of one data set are evaluated first and subsequently the corresponding quantiles for the other data set are established by interpolation. From Figure C.17 it is seen that the traffic flow is higher over the full range of observations for direction 2 as compared to direction 1. If the Q-Q plot would result in a line close to the $y = x$ then the data would have nearly identical distributions.

Another graphical representation that facilitates the comparison of the observations contained in two different data sets is the *Tukey mean-difference plot*. Here $y_i - x_i$ is plotted against $(y_i + x_i)/2$, where y and x are the observations of the data sets being compared. The evaluation of the means and differences for the traffic flow data is provided in Table C. 10.

Direction 1	Direction 2	y_i-x_i	$(y_i+x_i)/2$
3087	3677	590	3382.0
3578	4453	875	4015.5
3710	4480	770	4095.0
3737	4560	823	4148.5
3906	4635	729	4270.5
4029	4648	619	4338.5
4041	4672	631	4356.5
4085	4757	672	4421.0
4103	4791	688	4447.0
4164	4815	651	4489.5
4323	4880	557	4601.5
4359	4928	569	4643.5
4366	4946	580	4656.0
4368	5005	637	4686.5
4371	5013	642	4692.0
4419	5100	681	4759.5
4563	5220	657	4891.5
4588	5235	647	4911.5
4664	5281	617	4972.5
4667	5318	651	4992.5
4727	5398	671	5062.5
4739	5401	662	5070.0
4741	5679	938	5210.0
5001	5688	687	5344.5
5098	5729	631	5413.5
5193	6183	990	5688.0
5457	6308	851	5882.5
5892	7357	1465	6624.5
6551	9323	2772	7937.0
7118	10256	3138	8687.0
7974	11748	3774	9861.0

Table C. 10: Values for the Tukey mean-difference plot of the traffic flow data in the Gotthard tunnel.

In Figure C.18 the Tukey mean-difference plot is given for the traffic flow data.

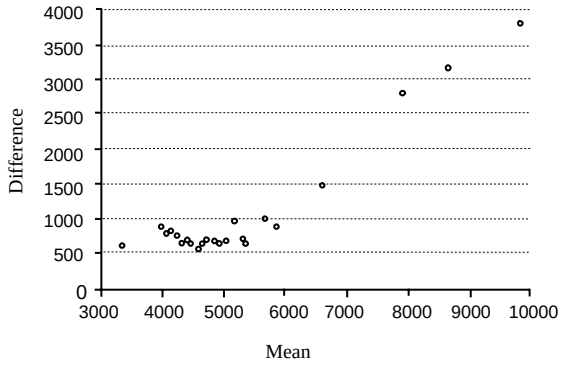


Figure C.18: Tukey mean-difference plot for the traffic flow data.

The Tukey mean-difference plot indicates that there is a systematic difference of 600 cars per day for traffic situations corresponding to a mean traffic flow of up to 6000 cars per day. Thereafter the differences in the two directions seem to be proportional in the mean traffic flow.

Self Assessment Questions/ Exercises

C.1 What is the purpose of descriptive statistics?

C.2 Within which interval the coefficient of correlation of two data sets may lie? What do the extreme values of the interval express?

C.3 What is the role of the interval width chosen for building up a histogram for the representation of a data set?

C.4 Which characteristics of a data set can be represented with a Tukey box plot?

C.5 Which means do Q-Q plots provide?

C.6 Provide a rough estimate of the correlation coefficient of the data sets plotted in the following figure.

A $r_{XY} \approx$ ☐

B $r_{XY} \approx$ ☐

C $r_{XY} \approx$ ☐

D $r_{XY} \approx$ ☐

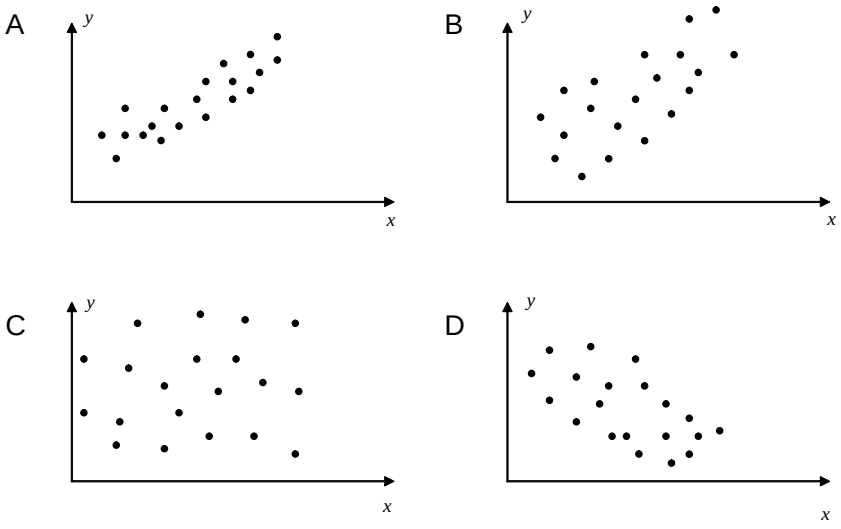


Figure C.19: Plotted data sets.

C.7 A number of statistical terms are shown in the following table. Check if the terms have something to do with (a) location parameter, (b) dispersion parameter or (c) none of the above.

	a	b	c
Mean	☐	☐	☐
Quartile	☐	☐	☐
Sample size	☐	☐	☐
Median	☐	☐	☐
Standard deviation	☐	☐	☐
Coefficient of variation	☐	☐	☐

C.8 Measurements were taken of the concrete cover depth of a bridge column. The histogram of the measured values has been plotted in Figure C.20.

If X represents the random variable for the concrete cover depth which one(s) of the following statement(s) is(are) correct?

The sample mean, $\bar{x}$, is equal to 0.16 cm. ☐

The sample mean, $\bar{x}$, is equal to 15 cm. ☐

The mode of the data set is equal to 15 cm. ☐

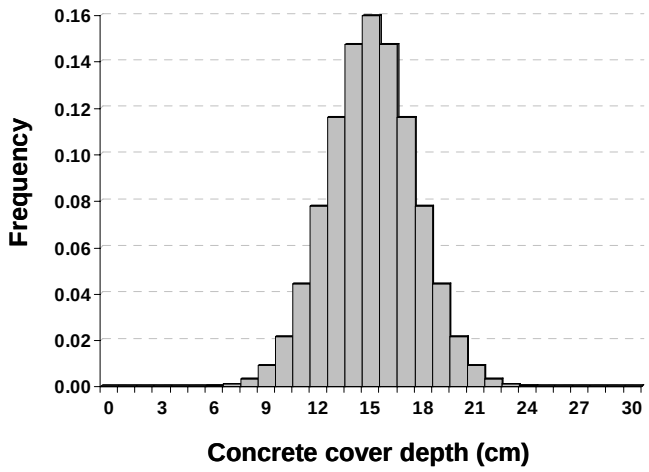


Figure C.20: Histogram of concrete cover depth measurements.

C.9 Which one(s) of the following are features of a symmetrical probability density function?

The variance is equal to the coefficient of variation. ☐

The mode is equal to the median. ☐

The skewness is equal to zero. ☐

None of the above. ☐

MODULE D – UNCERTAINTY MODELLING

4th Lecture

Aim of the present lecture

The aim of the present lecture is to provide a fundamental understanding of uncertainty and how this affects engineering decision making. Furthermore, random variables are introduced and it is explained how they may be characterized depending on the given situation.

On the basis of the lecture it is expected that the students should acquire knowledge and skills in regard to:

- Why do uncertainties influence engineering problems and decision making?
- Which are the principally different types of uncertainties?
- Why is it useful to differentiate between different types of uncertainties?
- Which types of uncertainties can be reduced?
- In what way may uncertainties depend on time?
- In what way might scale influence uncertainties?
- What is a random variable and how may it be characterized?
- How are cumulative distribution and probability density functions related?
- What is a discrete and what is a continuous probability distribution?
- How are the moments of a random variable defined?
- How is the expectation operation defined?

D.1 Introduction

A central role for engineers is to provide basis for decision making in regard to the cost efficient safeguarding of personnel, environment and assets in situations where uncertainties are at hand. A classical example is the decision problem of choosing the height of a dike. The risk of dike flooding can be reduced by increasing the height of the dike; however, due to the inherent natural variability in the water level a certain probability of dike flooding in a given reference period will always remain. Risk assessment within the theoretical framework of decision analysis (introduced in Lecture 13) can help us in deciding on the optimal dike height by weighing the benefits of reduced dike flooding risks with the costs of increasing the dike height. However, a prerequisite for the risk assessment is that the means for assessing the probability of dike flooding are established, and this in turn requires that a probabilistic model for the future water level is available.

D.2 Uncertainties in Engineering Problems

For the purpose of discussing the phenomenon uncertainty in more detail let us initially assume that the universe is deterministic and that our knowledge about the universe is perfect. This implies that it is possible by means of e.g. a set of exact equation systems and known boundary conditions by means of analysis to achieve perfect knowledge about any state, quantity or characteristic which otherwise cannot be directly observed or has yet not taken place. In principle following this line of reasoning the future as well as the past would be known or assessable with certainty. Considering the dike flooding problem it would thus be possible to assess the exact number of floods which would occur in a given reference period (the frequency of floods) for a given dike height and an optimal decision can be achieved by *cost benefit analysis*.

Whether the universe is deterministic or not is a rather deep philosophical question. Despite the obviously challenging aspects of this question its answer is, however, not a prerequisite for purposes of engineering decision making, the simple reason being that even though the universe would be deterministic our knowledge about it is still in part highly incomplete and/or uncertain.

In engineering decision analysis subject to uncertainties such as *Quantitative Risk Analysis* (QRA) and *Structural Reliability Analysis* (SRA) a commonly accepted view angle is that *uncertainties* should be interpreted and differentiated in regard to their type and origin. In this way it has become standard to differentiate between uncertainties due to *inherent natural variability*, *model uncertainties* and *statistical uncertainties*. Whereas the first mentioned type of uncertainty is often denoted *aleatory* (or Type 1) uncertainty, the two latter are referred to as *epistemic* (or Type 2) uncertainties. Without further discussion here it is just stated that in principle all prevailing types of uncertainties should be taken into account in engineering decision analysis within the framework of *Bayesian probability theory*.

Considering again the dike example it can be imagined that an *engineering model* might be formulated where future extreme water levels are predicted in terms of a regression of

previously observed annual extremes. In this case the uncertainty due to inherent natural variability would be the uncertainty associated with the annual extreme water level. The model chosen for the annual extreme water level events would by itself introduce model uncertainties and the parameters of the model would introduce statistical uncertainties as their estimation would be based on a limited number of observed annual extremes. Finally, the extrapolation of the annual extreme model to extremes over longer periods of time would introduce additional model uncertainties. The uncertainty associated with the future extreme water level is thus composed as illustrated in Figure D.1. Whereas the so-called inherent natural variability is often understood as the uncertainty caused by the fact that the universe is not deterministic it may also be interpreted simply as the uncertainty which cannot be reduced by means of collection of additional information. It is seen that this definition implies that the amount of uncertainty due to inherent natural variability depends on the models applied in the formulation of the engineering problem. Presuming that a refinement of models corresponds to looking more detailed at the problem at hand one could say that the uncertainty structure influencing a problem is scale dependent.

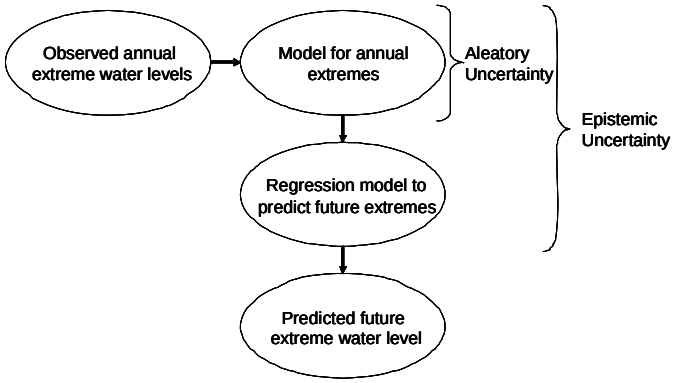


Figure D.1: Illustration of uncertainty composition in a typical engineering problem.

Having formulated a model for the prediction of future extreme water levels and taking into account the various prevailing types of uncertainties the probability of flooding within a given reference period can be assessed and just as in the case of a deterministic and perfectly known universe the optimum dike height can be decided, based on a cost benefit assessment.

It is interesting to notice that the type of uncertainty associated with the state of knowledge has a time dependency. Following Figure D.2 it is possible to observe an uncertain phenomenon when it has occurred. In principle, if the observation is perfect without any errors the knowledge about the phenomenon is perfect. The modelling of the same phenomenon in the future, however, is uncertain as this involves models subject to natural variability, model uncertainty and statistical uncertainty. Often but not always the models available tend to lose their precision rather fast so that phenomena lying just a few days or weeks ahead can be predicted only with significant uncertainty. An extreme example of this concerns the prediction of the weather.

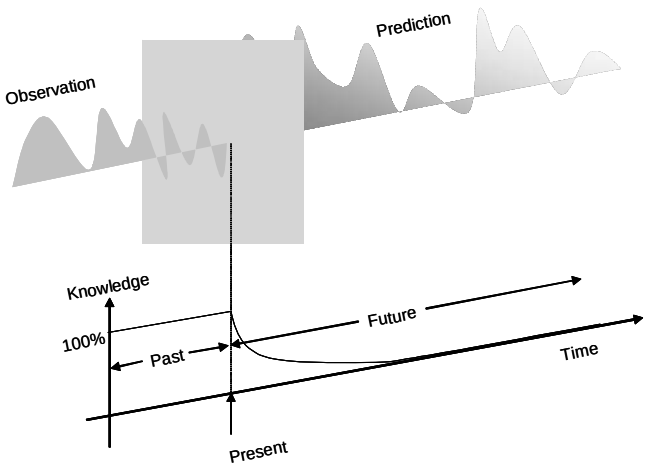


Figure D.2: Illustration of the time dependence of knowledge.

The above discussion shows another interesting effect, namely that the uncertainty associated with a model concerning the future transforms from a mixture of aleatory and epistemic uncertainty to a purely epistemic uncertainty when the modelled phenomenon is observed. This transition of the type of uncertainty has a significant importance because it facilitates that the uncertainty is reduced by utilization of observations - updating.

D.3 Random Variables

The performance of an engineering system, facility or installation (in the following referred to as system) may usually be modelled in mathematical physical terms in conjunction with empirical relations.

For a given set of model parameters the performance of the considered system can be determined on the basis of this model. The basic *random variables* are defined as the parameters that carry the entire uncertain input to the considered model.

The basic random variables must be able to represent all types of uncertainties that are included in the analysis. The uncertainties, which must be considered are as previously mentioned the physical uncertainty, the statistical uncertainty and the model uncertainty. The *physical uncertainties* are typically uncertainties associated with the loading environment, the geometry of the structure, the material properties and the repair qualities. The *statistical uncertainties* arise due to incomplete statistical information e.g. due to a small number of materials tests. Finally, the model uncertainties must be considered to take into account the uncertainty associated with the idealised mathematical descriptions used to approximate the actual physical behaviour of the structure.

Modern methods of reliability and risk analysis allow for a very general representation of these uncertainties ranging from non-stationary stochastic processes and fields to time-invariant random variables, see e.g. Melchers (1987). In most cases it is sufficient to model the uncertain quantities by random variables with given cumulative distribution functions and distribution parameters estimated on basis of statistical and/or subjective

information. Therefore the following is concerned with a basic description of the characteristics of random variables.

Cumulative Distribution and Probability Density Functions

A random variable, which can take on any value, is called a *continuous random variable*. The probability that such a random variable takes on a specific value is zero. The probability that a continuous random variable, X, is less than or equal to a value, x, is given by the *cumulative distribution function*:

$$F_X(x) = P(X \leq x) \tag{D.1}$$

In general capital letters denote a random variable and small letters denote an outcome or realization of a random variable. An example of a continuous cumulative distribution function is illustrated in Figure D.3.

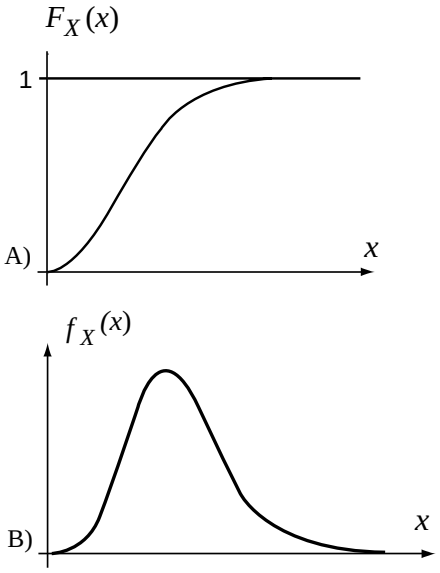


Figure D.3: Illustration of A) a cumulative distribution function and B) a probability density function for a continuous random variable.

For continuous random variables the *probability density function* is given by:

$$f_X(x) = \frac{\partial F_X(x)}{\partial x} \tag{D.2}$$

The probability of an outcome in the interval $[x; x+dx]$ where dx is small, is given by $P(x \in [x; x+dx]) = f_X(x)dx$.

Random variables with a finite or infinite countable sample space are called *discrete random variables*. For discrete random variables the cumulative distribution function is given as:

$$P_X(x) = \sum_{x_i < x} p_X(x_i) \tag{D.3}$$

where $p_X(x_i)$ is the probability density function given as:

$$p_X(x_i) = P(X = x_i) \tag{D.4}$$

A discrete cumulative distribution function and probability density function is illustrated in Figure D.4.

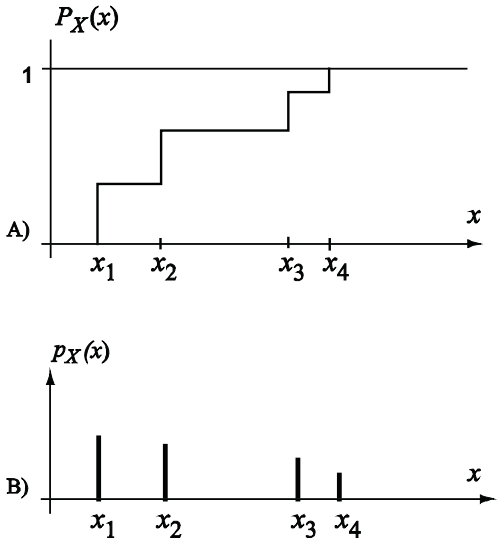


Figure D.4: Illustration of A) a cumulative distribution function and B) a probability density function for a discrete random variable.

Moments of Random Variables and the Expectation Operator

Probability distributions may be defined in terms of their *parameters* or *moments*. Often cumulative distribution functions and probability density functions are written as $F_X(x;\mathbf{p})$ and $f_X(x;\mathbf{p})$ respectively to indicate the parameters $\mathbf{p}$ (or moments) defining the functions.

The *i*'th moment m_i of a continuous random variable is defined by:

$$m_i = \int_{-\infty}^{\infty} x^i \, f_X(x) dx \tag{D.5}$$

and for a discrete random variable by:

$$m_i = \sum_{j=1}^{n} x_j^i \, p_X(x_j) \tag{D.6}$$

The mean (or *expected value*) of continuous and discrete random variables, X, are defined accordingly as the *first moment*, i.e.:

$$\mu_X = E[X] = \int_{-\infty}^{\infty} x\, f_X(x)\, dx \tag{D.7}$$

$$\mu_X = E[X] = \sum_{j=1}^{n} x_j\, p_X(x_j) \tag{D.8}$$

where $E[\cdot]$ denotes the expectation operator.

Similarly the *standard deviation*, σ_X, is defined through the *second central moment*, i.e. for continuous random variables as:

$$\sigma_X{}^2 = Var[X] = E\left[(X - \mu_X)^2\right] = \int_{-\infty}^{\infty} (x - \mu_X)^2\, f_X(x) dx \tag{D.9}$$

and for discrete random variables as:

$$\sigma_X^2 = Var[X] = \sum_{j=1}^{n} (x_j - \mu_X)^2\, p_X(x_j) \tag{D.10}$$

where $Var[X]$ denotes the *variance* of X.

The ratio between the standard deviation σ_X and the expected value μ_X of a random variable X is denoted the *coefficient of variation* $CoV[X]$ and is given by:

$$CoV[X] = \frac{\sigma_X}{\mu_X} \tag{D.11}$$

The coefficient of variation provides a useful descriptor of the variability of a random variable around its expected value.

Example D.1 – Uniform distribution

As an example consider a continuous random variable with a uniform (constant) probability density function in the interval $[a;b]$ as illustrated in Figure D.5.

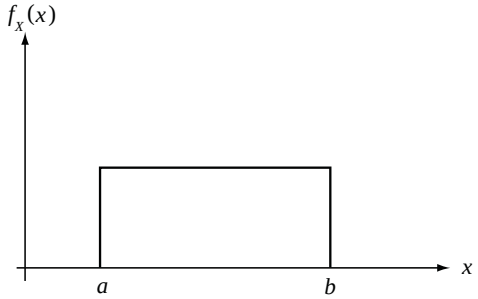


Figure D.5: Continuous random variable with a uniform density function.

The probability density function for a *uniformly distributed random variable* X is easily seen to be:

$$f_X(x) = \begin{cases} 0, & x < a \\ \dfrac{1}{b-a}, & a \le x \le b \\ 0, & x > b \end{cases} \tag{D.12}$$

remembering that the area under the probability density function must integrate to 1. In Equation (D.12) a and b are the parameters of the probability density function.

The cumulative distribution function for a uniformly distributed random variable X is thus:

$$F_X(x) = \begin{cases} 0, & x < a \\ \displaystyle\int_a^x f_X(y)dy = \int_a^x \frac{1}{b-a}dy = \frac{(x-a)}{(b-a)}, & a \le x \le b \\ 1, & x > b \end{cases} \tag{D.13}$$

The first moment i.e. the mean value (see Equation (D.5)) of a continuous random variable X with uniform distribution is thus:

$$\begin{aligned} \mu_X &= E[X] = \int_a^b x\ f_X(x)dx = \int_a^b \frac{x}{b-a}dx = \frac{x^2}{2(b-a)}\Bigg|_a^b \\ &= \frac{(b+a)}{2} \end{aligned} \tag{D.14}$$

and the standard deviation σ_X (see Equation (D.9)) is given through the second central moment:

$$\begin{aligned} \sigma_X^2 &= E\left[(X-\mu)^2\right] = \int_a^b (x-\mu)^2\ f_X(x)\,dx = \int_a^b \frac{(x-\mu)^2}{(b-a)}dx = \frac{\frac{1}{3}x^3 - x^2\mu + x\mu^2}{(b-a)}\Bigg|_a^b \\ &= \frac{1}{12}(b-a)^2 \end{aligned} \tag{D.15}$$

5th Lecture

Aim of the present lecture

The aim of the present lecture is to introduce the properties of the main characteristics of vectors of random variables and how to assess these. Furthermore, it is described how probabilistic characterizations of functions of random variables can be established.

On the basis of the lecture it is expected that the students should acquire knowledge and skills in regard to:

- How may the expectation operation be performed on a linear combination of random variables?
- How may the expectation operation be performed on a linear combination of functions of random variables?
- Which rule applies for the expectation operation of functions of random variables?
- What is the relation between the expectation operation and the variance operation?
- Which are the properties of the expectation and the variance operator?
- What is a random vector and what is a joint moment?
- How is the covariance between two random variables defined?
- How is the correlation coefficient defined and what information does it contain?
- What is a marginal probability distribution?
- What is a conditional probability distribution?
- How may the probability distribution for the sum of two random variables be established?
- How may the probability distribution for a function of random variables be established?

Properties of the Expectation Operator

It is useful to note that the *expectation operation* possesses the following properties, where a, b and c are constants and X is a random variable:

$$\begin{aligned} &E[c] = c \\ &E[cX] = cE[X] \\ &E[a + bX] = a + bE[X] \\ &E[g_1(X) + g_2(X)] = E[g_1(X)] + E[g_2(X)] \end{aligned} \tag{D.16}$$

The implication of the last equation is that expectation, like differentiation or integration, is a linear operation. This linearity property is useful since it can be used, for example, to find the following formula for the variance of a random variable X in terms of more easily calculated quantities:

$$\begin{aligned} Var[X] &= E\left[(X - \mu_X)^2\right] = E\left[X^2 + \mu_X^2 - 2\mu_X X\right] = \mu_X^2 + E\left[X^2\right] - 2\mu_X E[X] \\ &= \mu_X^2 + E\left[X^2\right] - 2\mu_X^2 = E\left[X^2\right] - \mu_X^2 \end{aligned} \tag{D.17}$$

By application of Equation (D.17) the following properties of the *variance operator* $Var[\cdot]$ can easily be derived:

$$\begin{aligned} &Var[c] = 0 \\ &Var[cX] = c^2 Var[X] \\ &Var[a + bX] = b^2 Var[X] \end{aligned} \tag{D.18}$$

where a, b and c are constants and X is a random variable.

From Equation (D.17) it is furthermore seen that in general it is $E[g(X)] \neq g(E[X])$. In fact for convex functions $g(x)$ it can be shown that the following inequality is valid (*Jensen's inequality*):

$$E[g(X)] \geq g(E[X]) \tag{D.19}$$

where the equality holds if $g(X)$ is linear.

Whether the cumulative distribution and density function are defined by their moments or by parameters is a matter of convenience and it is generally possible to establish the one from the other.

Random Vectors and Joint Moments

If a *n*-dimensional vector of continuous random variables $\mathbf{X} = (X_1, X_2, ..., X_n)^T$, is considered the *joint cumulative distribution function* is given by:

$$F_{\mathbf{X}}(\mathbf{x}) = P(X_1 \le x_1 \cap X_2 \le x_2 \cap \ldots \cap X_n \le x_n) \tag{D.20}$$

and the *joint probability density function* is:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\partial^n}{\partial x_1 \partial x_2 \ldots \partial x_n} F_{\mathbf{X}}(\mathbf{x}) \tag{D.21}$$

The *covariance* $C_{X_i X_j}$ between X_i and X_j is defined by:

$$C_{X_i X_j} = E\left[(X_i - \mu_{X_i})(X_j - \mu_{X_j})\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x_i - \mu_{X_i})(x_j - \mu_{X_j})\, f_{X_i X_j}(x_i, x_j)\, dx_i dx_j \tag{D.22}$$

and is also called the *joint central moment* between the variables X_i and X_j.

The covariance expresses the dependence between two variables. It is evident that $C_{X_i X_i} = Var[X_i]$. On the basis of the covariance the *correlation coefficient* is defined by:

$$\rho_{X_i X_j} = \frac{C_{X_i X_j}}{\sigma_{X_i} \sigma_{X_j}} \tag{D.23}$$

It is seen that $\rho_{X_i X_i} = 1$. The correlation coefficients can only take values in the interval $[-1;1]$. A negative correlation coefficient between two random variables implies that if the outcome of one variable is large compared to its mean value the outcome of the other variable is likely to be small compared to its mean value. A positive correlation coefficient between two variables implies that if the outcome of one variable is large compared to its mean value the outcome of the other variable is also likely to be large compared to its mean value. If two variables are independent their correlation coefficient is zero and the *joint density function* is the product of the 1-dimensional density functions. In many cases it is possible to obtain a sufficiently accurate approximation to the *n-dimensional cumulative distribution function* from the 1-dimensional distribution functions of the *n* variables and their parameters, and the correlation coefficients.

Finally using Equations (D.17), (D.18) and (D.22) it can be shown that the expected value $E[Y]$ and the variance $Var[Y]$, where Y is a linear function of the random vector $\mathbf{X} = (X_1, X_2, ..., X_n)^T$ i.e.:

$$Y = a_0 + \sum_{i=1}^{n} a_i X_i \tag{D.24}$$

are given by:

$$E[Y] = a_0 + \sum_{i=1}^{n} a_i E[X_i]$$
$$Var[Y] = \sum_{i=1}^{n} a_i^2 Var[X_i] + \sum_{\substack{i,j=1 \\ i \neq j}}^{n} a_i a_j C_{X_i X_j} \quad \text{(D.25)}$$

Conditional Distributions and Conditional Moments

The *conditional probability density function* for the random variable X_1, conditional on the outcome of the random variable X_2 is denoted $f_{X_1|X_2}(x_1|x_2)$ and defined by:

$$f_{X_1|X_2}(x_1|x_2) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)} \quad \text{(D.26)}$$

in accordance with the definition of conditional probability given previously.

As for the case when probabilities of events were considered two random variables X_1 and X_2 are said to be independent when there is:

$$f_{X_1|X_2}(x_1|x_2) = f_{X_1}(x_1) \quad \text{(D.27)}$$

By integration of Equation (D.26) the conditional cumulative distribution $F_{X_1|X_2}(x_1|x_2)$ is obtained:

$$F_{X_1|X_2}(x_1|x_2) = \frac{\int_{-\infty}^{x_1} f_{X_1,X_2}(z, x_2)dz}{f_{X_2}(x_2)} \quad \text{(D.28)}$$

and finally by integration of (D.28) weighed with the probability density function of X_2, i.e. $f_{X_2}(x_2)$ the unconditional cumulative distribution $F_{X_1}(x_1)$ is achieved by the *total probability theorem*:

$$F_{X_1}(x_1) = \int_{-\infty}^{\infty} F_{X_1|X_2}(x_1|x_2) f_{X_2}(x_2)dx_2 \quad \text{(D.29)}$$

The *conditional moments* of jointly distributed continuous random variables follow straightforwardly from Equation (D.7) by use of Equation (D.27) and the conditional expected value $\mu_{X_1|X_2}$ and the conditional variance of e.g. the jointly distributed random variables X_1 given X_2 are evaluated by:

$$\mu_{X_1|X_2} = E[X_1|X_2 = x_2] = \int_{-\infty}^{\infty} x_1 f_{X_1|X_2}(x_1|x_2)dx_1 \quad \text{(D.30)}$$

$$Var_{X_1|X_2} = E[(X_1 - \mu_{X_1|X_2})^2 \,|\, X_2 = x_2] = \int_{-\infty}^{\infty} (x_1 - \mu_{X_1|X_2})^2 \, f_{X_1|X_2}(x_1 \,|\, x_2) dx_1$$

The Probability Distribution for the Sum of two Random Variables

Based on the result in Equation (D.26) the probability density function for the random variable $Y = X_1 + X_2$ may be derived for a given joint probability density function $f_{X_1,X_2}(x_1, x_2)$.

First the conditional probability density function of Y given $X_1 = x_1$ is considered i.e.:

$$Y = x_1 + X_2 \tag{D.31}$$

where the conditional probability density function of X_2 given $X_1 = x_1$ is:

$$f_{X_2|X_1}(x_2 \,|\, x_1) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_1}(x_1)} \tag{D.32}$$

thus the probability density function for Y given $X_1 = x_1$ can be written as:

$$f_{Y|X_1}(y \,|\, x_1) = f_{X_2|X_1}(y - x_1 \,|\, x_1) \tag{D.33}$$

and the joint probability density function for $f_{Y,X_1}(y, x_1)$:

$$f_{Y,X_1}(y, x_1) = f_{X_2|X_1}(y - x_1 \,|\, x_1) \, f_{X_1}(x_1) = f_{X_2,X_1}(y - x_1, x_1) \tag{D.34}$$

from which one can get the so-called *marginal probability density* function of Y by integrating out over the domain of x_1 i.e.:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_2,X_1}(y - x_1, x_1) dx_1 \tag{D.35}$$

For the special case where the variables X_1 and X_2 are independent, Equation (D.35) can be written in the form of a so-called *convolution integral*:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_2}(y - x_1) \, f_{X_1}(x_1) dx_1 \tag{D.36}$$

The Probability Distribution for Functions of Random Variables

In some cases it is interesting to be able to derive the cumulative distribution function $F_Y(y)$ for a random variable Y which is given as a function of another random variable X i.e. $Y = g(X)$, with given cumulative distribution function $F_X(x)$. Under the condition that the function $g(x)$ is monotonically increasing and furthermore, represents a one-to-one mapping of x into y, a realization of Y is only smaller than y_0 if correspondingly the realization of

X is smaller than x_0 which in turn is given by $x_0 = g^{-1}(y_0)$. In this case the cumulative distribution function $F_Y(y)$ can be readily determined by:

$$F_Y(y) = P(Y \le y) = P(X \le g^{-1}(y)) \tag{D.37}$$

which is also written as:

$$F_Y(y) = F_X(g^{-1}(y)) \tag{D.38}$$

In accordance with Equation (D.2) the probability density function $f_Y(y)$ is simply given by:

$$f_Y(y) = \frac{\partial F_X(g^{-1}(y))}{\partial y} \tag{D.39}$$

which immediately leads to:

$$f_Y(y) = \frac{\partial g^{-1}(y)}{\partial y} f_X(g^{-1}(y)) \tag{D.40}$$

and:

$$f_Y(y) = \frac{\partial x}{\partial y} f_X(x) \tag{D.41}$$

It is noticed that the application of Equations (D.40) and (D.41) necessitates that $g(x)$ is at least one time differentiable in regard to x.

Now if the function $g(x)$ instead of being monotonically increasing is monotonically decreasing a realization of Y smaller that y_0 corresponds to a realization of X larger than x_0 in which case it is necessary to change the sign of the derivative $\partial x/\partial y$ in Equation (D.41). Generally for monotonically increasing or decreasing one-to-one functions $g(x)$ there is:

$$f_Y(y) = \left|\frac{\partial x}{\partial y}\right| f_X(x) \tag{D.42}$$

As shown in e.g. Thoft-Christensen and Baker (1982) the relationship given in Equation (D.34) can be generalized to consider the case of jointly distributed random variables.

Consider the random vector $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)^T$ with individual components given as one-to-one mapping monotonically increasing or decreasing functions $g_i, i = 1, 2, ..., n$ of the components of the random vector $\mathbf{X} = (X_1, X_2, ..., X_n)^T$ as:

$$Y_i = g_i(\mathbf{X}) \tag{D.43}$$

then there is:

$$f_{\mathbf{Y}}(\mathbf{y}) = \left|\mathbf{J}\right| f_{\mathbf{X}}(\mathbf{x}) \tag{D.44}$$

where $\left|\mathbf{J}\right|$ is the numerical value of the determinant of $\mathbf{J}$ given by:

$$\mathbf{J} = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_n} \end{bmatrix} \tag{D.45}$$

Finally the expected value $E[Y]$ of a function $g(\mathbf{X})$ of the random vector $\mathbf{X} = (X_1, X_2, ..., X_n)^T$ is given by:

$$E[Y] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\mathbf{x})\, f_{\mathbf{X}}(\mathbf{x}) dx_1 \cdots dx_n \tag{D.46}$$

6th Lecture

Aim of the present lecture

The aim of the present lecture is first to summarize typical probability distribution functions applied in engineering uncertainty modelling. Thereafter it is outlined how the Normal and the Lognormal probability distributions may be derived on the basis of the central limit theorem. Furthermore as an introduction on how to model uncertain phenomena with random variability over time, random sequences and their characterization are introduced.

On the basis of the lecture it is expected that the students should acquire knowledge and skills in regard to:

- What does the central limit theorem say?
- What is a standardized random variable?
- How may the variance of a linear combination of correlated Normal distributed random variables be calculated?
- How may the Lognormal distribution be derived?
- In what way may uncertain phenomena depend on “time”?
- What is a random sequence?
- What is a Bernoulli trial and what does it describe?
- For what can the Binomial distribution be used?
- What is a Geometric distribution and for what may it be applied?

Probability Density and Distribution Functions

In Table D.1 a selection of probability density and cumulative distribution functions is given with the definition of their distribution parameters and moments.

Distribution type	Parameters	Moments
Uniform, $a \le x \le b$ $f_X(x) = \frac{1}{b-a}$ $F_X(x) = \frac{x-a}{b-a}$	a b	$\mu = \frac{a+b}{2}$ $\sigma = \frac{b-a}{\sqrt{12}}$
Normal $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$ $F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right) dt$	μ $\sigma > 0$	μ σ
Shifted Lognormal, $x > \varepsilon$ $f_X(x) = \frac{1}{(x-\varepsilon)\zeta\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\ln(x-\varepsilon)-\lambda}{\zeta}\right)^2\right)$ $F_X(x) = \Phi\left(\frac{\ln(x-\varepsilon)-\lambda}{\zeta}\right)$	λ $\zeta > 0$ ε	$\mu = \varepsilon + \exp\left(\lambda + \frac{\zeta^2}{2}\right)$ $\sigma = \exp\left(\lambda + \frac{\zeta^2}{2}\right)\sqrt{\exp(\zeta^2)-1}$
Shifted Exponential, $x \ge \varepsilon$ $f_X(x) = \lambda \exp(-\lambda(x-\varepsilon))$ $F_X(x) = 1 - \exp\left(-\lambda\left(x-\varepsilon\right)\right)$	ε $\lambda > 0$	$\mu = \varepsilon + \frac{1}{\lambda}$ $\sigma = \frac{1}{\lambda}$
Gamma, $x \ge 0$ $f_X(x) = \frac{b^p}{\Gamma(p)} \exp(-bx) x^{p-1}$ $F_X(x) = \frac{\Gamma(bx, p)}{\Gamma(p)}$	$p > 0$ $b > 0$	$\mu = \frac{p}{b}$ $\sigma = \frac{\sqrt{p}}{b}$

Distribution type	Parameters	Moments
Beta, $a \le x \le b$ $f_X(x) = \frac{\Gamma(r+t)}{\Gamma(r)\Gamma(t)} \frac{(x-a)^{r-1}(b-x)^{t-1}}{(b-a)^{r+t-1}}$ $F_X(x) = \frac{\Gamma(r+t)}{\Gamma(r)\Gamma(t)} \int_a^x \frac{(u-a)^{r-1}(b-u)^{t-1}}{(b-a)^{r+t-1}} du$	a b $r > 0$ $t > 0$	$\mu = a + (b-a)\frac{r}{r+t}$ $\sigma = \frac{b-a}{r+t}\sqrt{\frac{rt}{r+t+1}}$

Table D.1: Probability distributions, Schneider (1994).

The relevance of the different distribution functions given in Table D.1 in connection with the probabilistic modelling of uncertainties in engineering risk and reliability analysis is strongly case dependent and the reader is suggested to consult the application specific literature for specific guidance. In the following, however, a brief introduction to the *central limit theorem* and the derived *Normal* and *Lognormal distributions* is given.

The Central Limit Theorem and Derived Distributions

The central limit theorem states:

The probability distribution for the sum of a number of random variables approaches the *Normal distribution* as the number becomes large.

This result, which indeed is one of the most import results in probability theory, will not be derived here but instead the in fact quite general conditions for the validity of the theorem will be outlined.

In principle the theorem is valid as long as the number of independent contributions to the sum is “large”. This implies that the sum may not be dominated by one or just a few random variables and furthermore, that the dependency between the random variables in the sum is not too strong. There is no requirement in regards to the type of distributions of the random variables entering the sum, but if the distributions are skew the number of variables in the sum, which is required for the validity of the theorem increases.

For the purpose of illustration consider the problem of assessing the accumulated error in repeated measurements. The length of a structural member is being measured using a ruler of length $2\,m$ with the smallest measuring unit equal to $1\,mm$. It is assumed that all measurements are being rounded off to the closest unit on the ruler and thus it is assumed that each measurement is subject to a measurement uncertainty which is uniformly distributed in the range $\pm 0.5\,mm$. If the length of a considered structural member is smaller or equal to $2\,m$ the length can be measured by one measurement. It is clear that in this case the measurement uncertainty is simply uniformly distributed as outlined in the above. However, if the member length is between $2\,m$ and $4\,m$ two measurements are required, if the member length is between $4\,m$ and $8\,m$ three measurements are required and so on. In Figure D.6 the

histograms of the corresponding resulting measurement errors are illustrated under the assumption that consecutive errors are independent.

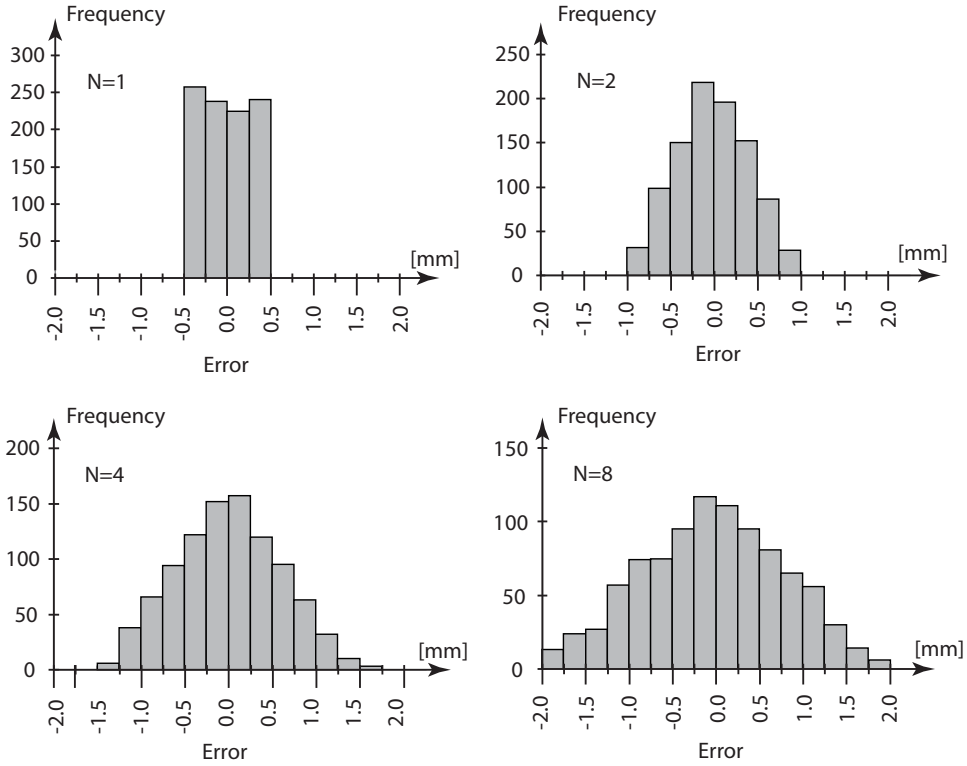


Figure D.6: Sample histograms for errors accumulated in 1, 2, 4 and 8 repeated measurements.

From Figure D.6 it is seen that whereas the sample histogram for one measurement is clearly uniform, the histogram approaches a bell shape already for four repeated measurements and for most practical purposes may be considered to be Normal distributed already for eight repeated measurements. The analytical form of the probability density function for the accumulated errors may be derived by repeated use of the result concerning the probability density function for the sum of random variables given in Equation (D.34). In Benjamin and Cornell (1971) it is heuristically shown that the analytical probability density functions has the form of a Normal distribution.

The Normal Distribution

The significant practical importance of the central limit theorem lies in the fact that even though only weak information is available regarding the number of contributions and their joint probability density function rather strong information is achieved in regard to the distribution of sum of the contributions.

The *Normal probability distribution* is thus applied very frequently in practical problems for the probabilistic modelling of uncertain phenomena which may be considered to originate from a cumulative effect of several uncertain contributions.

The Normal distribution has the property that the linear combination S of n Normal distributed random variables X_i, $i = 1, 2, ..., n$:

$$S = a_0 + \sum_{i=1}^{n} a_i X_i \tag{D.47}$$

is also Normal distributed. The distribution is said to be *closed* in respect to summation.

One special version of the Normal distribution should be mentioned, namely the *Standard Normal distribution*. In general a standardized (some times referred to as a reduced) random variable is a random variable which has been transformed such that it has an expected value equal to zero and a variance equal to one, i.e. the random variable Y defined by:

$$Y = \frac{X - \mu_X}{\sigma_X} \tag{D.48}$$

is a standardized random variable. If the random variable X follows the Normal distribution the random variable Y is standard Normal distributed. In Figure D.7 the process of standardization is illustrated.

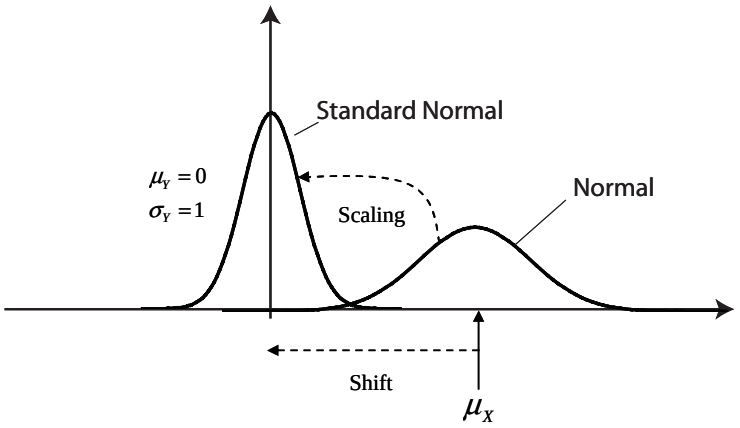


Figure D.7 Illustration of the relationship between a Normal distributed random variable and a standard Normal distributed random variable.

It is common practice to denote the cumulative distribution function for the standard Normal distribution by $\Phi(x)$ and the corresponding density function by $\varphi(x)$. These functions are broadly available in software packages such as MS Excel.

The Lognormal Distribution

A random variable Y is said to be *Lognormal distributed* if the variable $Z = \ln(Y)$ is Normal distributed. It thus follows that if an uncertain phenomenon can be assumed to originate from a multiplicative effect of several uncertain contributions then the probability distribution for the phenomenon can be assumed to be Lognormal distributed.

The Lognormal distribution has the property that if:

$$P = \prod_{i=1}^{n} Y_i^{a_i} \tag{D.49}$$

and all Y_i are independent Lognormal random variables with parameters λ_i, ζ_i and $\varepsilon_i = 0$ as given in Table 4.2 then also P is Lognormal with parameters:

$$\lambda_P = \sum_{i=1}^{n} a_i \lambda_i \tag{D.50}$$

$$\zeta_P^2 = \sum_{i=1}^{n} a_i^2 \zeta_i^2 \tag{D.51}$$

D.4 Stochastic Processes and Extremes

Random quantities may be "*time variant*" in the sense that they take on new realisations at new trials or at new times. If the new realizations of the time variant random quantity occur at discrete times and take on discrete realizations the random quantity is usually denoted a *random sequence*. Well known examples hereof are series of throws of dices - more engineering relevant examples are e.g. flooding events. If the realizations of the time variant quantity occur continuously in time and take on continuous realizations the random quantity is usually denoted a *random process* or *stochastic process*. Examples hereof are the wind velocity, wave heights, snowfall and water levels.

In some cases random sequences and random processes may be represented in a given problem context in terms of random variables e.g. for the modelling of the "*point in time*" value of the intensity of the wind velocity, or the maximum (extreme) wind velocity during one year. However, in many cases this is not possible and then it is necessary to model the uncertain phenomena by a random process. In the following first an important type of random sequence will be introduced, namely the sequence of *Bernoulli trials* from which the *Binomial distribution* has been derived. Thereafter a description of the *Poisson counting process* is given and finally the continuous Normal or *Gaussian processes* are described. It should be noted that numerous other types of random processes have been suggested in the literature of which most have been derived from the mentioned.

Random Sequences – Bernoulli Trials

A sequence of experiments with only two possible mutually exclusive outcomes is called a sequence of Bernoulli trials. Typically the two possible events of a Bernoulli trial are referred to as a *success* or a *failure.* If it is assumed that the probability of success of a Bernoulli trial is constant equal to p then the probability density of Y successes in n trials $p_Y(y)$ i.e. the *Binomial distribution* (or sometimes denoted $B(n, p)$) can be shown to be equal to:

$$p_Y(y) = \binom{n}{y} p^y (1-p)^{n-y}, \qquad y = 0,1,2...n \tag{D.52}$$

where $\binom{n}{y}$ is the so-called binomial operator defined as:

$$\binom{n}{y} = \frac{n!}{y!(n-y)!} \tag{D.53}$$

The cummulative distribution function for Y is thus given as:

$$P_Y(y) = \sum_{i=0}^{y} \binom{n}{i} p^i (1-p)^{n-i}, \qquad y = 0,1,2,...n \tag{D.54}$$

In Figure D.8 some examples of the *Binomial distribution* are shown for $n = 5$.

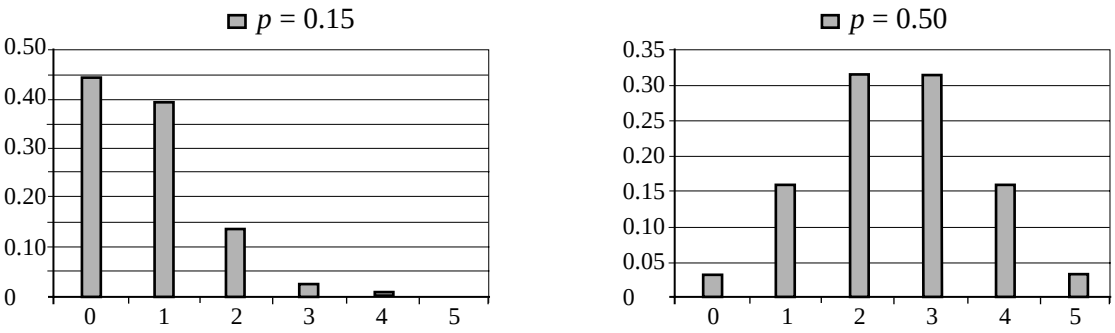


Figure D.8: Binomial distribution for *p* = 0.15 and *p* = 0.50, respectively.

The expected value and the variance of Y, i.e. $E[Y]$ and $Var[Y]$ can be shown to be given as:

$$E[Y] = np \tag{D.55}$$

$$Var[Y] = np(1-p) \tag{D.56}$$

It is often of significant interest to assess the statistical characteristics of the random "time" or random number of trials n until the first success occurs. The probability density of this event, provided that the trials are independent, is given by the so-called *Geometric distribution*:

$$p_N(n) = p(1-p)^{n-1} \tag{D.57}$$

and the corresponding cumulative distribution function by:

$$P_N(n) = \sum_{i=1}^{n} p(1-p)^{i-1} = 1-(1-p)^n \tag{D.58}$$

The mean value and the variance of the *Geometric distribution* are given by:

$$E[N] = \frac{1}{p} \tag{D.59}$$

$$Var[N] = \frac{1-p}{p^2} \tag{D.60}$$

Especially the result given in Equation (D.59) is of practical value as it gives the average "time" until success. If p is the annual probability of an extreme rainfall exceeding the capacity of a reservoir it means that in average such an event will occur with a *return period* of $1/p$ years.

7th Lecture

Aim of the present lecture

The aim of the present lecture is to provide an understanding on how to model events occurring discretely in time and to describe these probabilistically. In addition continuous random processes are introduced and their main characteristics are provided. Finally extreme events and their modelling are introduced and the concept of return period is explained. On the basis of the lecture it is expected that the students should acquire knowledge and skills in regard to:

- What is a simple Poisson process and for what may it be applied?
- Which are the properties which must be fulfilled before we can assume a Poisson process?
- What does homogeneity refer to for a Poisson process?
- According to what distribution can the time between realizations of a Poisson process be modelled?
- Which distribution does the sum of independent exponentially distributed variables follow and for what can this distribution be applied?
- What is a continuous random process?
- How to characterize a Normal process?
- What does stationarity mean and how is it defined
- What is an extreme value and what is required to model it probabilistically?
- Which are the different types of extreme value distributions?
- How are extreme value models and return periods related?

The Poisson Counting Process

The most commonly applied family of discrete processes in structural reliability are the *Poisson processes*. Due to the fact that Poisson processes have found applications in many different types of engineering problems a large number of different variants of Poisson processes has evolved. In general the process $N(t)$ denoting the number of points in the interval $[0;t[$ is called a *simple Poisson process* if it satisfies the following conditions:

- The probability of one event in the interval $[t;t+\Delta t[$ is asymptotically proportional to the length of the interval Δt.
- The probability of more than one event in the interval $[t;t+\Delta t[$ is a function of a higher order term of Δt for $\Delta t \to 0$.
- Events in disjoint intervals are mutually independent.

The Poisson process may be defined completely by its intensity $\nu(t)$:

$$\nu(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P \text{ (one event in} [t;t+\Delta t[\text{)} \tag{D.61}$$

If $\nu(t)$ is constant in time the Poisson process is said to be *homogeneous*, otherwise it is *inhomogeneous*.

In general the probability of n events in the interval $[0;t[$ of a Poisson process with *intensity* $\nu(t)$ can be shown to be given as:

$$P_n(t) = \frac{\left(\int_0^t \nu(\tau)d\tau\right)^n}{n!} \exp\left(-\int_0^t \nu(\tau)d\tau\right) \tag{D.62}$$

with mean value $E[N(t)]$ and variance $Var[N(t)]$:

$$E[N(t)] = Var[N(t)] = \int_0^t \nu(\tau)d\tau \tag{D.63}$$

The probability of no events in the interval $[0;t[$ i.e. $P_0(t)$ is especially interesting considering reliability problems. This probability may be determined directly from Equation (D.62) as:

$$P_0(t) = \exp\left(-\int_0^t \nu(\tau)d\tau\right) \tag{D.64}$$

implying that the time till and between events is *Exponential distributed*.

From Equation (D.64) the cumulative distribution function of the *waiting time* till the first event T_1, i.e. $F_{T_1}(t_1)$ may be straightforwardly derived. Recognising that the probability of $T_1 > t$ is $P_0(t)$ there is:

$$F_{T_1}(t_1) = 1 - \exp\left(-\int_0^{t_1} \nu(\tau) d\tau \right) \tag{D.65}$$

Consider now the sum of n independent and exponential distributed *waiting times* T given as:

$$T = T_1 + T_2 + ... + T_n \tag{D.66}$$

It can be shown by repeated application of the result on the probability distribution for the sum of two random variables (see Equation (D.34)) that T is *Gamma distributed*:

$$f_T(t) = \frac{\nu(\nu t)^{(n-1)} \exp(-\nu t)}{(n-1)!} \tag{D.67}$$

Continuous Random Processes

A *random process* $X(t)$ is as mentioned a random function of time meaning that for any point in time the value of $X(t)$ is a random variable. A realisation of a random process (e.g. water level variation) is illustrated in Figure D.9.

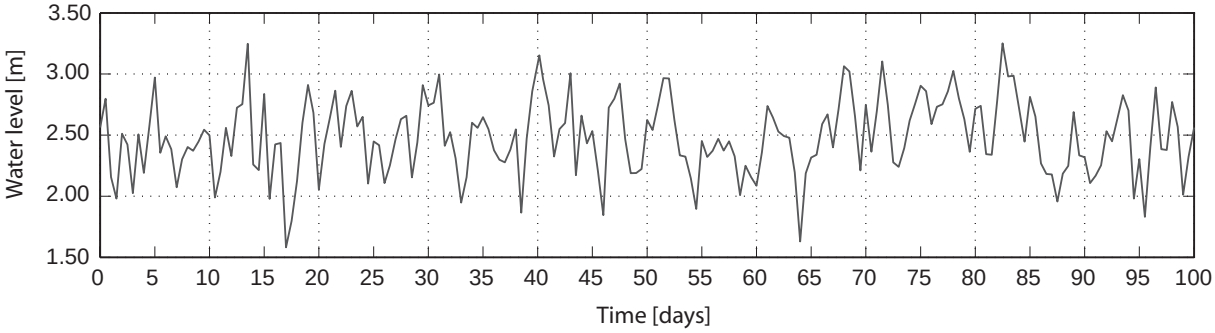


Figure D.9: Realization of the water level variation as function of time.

In accordance with the definition of the mean value of a random variable the mean value of all the possible realisations of the stochastic process at time *t* is given by:

$$\mu_X(t) = \int_{-\infty}^{\infty} x \, f_X(x;t) dx \tag{D.68}$$

The correlation between all the possible realisations at two points in time t_1 and t_2 is described through the so-called *autocorrelation function* $R_{XX}(t_1, t_2)$. Auto means that the function refers to only one realisation. The autocorrelation function is defined by:

$$R_{XX}(t_1,t_2) = E\left[X(t_1)X(t_2)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1\, x_2\; f_{XX}(x_1,x_2;t_1,t_2)dx_1dx_2 \tag{D.69}$$

The auto-covariance function is defined as:

$$\begin{aligned} C_{XX}(t_1,t_2) &= E\left[(X(t_1)-\mu_X(t_1))(X(t_2)-\mu_X(t_2))\right] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x_1-\mu_X(t_1))\;(x_2-\mu_X(t_2))\;\; f_{XX}(x_1,x_2;t_1,t_2)dx_1dx_2 \end{aligned} \tag{D.70}$$

for $t_1 = t_2 = t$ the autocovariance function becomes the *covariance function*:

$$\sigma_X^2(t) = C_{XX}(t,t) = R_{XX}(t,t) - \mu_X^2(t) \tag{D.71}$$

where $\sigma_X(t)$ is the *standard deviation function*.

The above definitions for the scalar process $X(t)$ may be extended to cover also *vector valued processes* $\mathbf{X}(t) = (X_1(t), X_2(t), ..., X_n(t))^T$ having covariance functions $C_{X_iX_j} = \mathrm{cov}\left[X_i(t_1), X_j(t_2)\right]$. For $i = j$ these become the auto-covariance functions and when $i \neq j$ these are termed the *cross-covariance functions*. Finally the *correlation function* may be defined as:

$$\rho\left[X_i(t_1), X_j(t_2)\right] = \frac{\mathrm{cov}\left[X_i(t_1), X_j(t_2)\right]}{\sigma_{X_i}(t_1)\;\sigma_{X_j}(t_2)} \tag{D.72}$$

Typically the correlation function is an exponentially decaying function in time.

Having defined the mean value function and the cross-correlation function for the stochastic process $X(t)$ the probability that the process remains within a certain safe domain D in the time interval $[0;t]$ may be evaluated by:

$$P_f(t) = 1 - P(N(t) = 0 \,|\, X(0) \in D)\;\; P(X(0) \in D) \tag{D.73}$$

where *N(t)* is the number of out-crossings of the random process out of the domain D in the time interval $[0,t]$.

Stationarity and Ergodicity

When the mean value function $\mu_X(t)$ and the autocorrelation function $R_{XX}(t)$ of a stochastic process $X(t)$ do not depend on time the process is said to be *weakly stationary*. Only if all the moments of a random process are independent of time the random process is said to be *strictly stationary*.

A consequence of stationarity is that the autocovariance functions and autocorrelation function only depend on the time difference $\tau = t_1 - t_2$. In this case Equation (D.69) may be written as:

$$R_{XX}(\tau) = E\left[X(t)X(t+\tau)\right] \tag{D.74}$$

It should be noted that for weakly stationary Normal stochastic processes the requirements for strict stationarity are automatically fulfilled as the Normal distribution function is completely defined by the first two moments.

Stationarity in principle implies that the process cannot start or stop, however, for practical purposes this requirement may be relaxed if the process is considered at a sufficient time after its start and/or before its end. Also stationarity may be assumed even for slowly varying stochastic processes if sufficiently short time intervals are considered.

If in addition to stationarity the mean value function and the autocorrelation function of a stochastic process may be defined by a time average over one realisation of the stochastic process the process is said to be *weakly ergodic*. If all moments of the process may be defined in this way the process is said to be *strictly ergodic*.

The assumption of *ergodicity* is especially important for the estimation of the statistical characteristics of stochastic processes when only one (or a few sufficiently long) realisation of the process is available. In practice ergodicity is in such cases often assumed unless of course evidence indicates the contrary.

Statistical Assessment of Extreme Values

In risk and reliability assessments *extreme values* (small and large) of random processes in a specified reference period are often of special interest. This is e.g. the case when considering the maximum sea water level, maximum wave heights, minimum ground water reservoir level, maximum wind pressures, strength of weakest link systems, maximum snow loads, etc.

For continuous time-varying loads, which can be described by a scalar, i.e. the water level or the wind pressure, one can define a number of related probability distributions. Often the simplest, namely the "*arbitrary point in time*", distribution is considered.

If $x(t^*)$ is a realisation of a single time-varying load at time t^* then $F_X(x)$ is the arbitrary point in time cumulative distribution function of $X(t)$ defined by:

$$F_X = P(X(t^*) \leq 0) \tag{D.75}$$

In Figure D.10 first observations of half yearly maximum values of wind speeds are plotted together with histograms showing the corresponding sample frequency distributions. In the figure also the equivalent presentations are provided for observations corresponding to maximums observed over periods of one and five years.

From Figure D.10 it is seen that there is a clear tendency that the mean value of the sample frequency histograms increase for increasing length of the considered period. At the same time the standard deviation is seen to be decreasing.

For practical purposes the observations of half yearly maxima may be assumed to be statistically independent and provide the basis (random "half yearly" point in time model) for the further modelling of the statistical characteristics of extremes for longer periods by extreme value considerations.

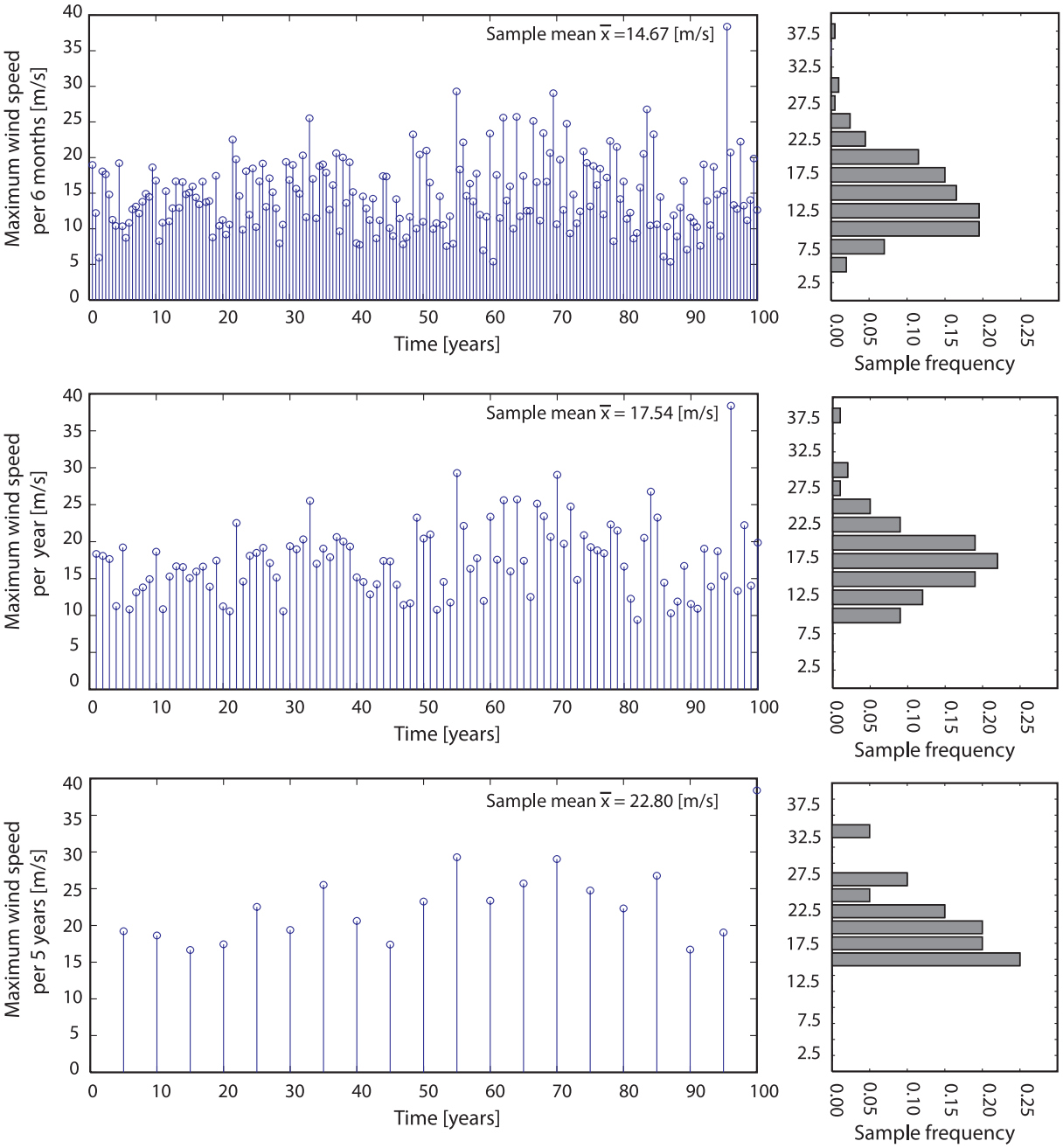


Figure D.10: Time series and corresponding sample frequency histograms of recorded half yearly, annual and five year maximum observed wind velocities.

In the following some results are given concerning the extreme events of trials of random variables and random processes, see also Madsen et al. (1986) and Benjamin and Cornell

(1971). Taking basis in the tail behaviour of cumulative distribution functions asymptotic results are given leading to the so-called *extreme value distributions*.

Extreme Value Distributions

When *extreme events* are of interest the arbitrary point in time distribution of the load variable is not of immediate relevance but rather the distribution of the maximal values of the considered quantity over a given reference period.

If the random process $X(t)$ may be assumed to be ergodic the distribution of the largest extreme in a *reference period T*, $F_{X,T}^{\max}(x)$ can be thought of as being generated by sampling values of the maximal realisation $x_{\max}$ from successive reference periods T. If the values of $x_{\max}$ are represented by the random variable Y, the cumulative distribution function $F_Y(y)$ is the cumulative distribution function of the extreme maximum realisation corresponding to the considered reference period T.

In the same way the cumulative distribution function of the largest extreme in a period of nT, $F_{X,nT}^{\max}(x)$, (with n being an integer) may be determined from the cumulative distribution function of the largest extreme in the period *T*, $F_{X,T}^{\max}(x)$, by:

$$F_{X,nT}^{\max}(x) = F_{X,T}^{\max}(x)^n \tag{D.76}$$

which follows from the multiplication law for independent events. The corresponding probability density function may be established by differentiation of Equation (D.76) yielding:

$$f_{X,nT}^{\max}(x) = n\, F_{X,T}^{\max}(x)^{n-1}\, f_{X,T}^{\max}(x) \tag{D.77}$$

In Figure D.11 the case of a Normal distribution with mean value equal to 10 and standard deviation equal to 3 is illustrated for increasing *n*.

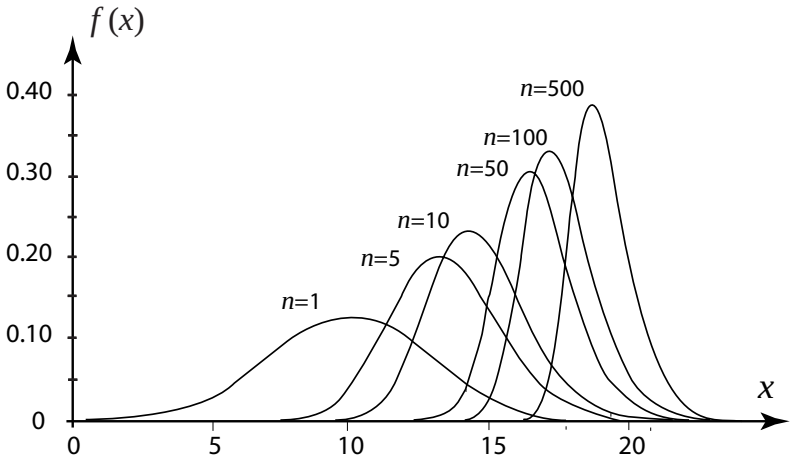


Figure D.11: Normal extreme value probability density functions.

Similarly to the derivation of Equation (D.76) the cumulative distribution function for the extreme minimum value in a considered reference period T, $F_{X,nT}^{\min}(x)$ may be found as:

$$F_{X,nT}^{\min}(x) = 1-(1-F_{X,T}^{\min}(x))^n \tag{D.78}$$

Subject to the assumption that the considered process is ergodic it can be shown that the cumulative function for an extreme event $F_{X,nT}^{\max}(x)$ converges asymptotically (as the reference period nT increases) to one of three types of *extreme value distributions*, i.e. type I, type II or type III. To which type the distribution converges depends only on the tail behaviour (upper or lower) of the considered random variable generating the extremes, i.e. $F_{X,T}^{\max}(x)$. In the following the three types or extreme value distributions will be introduced and it will be discussed under what conditions they may be assumed. In Table D.2 the definition of the extreme value probability distributions and their parameters and moments is summarised.

Type I Extreme Maximum Value Distribution – Gumbel max

For upwards unbounded distribution functions $F_X(x)$ where the upper tail falls off in an exponential manner such as it is the case for the exponential function, the Normal distribution and the Gamma distribution the cumulative distribution of extremes in the reference period T i.e. $F_{X,T}^{\max}(x)$ has the following form:

$$F_{X,T}^{\max}(x) = \exp(-\exp(-\alpha(x-u))) \tag{D.79}$$

with corresponding probability density function:

$$f_{X,T}^{\max}(x) = \alpha\exp(-\alpha(x-u)-\exp(-\alpha(x-u))) \tag{D.80}$$

which is also denoted the *Gumbel distribution* for extreme maxima. The mean value and the variance of the Gumbel distribution may be related to the parameters u and α as:

$$\begin{aligned}\mu_{X_T^{\max}} &= u+\frac{\gamma}{\alpha} = u+\frac{0.577216}{\alpha}\\ \sigma_{X_T^{\max}} &= \frac{\pi}{\alpha\sqrt{6}}\end{aligned} \tag{D.81}$$

where γ is Euler's constant.

The Gumbel distribution has the useful property that the standard deviation is independent of the considered reference period, i.e. $\sigma_{X_{nT}^{\max}} = \sigma_{X_T^{\max}}$ and that the mean value $\mu_{X_{nT}^{\max}}$ depends on n in the following simple way:

$$\mu_{X_{nT}^{\max}} = \mu_{X_T^{\max}} + \frac{\sqrt{6}}{\pi}\sigma_{X_T^{\max}}\ln(n) \tag{D.82}$$

Finally by manipulation of Equation (D.79) it can be shown, by utilising a Taylor expansion to the first order of $\ln(p)$ in $p=1$, that the characteristic value x_c corresponding to an annual exceedance probability of p and corresponding return period $T_R = 1/p$ for a Gumbel max distribution for large return periods can be written as:

$$x_c \approx u + \frac{1}{\alpha}\ln(T_R) \tag{D.83}$$

which shows that the characteristic value, a typical engineering decision parameter, increases with the logarithm of the considered return period.

Type I Extreme Minimum Value Distribution – Gumbel min

In case that the cumulative distribution function $F_X(x)$ is downwards unbounded and the lower tail falls off in an exponential manner, symmetry considerations lead to a cumulative distribution function for the extreme minimum $F_{X,T}^{\min}(x)$ within the reference period T of the following form:

$$F_{X,T}^{\min}(x) = 1 - \exp(-\exp(\alpha(x-u))) \tag{D.84}$$

with corresponding probability density function:

$$f_{X,T}^{\min}(x) = \alpha \exp(\alpha(x-u) - \exp(\alpha(x-u))) \tag{D.85}$$

which is also denoted the *Gumbel distribution* for extreme minima. The mean value and the variance of the Gumbel distribution can be related to the parameters u and α as:

$$\begin{aligned} \mu_{X_T^{\min}} &= u - \frac{\gamma}{\alpha} = u - \frac{0.577216}{\alpha} \\ \sigma_{X_T^{\min}} &= \frac{\pi}{\alpha\sqrt{6}} \end{aligned} \tag{D.86}$$

Type II Extreme Maximum Value Distribution – Frechet max

For cumulative distribution functions downwards limited at zero and upwards unlimited with a tail falling off in the form:

$$F_X(x) = 1 - \beta\left(\frac{1}{x}\right)^k \tag{D.87}$$

the cumulative distribution function of extreme maxima in the reference period T i.e. $F_{X,T}^{\max}(x)$ has the following form:

$$F_{X,T}^{\max}(x) = \exp(-\left(\frac{u}{x}\right)^{k}) \tag{D.88}$$

with corresponding probability density function:

$$f_{X,T}^{\max}(x) = \frac{k}{u}\left(\frac{u}{x}\right)^{k+1} \exp(-\left(\frac{u}{x}\right)^{k}) \tag{D.89}$$

which is also denoted the *Frechet distribution* for extreme maxima. The mean value and the variance of the Frechet distribution can be related to the parameters u and k as:

$$\begin{aligned} \mu_{X_T^{\max}} &= u\Gamma(1-\frac{1}{k}) \\ \sigma^2_{X_T^{\max}} &= u^2\left[\Gamma(1-\frac{2}{k}) - \Gamma^2(1-\frac{1}{k})\right] \end{aligned} \tag{D.90}$$

where it is noticed that the mean value only exists for $k > 1$ and the standard deviation only exist for $k > 2$. In general it can be shown that the *i*'th moment of the Frechet distribution exists only when $k > i$.

Type III Extreme Minimum Value Distribution – Weibull min

Finally in the case where the cumulative distribution function $F_X(x)$ is downwards limited at ε and the lower tail falls off towards ε in the form:

$$F(x) = c(x-\varepsilon)^k \tag{D.91}$$

leads to a cumulative distribution function for the extreme minimum $F_{X,T}^{\min}(x)$ within the reference period T of the following form:

$$F_{X,T}^{\min}(x) = 1 - \exp\left(-\left(\frac{x-\varepsilon}{u-\varepsilon}\right)^{k}\right) \tag{D.92}$$

with corresponding probability density function:

$$f_{X,T}^{\min}(x) = \frac{k}{u-\varepsilon}\left(\frac{x-\varepsilon}{u-\varepsilon}\right)^{k-1} \exp\left(-\left(\frac{x-\varepsilon}{u-\varepsilon}\right)^{k}\right) \tag{D.93}$$

which is also denoted the *Weibull distribution* for extreme minima. The mean value and the variance of the Weibull distribution can be related to the parameters u, k and ε as:

$$\begin{aligned} \mu_{X_T^{\min}} &= \varepsilon + (u-\varepsilon)\Gamma(1+\frac{1}{k}) \\ \sigma^2_{X_T^{\min}} &= (u-\varepsilon)^2\left[\Gamma(1+\frac{2}{k}) - \Gamma^2(1+\frac{1}{k})\right] \end{aligned} \tag{D.94}$$

Distribution type	Parameters	Moments
Extreme Type I Gumbel max $-\infty \le x \le \infty$ $f_X(x) = \alpha \exp(-\alpha(x-u) - \exp(-\alpha(x-u)))$ $F_X(x) = \exp(-\exp(-\alpha(x-u)))$	u $a > 0$	$\mu = u + \frac{0.577216}{\alpha}$ $\sigma = \frac{\pi}{\alpha\sqrt{6}}$
Extreme Type I Gumbel min $-\infty \le x \le \infty$ $f_X(x) = \alpha \exp(\alpha(x-u) - \exp(\alpha(x-u)))$ $F_X(x) = 1 - \exp(-\exp(\alpha(x-u)))$	u $a > 0$	$\mu = u - \frac{0.577216}{\alpha}$ $\sigma = \frac{\pi}{\alpha\sqrt{6}}$
Extreme Type II Frechet max $\varepsilon \le x \le \infty,\ \ u, k > 0$ $f_X(x) = \frac{k}{u-\varepsilon}\left(\frac{u-\varepsilon}{x-\varepsilon}\right)^{k+1} \exp\left(-\left(\frac{u-\varepsilon}{x-\varepsilon}\right)^k\right)$ $F_X(x) = \exp\left(-\left(\frac{u-\varepsilon}{x-\varepsilon}\right)^k\right)$	$u > 0$ $k > 0$ ε	$\mu = \varepsilon + (u-\varepsilon)\Gamma\left(1-\frac{1}{k}\right),\ k > 1$ $\sigma = (u-\varepsilon)$ $\sqrt{\Gamma\left(1-\frac{2}{k}\right) - \Gamma^2\left(1-\frac{1}{k}\right)},\ k > 2$
Extreme Type III Weibull min $\varepsilon \le x \le \infty,\ \ u, k > 0$ $f_X(x) = \frac{k}{u-\varepsilon}\left(\frac{x-\varepsilon}{u-\varepsilon}\right)^{k-1} \exp\left(-\left(\frac{x-\varepsilon}{u-\varepsilon}\right)^k\right)$ $F_X(x) = 1 - \exp\left(-\left(\frac{x-\varepsilon}{u-\varepsilon}\right)^k\right)$	$u > 0$ $k > 0$ ε	$\mu = \varepsilon + (u-\varepsilon)\Gamma\left(1+\frac{1}{k}\right)$ $\sigma = (u-\varepsilon)\sqrt{\Gamma\left(1+\frac{2}{k}\right) - \Gamma^2\left(1+\frac{1}{k}\right)}$

Table D.2: Probability distributions, Schneider (1994).

Return Period for Extreme Events

The *return period* T_R for an extreme event corresponding to *x* may be defined by:

$$T_R = nT = \frac{1}{(1 - F_{X,T}^{\max}(x))} \tag{D.95}$$

where T is the reference period for the cumulative distribution function of the extreme events $F_{X,T}^{\max}(x)$. If as an example the annual probability of an extreme load event is 0.02 the return period for this load event is 50 years.

Self Assessment Questions/Exercises

D.1 What types of uncertainties may be distinguished and how do these depend on the time and scale of modelling?

D.2 What is understood by the terms aleatory and epistemic uncertainties?

D.3 What is meant by the term "continuous random variable"?

D.4 Using the properties of the expectation operator how are the following notations may be rewritten? (Note that a and b are constants and X is a random variable)

a. $E[a+bX]$ b. $Var[a+bX]$

D.5 Write down the names of the axes of the probability density and the cumulative distribution functions of the random variable X illustrated in the following. Identify the locations of the mean, the mode and the median in the illustration of the probability density function Show also the value of the median in the illustration of the cumulative distribution function.

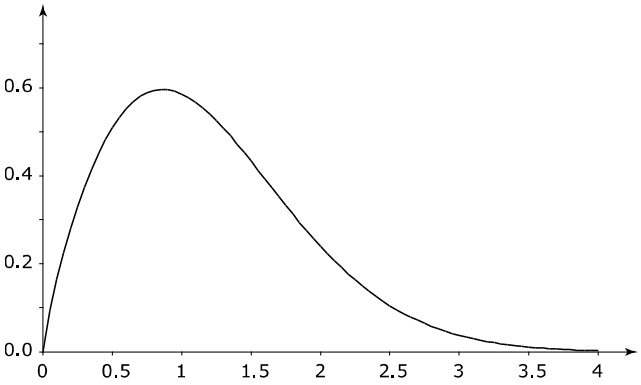


Figure D.12.: Illustration of a probability density function.

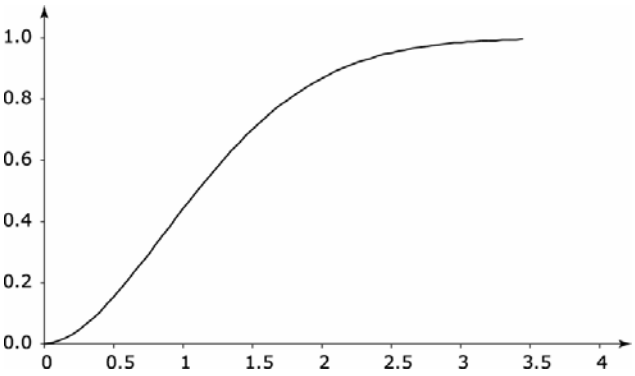


Figure D.13.: Illustration of a cumulative distribution function.

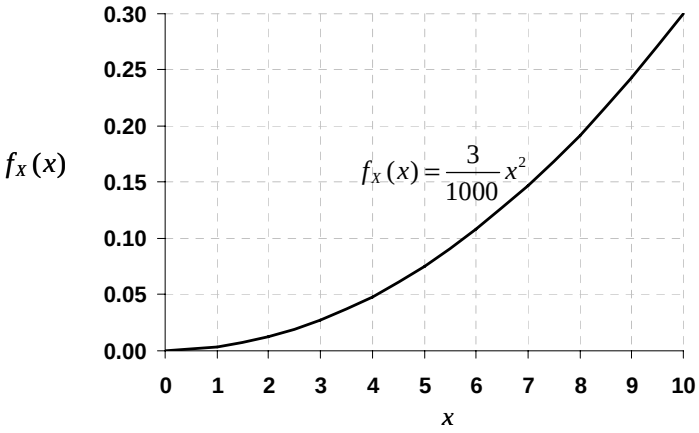
D.6 What is stated by the central limit theorem?

D.7 What is a standardized random variable and how is it defined?.

D.8 What is a Bernoulli trial and what does it describe?

D.9 What is a Poisson process and for what may it be applied?

D.10 The probability density function of a continuous random variable X, defined in the interval $[0,10]$, is illustrated in the following diagram. Calculate the probability that X may exceed the value of 5.



D.11 It is given that the operational life (until breakdown) T of a diesel engine follows an exponential distribution, $F_T(t) = 1 - e^{-\lambda t}$, with parameter λ and mean value, $\mu_T = 1/\lambda$, equal to 10 years. Calculate the probability that the engine breaks down within 2 years after placed in operation.

D.12 In a city there are on average 5 snowfalls a year. Assume that the occurrence of snowfalls follows a Poisson process. The number of snowfalls in t years, X, is described by the discrete cumulative distribution function $P(X = k) = \frac{(\nu t)^k}{k!} e^{-\nu t}$ and with annual mean rate ν. How large is the probability of no snowfall in the next year? How large is the probability of exactly 5 snowfalls in the next year?

MODULE E – ESTIMATION AND MODEL BUILDING

8th Lecture

Aim of the present lecture

The aim of the present lecture is to provide an overview of how to establish probabilistic models and to introduce the basic tools for assessing the validity of model assumptions. It is further explained how the statistical properties of sample characteristics depend on the sample size, i.e. the amount of data available and in this way the quantification of statistical uncertainty is addressed.

On the basis of the lecture it is expected that the students should acquire knowledge and skills in regard to:

- Which are the steps and constituents in establishing a probabilistic model?
- Which are the distributions function applied in sample statistics and which are the principles on the basis of which they are derived?
- What does the degree of freedom for a *t*-distributed random variable refer to?
- What distribution does a *t*-distributed random variable converge towards for increasing number of degree of freedom?
- What does it mean that the Chi-Square distribution is regenerative?
- How is the Chi-distribution related to the law of Pythagoras or the evaluation of Euclidean norms?
- How does the expected value of the sample mean depend on the number of samples?
- How does the variance of the sample mean depend on the number of samples?
- How may an unbiased estimator for the sample variance be established?
- What is a confidence interval and how can it be established?

E.1 Introduction

An important task in risk and reliability analysis is to establish probabilistic models for the further statistical treatment of uncertain variables.

In the literature a large number of probabilistic models for load and resistance variables may be found. E.g. in the Probabilistic Model Code by the Joint Committee on Structural Safety (JCSS, 2001) where probabilistic models may be found for the description of the strength and stiffness characteristics of steel and concrete materials, soil characteristics and for the description of load and load effects covering many engineering application areas. However it is not always the case that an appropriate probabilistic model for the considered problem is available. Moreover in other engineering fields, such as in environmental engineering and hydrology, standardization of the probabilistic modelling is far less progressed. In such situations it is necessary that methodologies and tools are readily available for the statistical assessment of *frequentistic information* (e.g. observations and test results) and the formulation of *probabilistic models* of uncertain variables.

In practice two situations may thus be distinguished namely, the situation where a new probabilistic model is formulated from the very beginning and the situation where an already existing probabilistic model is updated on the basis of new information, e.g. observations or experimental results. The formulation of probabilistic models may be based on data (frequentistic information) alone, but most often data are not available to the extent where this is possible. In such cases it is usually possible to base the model building on physical arguments, experience and judgement (*subjective information*). If also some data are available the subjective information may be combined with the frequentistic information and the resulting probabilistic model is in effect of a Bayesian nature.

It should be emphasised that on the one hand the probabilistic model should aim for simplicity and, on the other hand the model should be accurate enough to allow for including important information collected during the lifetime of the considered technical system, and thereby facilitate the updating of the probabilistic model. In this way uncertainty models, which initially are based entirely on subjective information will, as new information is collected, eventually be based on objective information.

In essence the *model building* process consists of five steps, namely:

- Assessment and statistical quantification of the available data.
- Selection of distribution function.
- Estimation of distribution parameters.
- Model verification.
- Model updating.

Typically the initial choice of the model i.e. underlying assumptions regarding distributions and parameters may be based mainly on subjective information whereas the assessment of the parameters of the distribution function and not least the verification of the models is

performed on the basis of the available data. The principle for establishing a probabilistic model is illustrated in Figure E.1.

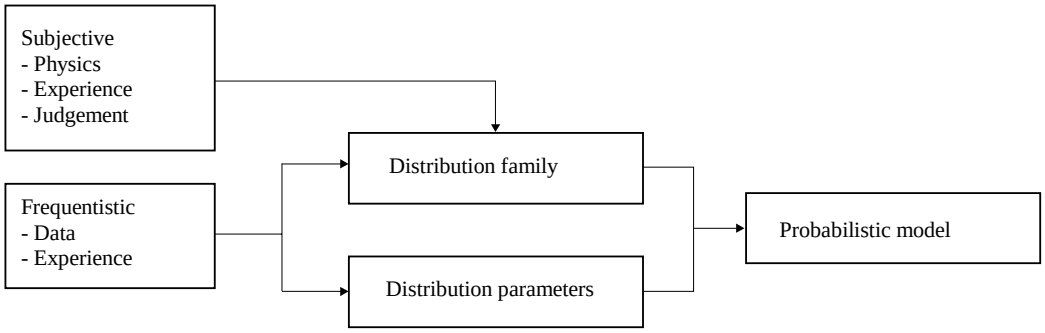


Figure E.1: Illustration of the formulation of probabilistic models for uncertain variables.

As the probabilistic models are based on both frequentistic information and subjective information these are Bayesian in nature.

In the following only the probabilistic modelling of random variables will be considered, but the described approach applies with some extensions also to the probabilistic modelling of random processes and random fields.

First some useful and necessary tools for the statistical analysis are introduced, including some special families of probability distributions together with the concept of *hypothesis testing*. After this the classical statistical theory of assessment of estimators is introduced including assessment of confidence and *significance testing*.

Thereafter the problem of choosing an appropriate distribution function family is addressed, and the task of estimating the parameters of the selected distribution function is considered. Having established models for distributions and parameters, a statistical framework for the verification of such models is given including the classical goodness of fit tests.

E.2 Probability Distributions in Statistics

Throughout the classical statistical theory some distribution functions are repeatedly used for assessment and testing purposes. These include the important *Chi-Square distribution*, the *Chi-distribution*, the *t-distribution* and the *F-distribution*, which hereafter are briefly introduced in accordance with Benjamin and Cornell (1971). The distributions are all related to the Normal distribution and may be derived from this as shown in e.g. Benjamin and Cornell (1971). The numerical evaluation of the distributions may be performed using standard commercial spread sheets such as e.g. Microsoft Excel or tabulations as given in Appendix T.

The Chi-Square (χ^2)- Distribution

When $X_i, i = 1, 2, ..., n$ are standard Normal distributed independent random variables the sum of the squares of the random variables Y_n i.e. :

$$Y_n = \sum_{i=1}^{n} X_i^2 \tag{E.1}$$

is said to be Chi-Square distributed (some times written as χ^2-distributed) with probability density function:

$$f_{Y_n}(y_n) = \frac{y_n^{(n/2-1)}}{2^{n/2}\Gamma(n/2)} \exp(-y_n/2), \qquad y_n \geq 0 \tag{E.2}$$

with mean value $\mu_{Y_n} = n$ (also referred to as the *degrees of freedom*) and variance $\sigma_{Y_n}^2 = 2n$. In Equation (E.2) $\Gamma(\cdot)$ is the complete *Gamma function* defined by:

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt \tag{E.3}$$

As it shall be seen later the Chi-Square distribution is often applied for assessing the statistical characteristics of squared errors but can also be applied in various engineering assessments involving squares of Normal distributed variables such as e.g. the drag component of wave and wind loads and kinetic energy components.

The Chi-Square distribution is regenerative in the sense that the sum of two Chi-Square distributed variables i.e. $Y_{n_1} + Y_{n_2}$ is also Chi-Square distributed with $n_1 + n_2$ degrees of freedom.

From the additive character of the Chi-Square distribution (Equation (E.1)) it is seen from application of the central limit theorem that for sufficiently large *n* the Chi-Square distribution converges towards a Normal distribution with mean value $\mu_{Y_n} = n$ and variance $\sigma_{Y_n}^2 = 2n$.

The Chi (χ)- Distribution

When a random variable Z is given as the square root of a Chi-Square distributed random variable Y_n, the variable Z is said to follow a *Chi-distribution* (sometimes written as χ-distributed) with probability density function:

$$f_Z(z) = \frac{z^{(n-1)}}{2^{n/2-1}\Gamma(n/2)} \exp(-z^2/2), \qquad z \geq 0 \tag{E.4}$$

The mean value μ_z and the variance σ_z^2 are given by:

$$\mu_z = \sqrt{2}\frac{\Gamma\left((n+1)/2\right)}{\Gamma\left(n/2\right)} \tag{E.5}$$

$$\sigma_z^2 = n - 2\frac{\Gamma^2\left((n+1)/2\right)}{\Gamma^2(n/2)} \tag{E.6}$$

The Chi-distribution is e.g. used for the assessment of the distances measured using the principles of Pythagoras or Euclidean norms, and for the assessment of the statistical characteristics of standard deviations.

The *t*-Distribution

A random variable S defined by a Normal distributed random variable X divided by the ratio of the square root of the sum of the squares of n independent standard Normal random variables (i.e. a Chi-distributed random variable) to n, i.e.:

$$S = \frac{X}{\sqrt{\sum_{i=1}^{n} X_i^2}\Big/\sqrt{n}} = \frac{X}{\sqrt{Y_n}/\sqrt{n}} = \frac{\sqrt{n}X}{Z} \tag{E.7}$$

is said to be a *t-distributed random variable* S with n degrees of freedom and with probability density function given as:

$$f_S(s) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\ \Gamma(n/2)}\left(1+\frac{s^2}{n}\right)^{-(n+1)/2} \qquad -\infty \le s \le \infty \tag{E.8}$$

with zero mean and variance σ_S^2 given by:

$$\sigma_S^2 = \frac{n}{n-2} \tag{E.9}$$

As a consequence of the central limit theorem, t-distributed random variables converge to standard Normal distributed random variables for large number of degrees of freedom.

The F-Distribution

A random variable Q defined as the ratio between two Chi-Square distributed random variables Y_{n_1} and Y_{n_2}, i.e.:

$$Q = \frac{Y_{n_1}}{Y_{n_2}} \tag{E.10}$$

is said to be *F-distributed* with (n_1, n_2) degrees of freedom and with probability density function given as:

$$f_Q(q) = \frac{\Gamma((n_1+n_2)/2) q^{(n_1-2)/2} (1+q)^{-(n_1+n_2)/2}}{\Gamma(n_1/2)\Gamma(n_2/2)}, \qquad q \geq 0 \tag{E.11}$$

and mean value and variance given by:

$$\mu_Q = \frac{n_2}{n_2 - 2}, \qquad n_2 > 2 \tag{E.12}$$

$$\sigma_Q^2 = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}, \qquad n_2 > 4 \tag{E.13}$$

The F-distribution as well as the t-distribution are mostly applied in the context of statistical procedures as shall be seen later in Section E.4.

E.3 Estimators for Sample Descriptors – Sample Statistics

When frequentistic information becomes available e.g. in the form of experimental results, a first step is often to try to assess the data simply as they are, without too many assumptions regarding the probabilistic characteristic of the mechanism which generated them. Such an assessment typically concerns the numerical summaries as described in Module C, e.g. the sample moments, but could in principle be any *sample characteristic* of the observed data which is found of interest in a given situation. In statistical terms such characteristics are called *sample statistics*, and in the following the statistical characteristics of such sample statistics will be considered in some detail. To this end, the uncertainty associated with parameter estimators will be assessed and *confidence intervals* on the estimators will be introduced. It should be mentioned that some special sample statistics associated with extreme values are considered directly in the Section on extreme values, i.e. Section D.3. Finally *significance testing* is introduced as a means for evaluating the significance of the variability of statistical data.

Statistical Characteristics of the Sample Average

Consider as an example the case where the permeability of a particular soil is of interest in an engineering decision problem. Due to various effects such as inherent natural variability in the soil composition and the consolidation, the permeability of the considered soil is associated with significant uncertainty. As an engineering model it is assumed that this uncertainty can be taken into account in the formulation of the decision problem by modelling the permeability by a random variable X with distribution function $F_X(x;\mathbf{p})$. Having selected the family of distribution functions, i.e. the distribution function $F_X(x)$, it is still needed to estimate the parameters $\mathbf{p}$ and as will be seen in the following sections this can be done by

e.g. the method of moments or the maximum likelihood method, provided that experiment results $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, ..., \hat{x}_n)^T$ are made available.

In order to better appreciate the uncertainty associated with statistical characteristics such as distribution parameters $\mathbf{p}$, the statistical properties of these are now considered. It is assumed that experiment results of yet unknown values are collected in the vector $\mathbf{X}$. If the experiments are conducted independently, the realizations can be modelled as independent random variables $X_i,\ i = 1, 2, ..., n$ with cumulative distribution functions $F_{X_i}(x_i; \mathbf{p}) = F_X(x; \mathbf{p}), i = 1, 2, ..., n$. Based on the probabilistic model of the realizations $X_i,\ i = 1, 2, ..., n$ it is possible to assess the statistical characteristics of the unknown *sample mean* $\bar{X}$ and the unknown *sample variance* S^2 given by:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{E.14}$$

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2 \tag{E.15}$$

The sample mean $\bar{X}$ and the sample variance S^2 are random variables given in terms of functions of the experiment outcomes $X_i,\ i = 1, 2, ..n$. Such functions are in general called *sample statistics* and include as mentioned previously any characteristic of the considered distribution of interest.

In order to assess the uncertainty by which the sample mean $\bar{X}$ is associated it is interesting to consider its expected value $E\left[\bar{X}\right]$ and variance $Var\left[\bar{X}\right]$. The expected value is given as:

$$E\left[\bar{X}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] = \frac{1}{n} n\mu_X = \mu_X \tag{E.16}$$

which shows that the expected value of the sample mean indeed is equal to the expected value of the underlying random variable, in this case the soil permeability. It was expected that the sample mean is a good estimator for the expected value of a random variable. However, due to the fact that the sample mean is a realization of a random variable, it is clear that the sample mean will normally not turn out to be exactly equal to the expected value of the underlying random variable.

The variability of the sample mean around its expected value can be assessed through the variance of the sample mean $Var\left[\bar{X}\right]$ given by:

$$Var\left[\bar{X}\right] = E\left[(\bar{X} - \mu_X)^2\right] = E\left[(\bar{X} - E\left[\bar{X}\right])^2\right] \tag{E.17}$$

Equation (E.17) may be rewritten as:

$$Var\left[\bar{X}\right] = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2} Var\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var\left[X_i\right]$$
$$= \frac{1}{n^2}\sum_{i=1}^{n} E\left[(X_i - \mu_X)^2\right] = \frac{1}{n}\sigma_X^2 \qquad \text{(E.18)}$$

from which it is seen that the variance of the sample mean $Var\left[\bar{X}\right]$ decreases linearly as a function of the number of samples. Considering the probability that the sample mean $\bar{X}$ will lie within a certain range around the expected value of X i.e. $\mu_X \pm k\sigma_X$ it is seen from Equation (E.18) that the *band width factor* k may be reduced by a factor of 2 by increasing the number of experiments by a factor of 4, to reduce k by a factor of 4 the number of experiments must be increased by a factor of 16. It is seen that it becomes increasingly expensive in terms of experiments to reduce the uncertainty associated with the sample mean. In Figure E.2 the probability density function of a sample mean is illustrated as a function of the number of experiments n used to assess it.

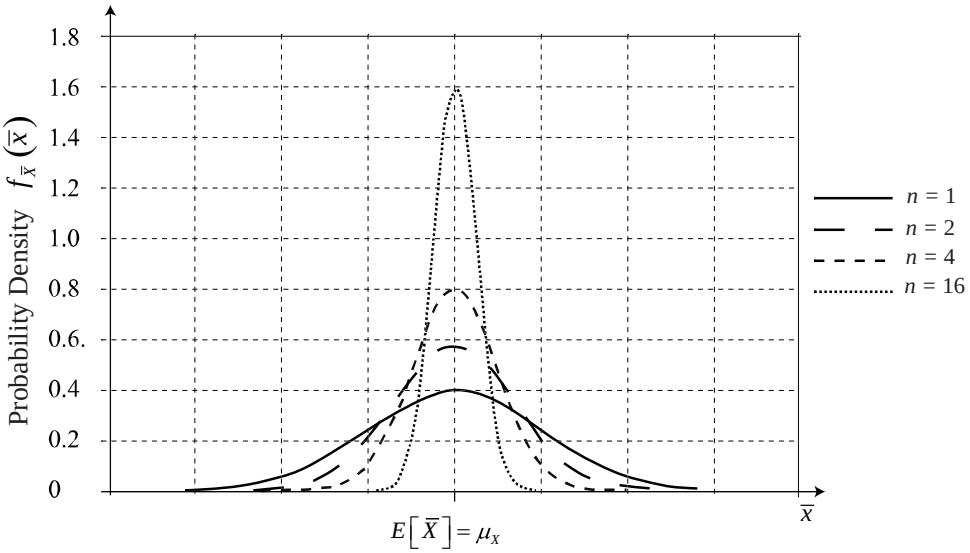


Figure E.2: Illustration of the probability density function of a sample average for different sample sizes *n*.

Statistical Characteristics of the Sample Variance

Whereas the *sample average* is of interest as an estimator of the mean value μ_X of a random variable, the sample variance S^2 is of interest as an estimator of the variance σ_X^2. The expected value of the *sample variance* is determined by taking the expectation of the sample variance as given by Equation (E.15), i.e.:

$$E\left[S^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{1}{n}E\left[\sum_{i=1}^{n}((X_i - \mu) - (\bar{X} - \mu))^2\right]$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n}E\left[(X_i - \mu)^2\right] - nE\left[(\bar{X} - \mu)^2\right]\right)$$

$$= \frac{1}{n}\left(nE\left[(X_i - \mu)^2\right] - nE\left[(\bar{X} - \mu)^2\right]\right) = \tag{E.19}$$

$$= \frac{1}{n}\left(n\sigma_X^2 - n\frac{\sigma_X^2}{n}\right)$$

$$= \sigma_X^2 - \frac{1}{n}\sigma_X^2 = \frac{(n-1)}{n}\sigma_X^2$$

In the step going from the third line to the fourth line in Equation (E.19) the assumption of independence has been used, i.e. using that $E\left[X_i X_j\right] = 0$ for $i \neq j$. From Equation (E.19) it is noticed that the expected value of the sample variance is different from the variance of the underlying random variable. Even though this difference is small for large sample sizes n this is disturbing and essentially means that the estimator S^2 is *biased*, i.e. its mean value is different from σ_X^2. An estimator of the variance σ_X^2 which is *unbiased* $S_{unbiased}^2$ may, however, easily be constructed from S^2 as:

$$S_{unbiased}^2 = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{E.20}$$

It is noted (see later in Section E.6) that the biased estimator S^2 for the variance is applied in both the methods of moments and the maximum likelihood method.

The goodness of an estimator cannot, however, be judged alone on the basis of whether or not it is biased. Another characteristic of estimators commonly used is the efficiency, i.e. the mean square error, associated with an estimator. If the estimator S^2 of the parameter σ_X^2 is considered, the *mean square error* is given by:

$$E\left[(S^2 - \sigma_X^2)^2\right] \tag{E.21}$$

The efficiency of the estimator S^2 can be shown to be better than the efficiency of the estimator $S_{unbiased}^2$ and then the choice stands between a less efficient but *unbiased estimator* or a more efficient but *biased estimator*.

A number of other criteria such as *invariance*, *consistency*, *sufficiency* and *robustness* may be considered when comparing estimators. These characteristics will not be considered here but it is simply noted that the maximum likelihood method estimators in general have equally good or better characteristics than any other estimator. For more details the reader is referred to Benjamin and Cornell (1971) where also further references to specialized literature are provided.

Confidence Intervals on Estimators

As seen in the previous, estimators are associated with *statistical uncertainty* and thus it is essential that this uncertainty is quantified and taken into account in the considered problem context. A classical approach for the quantification and the communication of this uncertainty is by means so-called *confidence intervals*. The $1\text{-}\alpha$ confidence interval on an estimate defines an interval within which the estimated parameter will occur with the predefined probability, with α being the so called *significance level*.

If the case is considered where the standard deviation σ_X of an uncertain variable X is known with certainty and the mean value is unknown, which could be the case e.g. for the yield stress of steel materials, then the so-called double sided and symmetrical $1-\alpha$ confidence interval on the mean value is given by:

$$P\left[-k_{\alpha/2} < \frac{\bar{X} - \mu_X}{\sigma_X \frac{1}{\sqrt{n}}} < k_{\alpha/2}\right] = 1-\alpha \tag{E.22}$$

where n is the number of samples planned for the estimation of the mean value.

Considering the case of a Normal distributed yield stress of a mild construction steel, and assuming that the standard deviation of the yield stress is known equal to 20 MPa and the mean value is unknown, the 0.95 confidence interval of the mean value is given by:

$$P\left[-1.96 < \frac{\bar{X} - \mu_X}{20 \frac{1}{\sqrt{n}}} < 1.96\right] = 1 - 0.05 \tag{E.23}$$

where -1.96 and 1.96 are the simple lower and upper 2.5 percentile values of the standard Normal distribution function i.e. determined by:

$$k_{\alpha/2} = \Phi^{-1}\left(1-\frac{\alpha}{2}\right) \tag{E.24}$$

where $\Phi(\cdot)$ is the standard Normal distribution function.

Assuming e.g. that 16 experiments are planned Equation (E.23) yields:

$$P\left[\mu_X - 9.8 < \bar{X} < \mu_X + 9.8\right] = 0.95 \tag{E.25}$$

which simply states that with 0.95 probability the sample average of the steel yield stress will lie within an interval of ± 9.8 MPa of the true mean value μ_X.

From Equation (E.22) it is seen that the confidence interval limits depend on α, n and σ_X. Typically α is chosen as 0.1, 0.05 and 0.01 in engineering applications. Narrow confidence intervals may be achieved by increasing the number of experiments, which on the other hand may be expensive to achieve and in some cases for practical reasons not even possible.

9th Lecture

Aim of the present lecture

The aim of the present lecture is to introduce the concept of testing for statistical significance. This concept is useful when it is attempted to draw conclusions in regard to the probabilistic characteristics (such as expected value and variance) on the basis of observed realizations of uncertain phenomena. It is explained how hypotheses related to the statistical properties of probabilistic models may be tested on the basis of data. Finally, also the problem of selecting appropriate probability distribution functions for the purpose of modelling uncertainties with basis in observations of uncertain phenomena is treated and it is shown how the concept of probability paper can provide a pragmatic basis for this.

On the basis of the lecture it is expected that the students should acquire knowledge and skills in regard to:

- What is a hypothesis and how may it be tested?
- What is a null-hypothesis?
- What is an alternate hypothesis?
- What is a Type I error and what is a type II error?
- What is the meaning of statistical significance?
- How to perform tests of the mean and variance of a random variable?
- How to derive conclusions on the statistical properties of two data sets?
- What is a probability paper?
- How to construct a probability paper?
- In what way does a probability paper relate to a Quantile plot?
- How to select a probability distribution to represent a random variable based on data?
- In what regions of the probability paper is it especially important that the plotted quantiles fit a straight line?

E.4 Testing for Statistical Significance

In practical engineering problems the engineer is often confronted with the challenge of deriving simple operational conclusions based on an often small set of data exhibiting a high degree of variability. An example of such a situation is the geotechnical engineer attempting by means of a limited number of "on site" vane tests to verify that an empirical soil strength model based on soil specimen laboratory tests is unbiased. Another example is the materials expert pursuing to verify, by analysis of the chloride content of drilled concrete cylinders samples, that the mean value of the surface concentration of chlorides on a concrete structure can be assumed equal to the value assumed in the design basis for the structure. Yet another kind of problem is the selection and/or verification of probabilistic models as shall be seen later in Section E.7.

It is essential that the basis for conclusions in problems such as those outlined in the above are made consistently from case to case (and from engineer to engineer) and moreover that the variability of the observed data and the amount of data is taken appropriately into account. One approach which facilitates the support of such conclusions is the formulation and testing of hypothesis – *hypothesis testing*, which will be introduced in the following.

Consider the example concerning the surface concentration of chlorides on a concrete structure. In the design basis for the structure it was assumed that the surface concentration of chlorides (measured in percentage of total concrete weight) would be 0.3%. Suppose now that the materials expert has studied the chloride contents of concrete cylinders taken from 10 different locations of the considered structure. Even though the materials expert has collected a data set of 10 surface concentrations the observed mean value, also called the sample average $\overline{X}$ (in general terms one of several possible *sample statistics*, i.e. functions of the tested or otherwise observed data) will in some cases be below and in some cases be above the true mean μ of the surface chloride concentration. The question is if on the basis of the observed statistic it can be concluded that the sample average deviates statistically significantly from the assumed mean value. To accommodate for the solution of this problem hypothesis testing includes so-called *operating rules* describing how to achieve a conclusion, which provides a means for assessing the percentage α of times where the achieved conclusions are wrong in one way or the other.

The Hypothesis Testing Procedure

Continuing with the example introduced in the forgoing, a first step is the formulation of the so-called *null-hypothesis* H_0 i.e. expressing that the true mean value μ of the surface chloride concentration is equal to the assumed value of 0.3%. The next step is to formulate an *operating rule* on the basis of which the null-hypothesis can be either accepted or rejected given the test results. An operating rule could be to accept the null-hypothesis H_0 if the sample average $\overline{X}$ of the surface chloride concentration is within the interval $0.3\% \pm \Delta$ and otherwise to reject it. Rejecting the null-hypothesis implies accepting the so-called *alternate hypothesis* H_1 that the true mean value μ is different from the assumed value. Typically the

value Δ is selected such that the probability α of the sample average $\bar{X}$ being outside the interval given by Δ is small, say 0.1.

Two types of errors may occur, namely, rejecting the null-hypothesis H_0 when it is true or accepting it when it is false. These two different types of errors are referred to as *Type I and Type II errors* respectively. It is important to note that performing *Type I* as well as *Type II* errors may be associated with severe consequences. The selection of an appropriate value for α should reflect this. A possible approach for the selection of α is Bayesian decision analysis. The general principles of Bayesian decision analysis are introduced in Module G.

In Summary the Procedure is:

- Formulate a *null-hypothesis* H_0 expressing that the desired condition is fulfilled and formulate the *alternate hypothesis* H_1. Both hypotheses should be formulated in terms of a sample statistic.
- Formulate an operating rule such that the formulated *null-hypothesis* H_0 may easily be either accepted or rejected on the basis of the observation of the sample statistic. Operating rules are typically formulated by means of a constant Δ.
- Select a significance level α for conducting the test (i.e. the probability by which Type I errors occur). This should be done with due considerations of the consequences of performing this type of error.
- By statistical analysis of the sample statistic identify the value of Δ resulting in a probability α of performing a Type I error.
- Perform the planned testing, evaluate the corresponding sample statistic and check which hypothesis is supported by the experiment.
- Provided that the *null-hypothesis* H_0 is not supported by the experiment it is classified as significant at the α-significance level and rejected. Otherwise it is accepted.

In the following a selection of typical cases will be presented considering significance testing.

Testing of the Mean with Known Variance

The example of the surface concentration of chlorides on a concrete structure is considered again. Based on an extensive experience obtained from the assessment of many structures the variance of the chloride surface concentration is assumed to be known equal to $\sigma^2 = 0.04^2$.

The null hypothesis H_0 can be formulated as the true mean μ being equal to 0.3. The alternate hypothesis H_1 is then simply given by $\mu \neq 0.3$. The considered statistic is the sample average $\bar{X}$ which may be assumed to be Normal distributed. The operating rule specifies that the null hypothesis H_0 should be accepted at the α-significance level if:

$$0.3 - \Delta \leq \bar{X} \leq 0.3 + \Delta \qquad \text{(E.26)}$$

where Δ is determined such that the probability of $\bar{X}$ being outside the interval is equal to α i.e.:

$$P\left(0.3-\Delta \le \bar{X} \le 0.3+\Delta\right)=1-\alpha \tag{E.27}$$

Choosing $\alpha=10\%$, Δ can be determined from Equation (E.27) as:

$$\begin{aligned}&\Phi\left(\frac{\Delta}{\frac{0.04}{\sqrt{10}}}\right)-\Phi\left(\frac{-\Delta}{\frac{0.04}{\sqrt{10}}}\right)=0.9 \Rightarrow \Phi\left(\frac{\Delta}{\frac{0.04}{\sqrt{10}}}\right)-\left(1-\Phi\left(\frac{\Delta}{\frac{0.04}{\sqrt{10}}}\right)\right)=0.9 \Rightarrow \\ &\Rightarrow 2\Phi\left(\frac{\Delta}{\frac{0.04}{\sqrt{10}}}\right)-1=0.9 \Rightarrow \frac{\Delta}{\frac{0.04}{\sqrt{10}}}=\Phi^{-1}\left(\frac{0.9+1}{2}\right) \Rightarrow \Delta=\frac{0.04}{\sqrt{10}}\cdot 1.645\end{aligned} \tag{E.28}$$

yielding $\Delta=0.0208$. If the observed sample average over 10 samples lies within the interval $[0.28;0.32]$ the null hypothesis H_0 cannot be rejected.

Assuming that investigations of the material expert resulted in the following data set $x=(0.33,0.32,0.25,0.31,0.28,0.27,0.29,0.3,0.27,0.28)^T$ the sample average is equal to $\bar{x}=0.29$. This is seen to be within the boundaries of the interval given by Δ and the null hypothesis H_0 cannot be rejected.

Testing of the Mean with Unknown Variance

It is not always the case that the variance of a random variable is known and then the appropriate sample statistic is no longer the sample average $\bar{X}$. In this case the following statistic should be used:

$$T=\frac{\bar{X}-\mu}{\frac{S_{unbiased}}{\sqrt{n}}} \tag{E.29}$$

which may be realized to be t-distributed with n-1 degrees of freedom (see e.g. Section E.2).

Similarly to Equation (E.27) the determination of the critical value Δ can be performed from

$$P\left(-\Delta \le T \le \Delta\right)=1-\alpha \tag{E.30}$$

yielding Δ = 1.83. If again the sample of the 10 values of the surface concentration of chlorides is considered, the unbiased sample variance is determined by:

$$s_{unbiased}=\sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2}=0.025 \tag{E.31}$$

whereby the *t* statistic becomes

$$t = \frac{(0.29 - 0.3)\sqrt{10}}{0.025} = -1.27$$

which is seen to be within the interval given by $\Delta = 1.83$. It may thus be concluded that the null hypothesis cannot be rejected.

Testing of the Variance

In some cases also the variance is of direct interest. Consider as an example the situation where it is attempted to reduce the variance σ_{old}^2 of the fatigue life of welded joints by means of a new weld surface treatment. Full scale fatigue experiments on welded joints are typically very expensive and the effect of the surface treatment is attempted to be verified from a relatively small number of experiments n.

For this case the null hypothesis H_0 is that the variance of the fatigue life of the welded joints with the new surface treatment is smaller than σ_{old}^2 i.e. $\sigma_{new}^2 \le \sigma_{old}^2$. The alternate hypothesis H_1 is then $\sigma_{new}^2 > \sigma_{old}^2$. The test statistic is in this case the sample variance S^2 and the operating rule is to accept H_0 if $S^2 \le \Delta$ where Δ can be determined from:

$$P\left[S^2 \le \Delta\right] = 1 - \alpha \tag{E.32}$$

For a given sample size and a given family of distribution of the population, the statistic can easily be evaluated and the critical values Δ determined for $\sigma_{new}^2 = \sigma_{old}^2$, i.e. the largest value of σ_{new}^2 fulfilling the null hypothesis H_0. For instance, if the fatigue life times are Normal distributed or can be transformed into Normal distributed random variables, the statistic $D^2 = nS^2 / \sigma_{old}^2$ is seen to become Chi-Square distributed with *n* degrees of freedom. After fatigue lives have been obtained from experiments the statistic may be evaluated and compared with Δ and thereafter the null hypothesis H_0 can be either accepted or rejected.

Test of Two or More Data Sets

Often it is interesting to be able to compare two or more data sets to see e.g. if it can be assumed that their mean values or their variations can be assumed to be identical or alternatively different.

In the following the two data sets $\hat{\mathbf{x}} = \left(\hat{x}_1, \hat{x}_2, .., \hat{x}_k\right)^T$ and $\hat{\mathbf{y}} = \left(\hat{y}_1, \hat{y}_2, .., \hat{y}_l\right)^T$ are considered. If it can be assumed that the random variables X and Y are Normal distributed with known variances equal to σ_X and σ_Y then a test for equal mean values of X and Y can easily be performed by means of the statistic $\overline{X} - \overline{Y}$ which then is also Normal distributed with mean value $\mu_{\overline{X}\text{-}\overline{Y}}$:

$$\mu_{\overline{X}\text{-}\overline{Y}} = \mu_X - \mu_Y \tag{E.33}$$

and variance $\sigma^2_{\bar{X}-\bar{Y}}$:

$$\sigma^2_{\bar{X}-\bar{Y}} = \frac{\sigma^2_X}{k} + \frac{\sigma^2_Y}{l} \tag{E.34}$$

For this test the null hypothesis H_0 could be given as $\mu_X \leq \mu_Y$ and the alternate hypothesis H_1 consequently as $\mu_X > \mu_Y$. The operating rule in this case could be to accept H_0 provided that $\bar{X} - \bar{Y} \leq \Delta$. In this case the critical value Δ is easily determined from:

$$P\left(\bar{X} - \bar{Y} \leq \Delta\right) = 1 - \alpha \tag{E.35}$$

which for α=10% implies that Δ can be calculated from:

$$\Delta = 1.28\sqrt{\frac{\sigma^2_X}{k} + \frac{\sigma^2_Y}{l}} \tag{E.36}$$

If both X and Y are Normal distributed a test for equal variances can be performed by consideration of the statistic T given by:

$$T = \frac{S^2_{X,unbiased}}{S^2_{Y,unbiased}} \tag{E.37}$$

which is seen to be the ratio between two Chi-Square distributed variables with *k* and *l* degrees of freedom respectively. The null hypothesis H_0 could be given as $\sigma_X = \sigma_Y$ and the alternate hypothesis consequently $\sigma_X > \sigma_Y$. The operating rule in this case could be to accept H_0 provided that $T \leq \Delta$ where Δ is determined from:

$$P\left(T \leq \Delta\right) = 1 - \alpha \tag{E.38}$$

which can be solved by calculation of the F-distribution with *k*, *l* degrees of freedom, using e.g. Microsoft Excel.

Finally also a test for zero correlation i.e. $\rho_{X,Y} = 0$ is given here for the case that X and Y are jointly Normal distributed. In this case it can be shown that the statistic given by the sample correlation coefficient i.e.:

$$R_{X,Y} = \frac{1}{n}\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{S_X S_Y} \tag{E.39}$$

is related to the t-distribution. The null hypothesis H_0 may be given as $\rho_{X,Y} = 0$ and the alternate hypothesis H_1 as $\rho_{X,Y} \neq 0$. The test statistic T which is t-distributed with $n-2$ degrees of freedom is given by:

$$T = \frac{R_{X,Y}\sqrt{n-2}}{\sqrt{1-R_{X,Y}^2}} \tag{E.40}$$

The operating rule is to accept the null hypothesis H_0 provided that $-\Delta \le t \le \Delta$ where Δ is determined from:

$$P(-\Delta \le T \le \Delta) = 1-\alpha \tag{E.41}$$

e.g. using Microsoft Excel.

Finally it should be noted that when more data sets are at hand that these may be compared pair wise.

Some Remarks on Testing

In the foregoing a number of different tests have been introduced for the assessment of observed data. As it is surely very clear by now such tests can be formulated in many different ways and conducted at different levels of significance α.

It should be noted that the different ways for formulating the null hypothesis H_0 and the different choices of the significance level α have impact on the probability of the Type I and Type II errors, respectively. The optimal choice is as already mentioned a decision problem which can be solved by considering a proper weighing of costs and benefits. This is also reflected in the way different organizations formulate their null hypothesis H_0. An organization buying goods from a producing organization tends to postulate that the quality of the goods is below a given criterion, unless it can be shown by testing to be statistically significantly above. This encourages the producing organization to attempt to reduce the variance of the quality of the produced goods.

E.5 Selection of Probability Distributions

In general the distribution function for a given random variable or stochastic process is not known and must thus be chosen on the basis of frequentistic information, physical arguments or a combination of both.

A formal classical approach (described in details in Section E.7) for the identification of an appropriate distribution function on the basis of statistical evidence is to:

- Postulate a hypothesis for the distribution family.
- Estimate the parameters for the selected distribution on the basis of statistical data.
- Perform a statistical test to attempt to reject the hypothesis.

If it is not possible to reject the hypothesis the selected distribution function may be considered to be appropriate for the modelling of the considered random variable. If the hypothesis is rejected a new hypothesis must be postulated and the process repeated.

This procedure follows closely the classical frequentistic approach to statistical analysis. However, in many practical engineering applications this procedure has limited value. This not least due to the fact that the amount of available data most often is too limited to form the solid basis for a statistical test, but also because the available tests applied in situations with little frequentistic information may lead to the false conclusions.

In practice it is, however, often the case that physical arguments can be formulated for the choice of distribution functions and statistical data are therefore merely used for the purpose of checking whether the anticipated distribution function is plausible.

A practically applicable approach for the selection of the distribution function for the modelling of a random variable is thus:

- first to consider the physical reasons why the quantity at hand may belong to one or the other distribution family;
- thereafter to check whether the statistical evidence is in gross contradiction with the assumed distribution; by using e.g. probability paper as explained in the subsequent or if relevant the more formal approaches given in Section E.7.

Model Selection by Use of Probability Paper

Having selected a probability distribution family for the probabilistic modelling of a random variable, *probability paper* is an extremely useful tool for the purpose of checking the plausibility of the selected distribution family.

A probability paper for a given distribution family is constructed such that the cumulative probability distribution function (or the complement) for that distribution family will have the shape of a straight line when plotted on the paper. A probability paper is thus constructed by a non-linear transformation of the *y*-axis.

For a Normal distributed random variable the cumulative distribution function is given as:

$$F_X(x) = \Phi\left(\frac{x-\mu_X}{\sigma_X}\right) \tag{E.42}$$

where μ_X and σ_X are the mean value and the standard deviation of the Normal distributed random variable and where $\Phi(\cdot)$ is the standard Normal probability distribution function. By inversion of Equation (E.42) there is:

$$x = \Phi^{-1}(F_X(x))\sigma_X + \mu_X \tag{E.43}$$

Now by plotting x against $\Phi^{-1}(F_X(x))$, see also Figure E.3, it is seen that a straight line is obtained with slope depending on the standard deviation of the random variable X and

crossing point with the y-axis depending on the mean value of the random variable. Such a plot is sometimes called a quantile plot, see also Section C.3 .

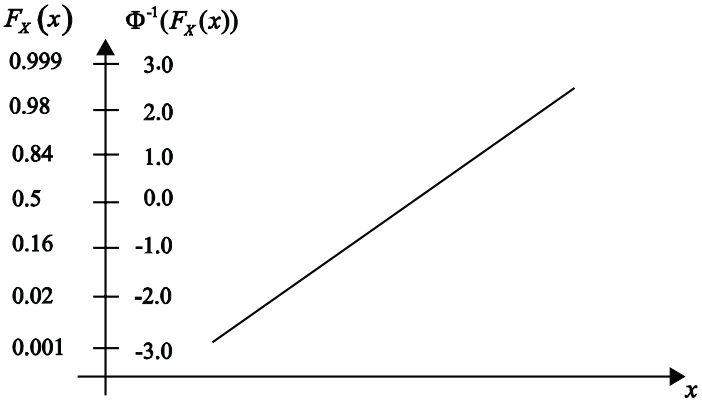


Figure E.3: Illustration of the non-linear scaling of the y-axis for a Normal distributed random variable.

Also in Figure E.3 the scale of the non-linear y-axis is given corresponding to the linear mapping of the observed cumulative probability densities. In probability papers typically only this non-linear scale is given.

Probability papers may also be constructed graphically. In Figure E.4 the graphical construction of a Normal probability paper is illustrated.

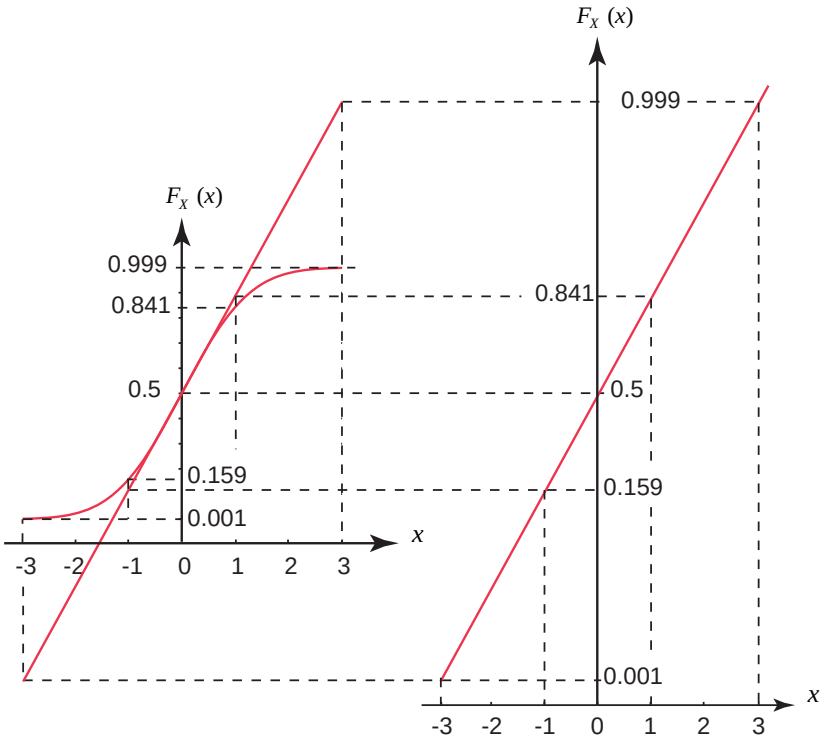


Figure E.4: Illustration of the graphical construction of a Normal distribution probability paper.

Various types of probability paper are readily available in the literature.

Given an ordered set of observed values $\mathbf{x} = (x_1, x_2, .., x_N)^T$ of a random variable the cumulative distribution function may be evaluated as:

$$F_X(x_i) = \frac{i}{N+1} \tag{E.44}$$

In Table E.1 an example is given for a set of observed concrete cube compressive strengths together with the cumulative distribution function values as calculated using Equation (E.44) In Figure E.5 the cumulative distribution values are plotted in a Normal distribution probability paper.

A first estimate of the distribution parameters may readily be determined from the slope and the position of the best straight line through the plotted cumulative distribution values. In Section E.6 the problem of parameter estimation is considered in more detail.

From Figure E.5 it is seen that the observed cumulative distribution function fits relatively well with a straight line. This might also be expected considering that the observed values of the concrete compressive strength are not really representative for the lower tail of the distribution, where due to the non-negativity of the compressive strength it might be assumed that a Lognormal distribution would be more suitable.

i	x_i^O	$F_X(x_i^O)$	$\Phi^{-1}(F_X(x_i^O))$
1	24.4	0.047619048	- 1.668390969
2	27.6	0.095238095	- 1.309172097
3	27.8	0.142857143	- 1.067570659
4	27.9	0.19047619	- 0.876142694
5	28.5	0.238095238	- 0.712442793
6	30.1	0.285714286	- 0.565948707
7	30.3	0.333333333	- 0.430727384
8	31.7	0.380952381	- 0.302980618
9	32.2	0.428571429	- 0.180012387
10	32.8	0.476190476	- 0.059716924
11	33.3	0.523809524	0.059716924
12	33.5	0.571428571	0.180012387
13	34.1	0.619047619	0.302980618
14	34.6	0.666666667	0.430727384
15	35.8	0.714285714	0.565948707
16	35.9	0.761904762	0.712442793
17	36.8	0.80952381	0.876142694
18	37.1	0.857142857	1.067570659
19	39.2	0.904761905	1.309172097
20	39.7	0.952380952	1.668390969

Table E.1: Ordered set of observed concrete cube compressive strengths and the calculated cumulative distribution values.

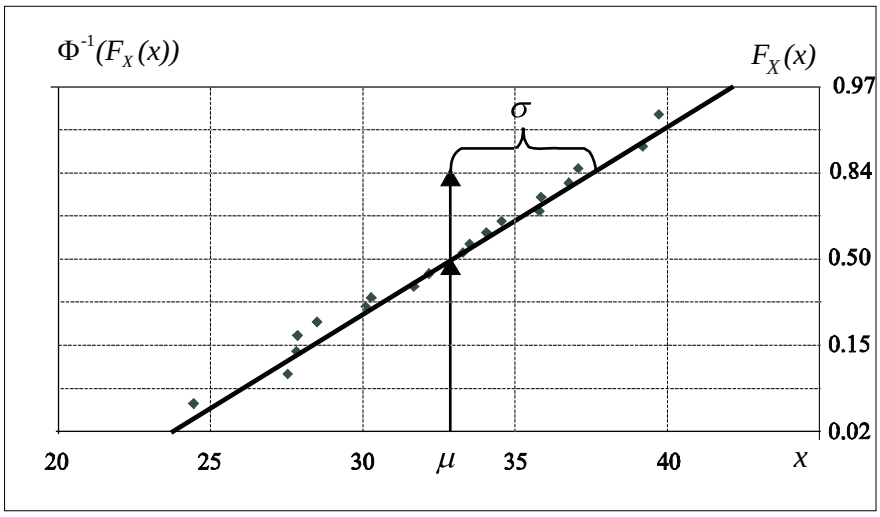


Figure E.5: Concrete cube compressive strength data plotted in Normal distribution paper.

When using probability paper for the consideration of extreme phenomena such as e.g. the maximum water level in a one year period, the probability paper may also be used for the purpose of estimating the values of the water level with a certain return period i.e. for the purpose of extrapolation (see e.g. Schneider, 1994). However, as always when extrapolating, extreme care must be exercised.

10th Lecture

Aim of the present lecture

The aim of the present lecture is, in the context of probabilistic model building, to provide the required theory and methodology for the estimation of parameters of probability distributions based on data. To this end the Method of Moments and the Maximum Likelihood Method are introduced and their limitations and application are discussed and illustrated.

On the basis of the lecture it is expected that the students should acquire knowledge and skills in regard to:

- What is the principle behind the Method of Moments?
- How may the method of moments be applied to estimate the parameters of a probability distribution?
- What kind of estimates does the Method of Moments provide?
- Does the Method of Moments provide for an assessment of the statistical uncertainty associated with the parameters?
- What is the principle behind the maximum likelihood method?
- How may the method of maximum likelihood be applied to estimate the parameters of a probability distribution?
- What is sample likelihood and how is it quantified?
- How can the statistical uncertainty associated with estimated distribution parameters be quantified?
- What is the information matrix and how does this relate to the covariance matrix of the estimated parameters?

E.6 Estimation of Distribution Parameters

There are in principle two different methods to estimate the distribution parameters on the basis of data, namely the methods of *point estimates* and the methods of *interval estimates*. In the following, however, only two of the methods of *point estimates* will be explained, namely the *method of moments* and the *method of maximum likelihood* as these have proven especially useful in practical risk and reliability engineering analysis.

The Method of Moments

Assuming that the considered random variable X may be modelled by the probability density function $f_X(x;\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, .., \theta_k)^T$ are the distribution parameters, the first k moments $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, .., \lambda_k)^T$ of the random variable X may be written as:

$$\begin{aligned} \lambda_j(\boldsymbol{\theta}) &= \int_{-\infty}^{\infty} x^j f_X(x|\boldsymbol{\theta}) dx \\ &= \lambda_j(\theta_1, \theta_2, .., \theta_k) \end{aligned} \tag{E.45}$$

If the random sample, from which the distribution parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, .., \theta_k)^T$ are to be estimated, is collected in the vector $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, .., \hat{x}_n)^T$ the corresponding k sample moments may be calculated as:

$$m_j = \frac{1}{n} \sum_{i=1}^{n} \hat{x}_i^j \tag{E.46}$$

By equating the k sample moments to the k equations for the moments of the random variable X a set of k equations with the k unknown distribution parameters is obtained, the solution of which gives the *point estimates* of the distribution parameters.

The Method of Maximum Likelihood

This method may be somewhat more difficult to use than the method of moments but has a number of very attractive properties, which makes this method especially applicable in engineering risk and reliability analysis.

The principle of the method is that the parameters of the distribution function are fitted such that the probability (likelihood) of the observed random sample is maximised.

Let the random variable of interest X have a probability density function $f_X(x;\boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \theta_2, .., \theta_k)^T$ are the distribution parameters to be estimated.

If the random sample, from which the distribution parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, .., \theta_k)^T$ are to be estimated are collected in the vector $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, .., \hat{x}_n)^T$ the likelihood $L(\boldsymbol{\theta}|\hat{\mathbf{x}})$ of the observed random sample is defined as:

$$L(\boldsymbol{\theta}|\hat{\mathbf{x}}) = \prod_{i=1}^{n} f_X(\hat{x}_i|\boldsymbol{\theta}) \tag{E.47}$$

The maximum likelihood point estimates of the parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, .., \theta_k)^T$ may now be obtained by solving the following optimisation problem:

$$\min_{\theta}(-L(\boldsymbol{\theta}|\hat{\mathbf{x}})) \tag{E.48}$$

Instead of the likelihood function it is advantageous to consider the log-likelihood $l(\boldsymbol{\theta}|\hat{\mathbf{x}})$ i.e.:

$$l(\boldsymbol{\theta}|\hat{\mathbf{x}}) = \sum_{i=1}^{n} \log(f_X(\hat{x}_i|\boldsymbol{\theta})) \tag{E.49}$$

One of the most attractive properties of the maximum likelihood method is that when the number of samples i.e. *n* is sufficiently large the distribution of the parameter estimates converges towards a Normal distribution with mean values $\boldsymbol{\mu}_{\boldsymbol{\Theta}}$ equal to the point estimates, i.e.:

$$\boldsymbol{\mu}_{\boldsymbol{\Theta}} = (\theta_1^*, \theta_2^*, .., \theta_n^*)^T \tag{E.50}$$

The covariance matrix $C_{\Theta\Theta}$ for the point estimates may readily be obtained by:

$$\mathbf{C}_{\boldsymbol{\Theta\Theta}} = \mathbf{H}^{-1} \tag{E.51}$$

where $\mathbf{H}$ is the *Fischer information matrix* with components determined by the second order partial derivatives of the log-likelihood function taken in the maximum, i.e.:

$$H_{ij} = \frac{\partial^2 - l(\boldsymbol{\theta}|\hat{\mathbf{x}})}{\partial\theta_i \partial\theta_j}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \tag{E.52}$$

Example E.1 – Parameter estimation

Consider again the experimental results of the concrete cube compressive strengths given in Table E.1. Assuming that the concrete cube compressive strength is Normal distributed it is required now to estimate the parameters on the basis of the experiment results.

It can be shown that the equations for the moments of a Normal distribution in terms of the distribution parameters are given as:

$$\begin{aligned} \lambda_1 &= \mu \\ \lambda_2 &= \mu^2 + \sigma^2 \end{aligned} \tag{E.53}$$

Analysing the sample data, the first two sample moments are found as:

$$m_1 = 32.67$$
$$m_2 = 1083.36 \tag{E.54}$$

The point estimates of the parameters μ, σ may now be determined by solving the equations:

$$\mu = 32.67$$
$$\mu^2 + \sigma^2 = 1083.36 \tag{E.55}$$

giving:

$$\mu = 32.67$$
$$\sigma = 4.05 \tag{E.56}$$

Using the method of maximum likelihood, the maximum likelihood function is readily written as:

$$L(\boldsymbol{\theta}|\hat{\mathbf{x}}) = \left(\frac{1}{\sqrt{2\pi}\theta_1}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(\hat{x}_i - \theta_2)^2}{\theta_1^2}\right) \tag{E.57}$$

and correspondingly the log-likelihood function as

$$l(\boldsymbol{\theta}|\hat{\mathbf{x}}) = n\ln\left(\frac{1}{\sqrt{2\pi}\theta_1}\right) - \frac{1}{2}\sum_{i=1}^{n}\frac{(\hat{x}_i - \theta_2)^2}{\theta_1^2} \tag{E.58}$$

The mean values of the estimates may be determined by solving the following equations:

$$\frac{\partial l}{\partial \theta_1} = -\frac{n}{\theta_1} + \frac{1}{\theta_1^3}\sum_{i=1}^{n}(\hat{x}_i - \theta_2)^2 = 0$$
$$\frac{\partial l}{\partial \theta_2} = \frac{1}{\theta_1^2}\sum_{i=1}^{n}(\hat{x}_i - \theta_2) = 0 \tag{E.59}$$

yielding:

$$\theta_1 = \sqrt{\frac{\sum_{i=1}^{n}(\hat{x}_i - \theta_2)^2}{n}}$$
$$\theta_2 = \frac{1}{n}\sum_{i=1}^{n}\hat{x}_i \tag{E.60}$$

which by using the sample data gives:

$$\theta_1 = \sigma = 4.04$$

$$\theta_2 = \mu = 32.665$$

Not surprisingly the same result as the method of moments.

As mentioned previously the covariance matrix $\mathbf{C}_{\Theta\Theta}$ for the parameters estimates may be determined through the *information matrix* **H** containing the second-order partial derivatives of the log-likelihood function, see Equation (E.58). The information matrix may be found to be:

$$\mathbf{H} = \begin{pmatrix} \dfrac{n}{\theta_1^2} - \dfrac{3\sum_{i=1}^{n}(x_i - \theta_2)^2}{\theta_1^4} & \dfrac{2\sum_{i=1}^{n}(x_i - \theta_2)}{\theta_1^3} \\ \dfrac{2\sum_{i=1}^{n}(x_i - \theta_2)}{\theta_1^3} & \dfrac{n}{\theta_1^2} \end{pmatrix} \tag{E.61}$$

whereby the covariance matrix is evaluated using the sample data as:

$$\mathbf{C}_{\Theta\Theta} = \mathbf{H}^{-1} = \begin{pmatrix} 0.836 & 0 \\ 0 & 0.1647 \end{pmatrix} \tag{E.62}$$

In a probabilistic modelling where the concrete cube compressive strength enters as a random variable it is then possible to take into account the statistical uncertainty associated with the estimates of the distribution parameters for the distribution function, simply by including the distribution parameters in the reliability analysis as Normal distributed variables with the evaluated mean values and covariance's.

□

11th Lecture

Aim of the present lecture

The aim of the present lecture is to address the problem of model verification and comparison. Having established a probabilistic model in terms of a probability distribution and estimated probability distribution parameters, the issue here is how to evaluate the appropriateness of the established model by means of data. Furthermore, in order to provide guidance on the comparison on the goodness of in principally equally acceptable models a basis for the comparison of two or more models is provided.

On the basis of the lecture it is expected that the students should acquire knowledge and skills in regard to:

- How may probabilistic models be evaluated and validated on the basis of statistical tests?
- What is the idea behind the χ^2 goodness of fit test and how is it performed?
- How is it that the χ^2 goodness of fit test is applied for the testing of the validity of continuous random variable models?
- What is the idea behind the Kolmogorov-Smirnoff goodness of fit test and how is it performed?
- How is the Quantile-plot related to the Kolmogorov-Smirnoff goodness of fit test?
- How conclusive are statistical tests for the purpose of model verification and what must be kept in mind?
- How may probabilistic models be compared in regard to appropriateness?

E.7 Model Evaluation by Statistical Testing

In the foregoing first sections a highly qualitative method – the probability paper - was introduced for the identification of a family or type of probability distributions representing data obtained from observations or experimental results. This method in conjunction with a physical understanding of the mechanism generating the observed data, as already outlined, represents a very pragmatic approach for establishing at least a first model assumption. The next step in model building typically concerns the assessment of the parameters of the assumed distribution function and to this end first the Method of Moments and thereafter the Maximum Likelihood Method were introduced. From the foregoing sections it is obvious that the quality of the established model is a product of the appropriateness of the selected probability distribution and the estimated parameters – i.e. the so-called goodness of fit. It would thus be of significant interest to be able to assess the goodness of fit in a quantitative way allowing for a systematic and consistent way of justifying or rejecting model assumptions. For this purpose the classical statistical distribution tests, i.e. the goodness of fit tests were developed. A number of different types of tests have been developed in the past, in part with very specialized and consequently limited applicability over different application areas. In the following two such tests, namely the χ^2 - and the *Kolmogorov-Smirnov* goodness of fit tests will be explained as these have gained some importance in a broader range of engineering application areas.

In principle the χ^2-goodness of fit test is applicable only for discrete probability distributions, however, it may easily be adapted to continuous probability distributions as shall be seen in the following. The Kolmogorov-Smirnov goodness of fit test on the other hand is only applicable for continuous probability distributions.

The χ^2-Goodness of Fit Test

As mentioned above the χ^2-goodness of fit test is applicable for discrete cumulative distribution functions $P(x_i)$ e.g. defined by:

$$P(x_i) = \sum_{j=1}^{i-1} p(x_j) \tag{E.63}$$

Postulating a cumulative distribution function of the type as given in Equation (E.63) intuitively the differences between predicted frequencies $N_{p,i}$ (using the assumed model) and the observed frequencies $N_{o,i}$ should indicate the quality of the postulated cumulative distribution function and this is indeed the idea behind the χ^2-test.

Assume that the random variable X_j is sampled n times. Then the expected value and the variance of X_j i.e. $E\left[X_j\right]$ and $Var\left[X_j\right]$ are given by (see also Section D.3):

$$E\left[X_j\right] = np(x_j) = N_{p,j}$$
$$Var\left[X_j\right] = np(x_j)(1-p(x_j)) = N_{p,j}(1-p(x_j)) \quad \text{(E.64)}$$

In accordance with the *central limit theorem* and provided that the postulated model is correct, it is reasonable to assume that the standardized random deviations of the sample frequency histogram from the postulated frequency histogram ε_j i.e.:

$$\varepsilon_j = \frac{N_{o,j} - N_{p,j}}{\sqrt{N_{p,j}(1-p(x_j))}} \quad \text{(E.65)}$$

are standard Normal distributed. This however assumes that the number of samples of each of the x_j values is large enough for the *central limit theorem* to be valid. If, however, not just the absolute values of the deviations but rather the squared deviations ε_j^2 , summed up over all possible values of the discrete random variable i.e. for $j = 1, 2, ..., k$, are considered, it is known from section E.2 that this sample statistic is Chi-Square distributed:

$$\varepsilon^2 = \sum_{j=1}^{k} \varepsilon_j^2 = \sum_{j=1}^{k} \frac{(N_{o,j} - N_{p,j})^2}{N_{p,j}(1-p(x_j))} \quad \text{(E.66)}$$

Due to the fact that the numbers of realizations of the discrete random variables are dependent the statistic given by Equation (E.66) does in fact not have k *degrees of freedom* but only $k-1$. Furthermore for the same reason each term in Equation (E.66) shall be reduced with the factor $(1-p(x_j))$ whereby finally the modified statistic ε_m^2 is obtained:

$$\varepsilon_m^2 = \sum_{j=1}^{k} \frac{(N_{o,j} - N_{p,j})^2}{N_{p,j}} \quad \text{(E.67)}$$

which is Chi-Square distributed with k -1 degrees of freedom.

Following the principles given in Section E.4 it is thus possible to formulate and test, at the α -significance level, the null-hypothesis H_0 that the postulated distribution function is not in contradiction with the observed data. The operating rule i.e. the critical value Δ with which the sample statistic shall be compared, can be calculated from:

$$P(\varepsilon_m^2 \geq \Delta) = \alpha \quad \text{(E.68)}$$

It should be underlined that the alternate hypothesis H_1 is less informative in the sense that this hypothesis in principle envelopes all possible distributions and distribution parameters except those of the postulated probability distribution.

Assume as an example that a Normal distribution with mean value $\mu = 33$ and standard deviation $\sigma = 5$ is postulated as representative for the data of the observed concrete compressive strengths presented in Table E.1 – this postulate is the so-called null hypothesis H_0 . It is clear that the concrete compressive strength is a continuous variable but

this can be descritized by dividing the continuous sample space into intervals. The probability of a realization of the continuous random variable in each of the intervals is given as the probability that the outcome of the random variable is smaller that the upper boundary of the interval minus the probability that the outcome of the random variable is smaller than the lower boundary of the interval. It is then possible along the same procedure as explained in the above for discrete random variables to plot the histograms with the observed and predicted frequencies, $N_{o,i}$ and $N_{p,i}$, respectively, for the different data ranges $i = 1, 2, .., k$ in one figure, see Figure E.6.

From Figure E.6 it is seen that the chosen discretization implies that the number of different data ranges k is equal to 4. However, it is noted that the observed and the predicted frequencies in the lower interval is relatively small and it is doubtful if the conditions prevailing the Normal distribution assumption are fulfilled. To overcome this problem it is recommended in the literature, see e.g. Benjamin and Cornell (1971) to lump the data in adjacent intervals such that the number of observations is about 5 or larger. Lumping the frequencies in the two lower intervals yields the histograms shown in Figure E.7.

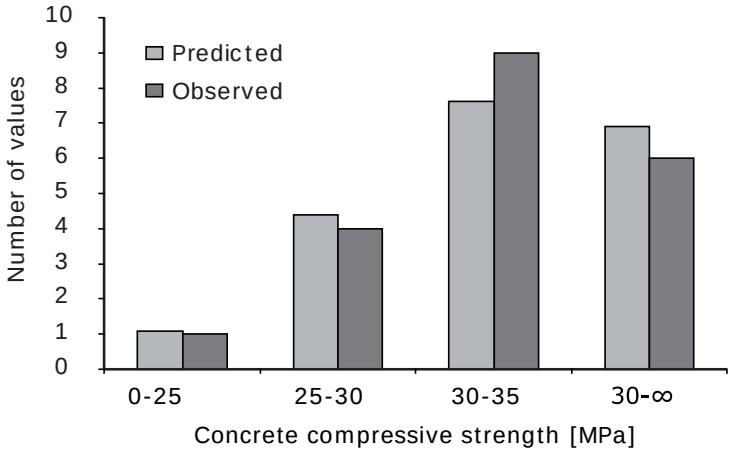


Figure E.6: Predicted and sample histograms for the compressive strength of concrete (data from Table E.1).

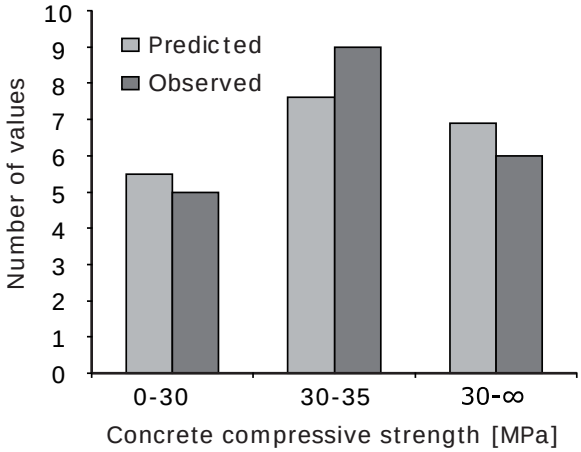


Figure E.7: Predicted and sample histograms for the compressive strength of concrete with lumped data for the lower interval (see Figure E.6).

Following the approach outlined in the foregoing the statistic given in Equation (E.67) is now evaluated as summarized in Table E.2.

Interval - x_j	Number of observed values $N_{o,j}$	Predicted probability $p(x_j)$	Predicted number of observations $N_{p,j} = 20\,p(x_j)$	Sample statistic Equation (E.67)
0 -30	5	0.274253	5.485061	0.042896
30-35	9	0.381169	7.623373	0.248591
35-∞	6	0.344578	6.891566	0.115342
			Sum	0.406829

Table E.2: Calculation sheet for the χ^2 -goodness of fit test of the concrete compressive strength example (raw data in Table E.1).

Finally in order to support a decision on either rejecting or accepting the null-hypothesis H_0 the sample statistic as calculated in Table E.2 must be compared with a critical value Δ given by Equation (E.68). In the present case if the test is performed at a 5%-significance level, the value $\Delta = 5.9915$ can be calculated from a Chi-Square distribution with 2 degrees of freedom using e.g. Microsoft Excel. By comparison of the sample statistic in Table E.2, i.e. 0.406829 with the critical value $\Delta = 5.9915$ it is seen that the null hypothesis cannot be rejected at the 5% significance level.

In the foregoing example not only the type of distribution but also the parameters of the distribution were postulated. In practice it is often the case that the parameters of the distribution are estimated first and thereafter the test for distribution type is performed. This is in principle possible following exactly the same approach as outlined in the above with the modification that the number of degrees of freedom is reduced with the number of parameters estimated from the available data. If e.g. in the example concerning the concrete cube compressive strength it is assumed that first the experiment data are used to assess the standard deviation of the distribution as shown in Section E.6 the number of degrees of freedom is reduced to 1.

Postulating as before a Normal distribution with mean equal to $\mu = 33$ but now with the standard deviation found in Section E.6, i.e. $\sigma = 4.05$ the calculations are modified as shown in Table E.3.

Interval - x_j	Number of observed values $N_{o,j}$	Predicted probability $p(x_j)$	Predicted number of observations $N_{p,j} = 20\,p(x_j)$	Sample statistic Equation (E.67)
0 -30	5	0.229425	4.588507	0.036902
30-35	9	0.459861	9.197211	0.004229
35-∞	6	0.310714	6.214283	0.007389
			Sum	0.044852

Table E.3: Calculation sheet for the χ^2 -goodness of fit test of the concrete compressive strength example with reduced number of degree of freedom (raw data in Table E.1).

With only 1 degree of freedom the critical level Δ is reduced to $\Delta = 3.84$ but the null-hypothesis can still not be rejected at the 5%-significance level.

From the above it is seen that the available data simply do not permit that both of the distribution parameters are first estimated and thereafter the distribution and parameters postulates tested. The number of degrees of freedom is not sufficient. In engineering applications this problem is not unusual as the available data are generally sparse. In other areas such as in producing industries the available amount of data is generally very substantial and the merits of statistical testing are more obvious.

The Kolmogorov-Smirnov Goodness of Fit Test

Whereas the χ^2-goodness of fit test took basis in a statistic quantifying the squared errors between the observed sample histogram and the predicted postulated histogram the goodness of fit test due to Kolmogorov-Smirnov is utilising a sample statistic formulated in terms of the cumulative distributions. However, where the χ^2-goodness of fit test can be applied in cases where both distribution and parameters are postulated, as well as in cases where only the distribution is postulated and the parameters estimated, using the same data utilized for the testing. This is not the case for the Kolmogorov-Smirnov test. Here both the distribution family and parameters must be postulated.

If the observed cumulative distribution function $F_o(x_i)$ is written as:

$$F_o(x_i^O) = \frac{i}{n} \tag{E.69}$$

where x_i is the *i*'th smallest observation in the sample of size *n* and the postulated cumulative distribution function is $F_p(x_i)$ then the following statistic may be utilised for significance testing of the null hypothesis that the observed data do not deviate statistically significantly from the postulated distribution function:

$$\varepsilon_{\max} = \max_{i=1}^{n}\left[\left|F_o(x_i^O) - F_p(x_i^O)\right|\right] = \max_{i=1}^{n}\left[\left|\frac{i}{n} - F_p(x_i^O)\right|\right] \tag{E.70}$$

The distribution of the statistic $\varepsilon_{\max}$ is tabulated for most practical purposes in Table E.4.

	n											
α	1	5	10	15	20	25	30	40	50	60	70	80
0.01	0.9950	0.6686	0.4889	0.4042	0.3524	0.3166	0.2899	0.2521	0.2260	0.2067	0.1917	0.1795
0.05	0.9750	0.5633	0.4093	0.3376	0.2941	0.2640	0.2417	0.2101	0.1884	0.1723	0.1598	0.1496
0.1	0.9500	0.5095	0.3687	0.3040	0.2647	0.2377	0.2176	0.1891	0.1696	0.1551	0.1438	0.1347
0.2	0.9000	0.4470	0.3226	0.2659	0.2315	0.2079	0.1903	0.1654	0.1484	0.1357	0.1258	0.1179

Table E.4: Tabulated values of the Kolmogorov-Smirnov statistic for different significance levels α and sample sizes *n*.

The H_0 null hypothesis may be formulated expressing that the observed data follow the postulated cumulative distribution function and the alternate hypothesis H_1 that the observed data follow some other distribution.

The operating rule, i.e. the critical value Δ the sample statistic $\varepsilon_{\max}$ shall be compared with, can be calculated from:

$$P(\varepsilon_{\max} \geq \Delta) = \alpha \tag{E.71}$$

where Δ is determined from Table E.4.

Consider again the example concerning the concrete compressive strength. As before, it is postulated that the data are representative for a Normal distribution with mean value $\mu = 33$ and a standard deviation equal to $\sigma = 5$. By inspection of Figure E.5 it is seen that the largest deviation between the observed cumulative distribution function and the predicted postulated probability distribution occurs for the 18th data point corresponding to a concrete cube compressive strength of 37.1 MPa. For this value the postulated cumulative distribution function yields:

$$F_p(x_{18}^O) = \Phi\left(\frac{37.1-33}{5}\right) = \Phi(0.8) = 0.794 \tag{E.72}$$

and the observed cumulative distribution function yields:

$$F_o(x_{18}^O) = \frac{i}{n} = \frac{18}{20} = 0.9 \tag{E.73}$$

whereby the sample statistic becomes $\varepsilon_{\max} = 0.9 - 0.794 = 0.106$. From Table E.4 the critical value Δ for $n = 20$ and a 5% significance level is 0.29. Since the sample statistic 0.106 is smaller than the critical value 0.29, the null hypothesis should not be rejected.

Model Comparison

In the foregoing statistical tests were introduced as means for evaluating the goodness of fit of a given postulated distribution function to observed data. These tests can, as mentioned, however, only be applied to assess the plausibility of a given distribution being representative for the observed data. Other postulated distribution functions could also be representative for the observed data and the question thus remains how to select between two postulated distributions which both cannot be rejected by testing as possible candidates. To this end two possibilities might be considered, namely by comparison of the *sample likelihood* defined by Equation (E.47) or by comparison of the likelihood of the sample statistics Equations (E.67) or (E.70). Direct comparison of the sample statistic for the χ^2-goodness of fit test is not a consistent means for comparison, as the number of degrees of freedom may be different for the cases considered.

As an example consider the two cases where the χ^2-goodness of fit test was applied first for testing the goodness of fit for a postulated Normal distribution with postulated parameters $\mu = 33$, $\sigma = 5$ and thereafter a postulated Normal distribution for which the parameters were estimated from the data set, i.e. $\mu = 33$, $\sigma = 4.05$. For the first case the sample statistic is equal to 0.4099 and the number of degrees of freedom is 2. For the second

case the sample statistic is equal to 0.4068 and the number of degrees of freedom is equal to 1. The corresponding likelihoods are equal to 0.8151 and 0.5236 respectively and it is thus seen that the first postulate is more likely that the second postulate given the observed data.

Self Assessment Questions/Exercises

E.1 Which are the steps and constituents in establishing a probabilistic model?

E.2 Express in words the following mathematical expression, where $\bar{X}$ is the sample average of a random variable X and μ_X is the true mean of the random variable:

$$P\left[\mu_X - 9.8 < \bar{X} < \mu_X + 9.8\right] = 0.95$$

E.3 Which are the main steps of hypothesis testing?

E.4 An engineer tests the null hypothesis that the mean value of the concrete cover depth of a concrete structure corresponds to design assumptions. In a preliminary assessment a limited number of measurements of the concrete cover depth are made, and after performing the hypothesis test the engineer accepts the null hypothesis. After a few years, a comprehensive survey of the concrete cover depth is carried out, i.e. many measurements are made. The survey shows that the mean value of the concrete cover depth does not fulfill the design assumptions. Which of the following statement(s) is(are) correct?

In the preliminary survey the engineer has performed a Type I error. ☐

In the preliminary survey the engineer has performed a Type II error. ☐

In the preliminary survey the engineer has performed a Type I and a Type II error. ☐

E.5 Describe in a few words the significance of the probability paper in model selection and how can it be constructed.

E.6 In the following figure data of the annual observed maximum values of precipitation per hour (rainfall) are plotted on a probability paper for the Gumbel distribution. The "best-fit" line is also shown. Could an engineer accept the Gumbel distribution as being suitable for the modelling of the annual maximum precipitation per hour?

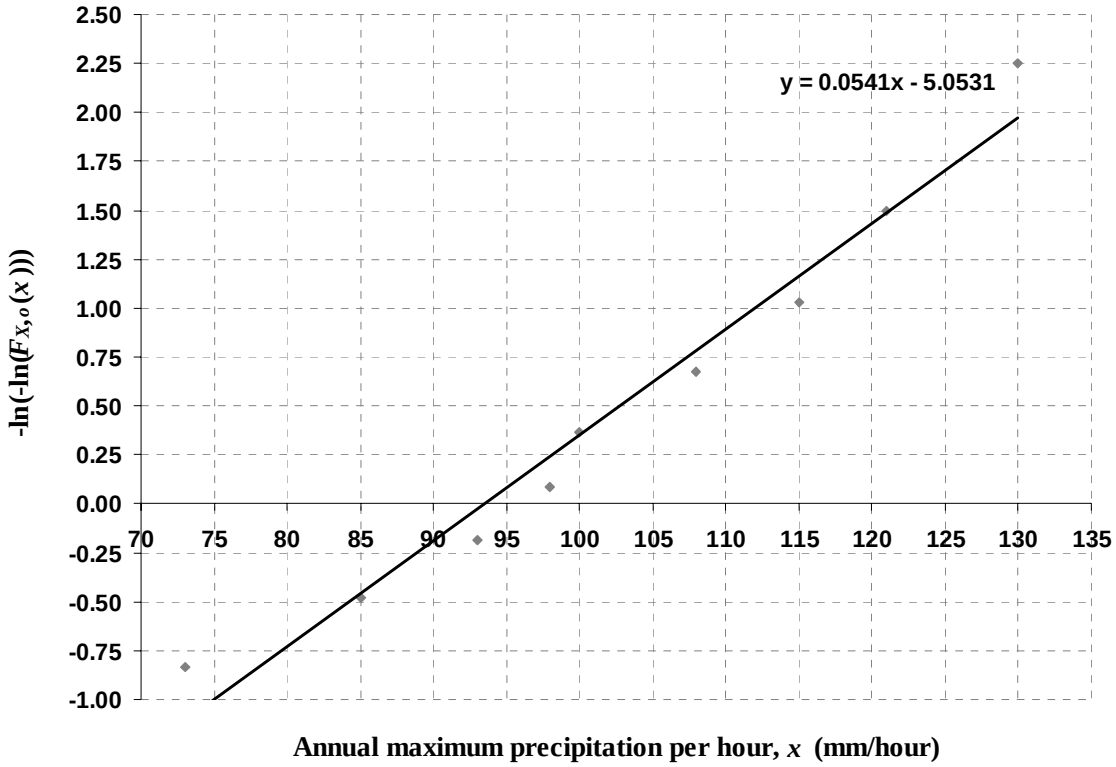


Figure E.8: Probability paper for the Gumbel distribution.

E.7 The Maximum Likelihood Method (MLM) enables engineers to calculate the distribution parameters of a random variable on the basis of data. Which of the following statement(s) is(are) correct?

The MLM provides point estimates of the distribution parameters. ☐

The MLM provides information about the uncertainty associated with the estimated parameters. ☐

The MLM provides no information about the uncertainty associated with the estimated parameters. ☐

E.8 From past experience it is known that the shear strength of soil can be described by a Lognormal distribution. 15 samples of soil are taken from a site and an engineer wants to use the data in order to estimate the parameters of the Lognormal distribution. The engineer:

may use a probability paper to estimate the parameters of the Lognormal distribution. ☐

may use the maximum likelihood method to estimate the parameters of the Lognormal distribution. ☐

may use the method of moments to estimate the parameters of the Lognormal distribution. ☐

None of the above. ☐

E.9 In order to perform a χ^2 test, how do the data need to be divided?

E.10 Which are the main differences between the χ^2 and the Kolmogorov-Smirnov goodness of fit tests?

E.11 Based on experience it is known that the concrete compressive strength may be modelled by a Normal random variable X with mean value $\mu_X = 30MPa$ and standard deviation $\sigma_X = 5MPa$. The compressive strengths of 20 concrete cylinders are measured. An engineer wants to test the null hypothesis H_0 that X follows a Normal distribution with the above given parameters. He/she carries out a χ^2 goodness of fit test by dividing the samples into 3 intervals. He/she calculates a Chi-square sample statistic equal to $\varepsilon_m^2 = 0.41$. Can the engineer accept the null hypothesis at the 5% significance level?

E.12 An engineer wants to examine and compare the suitability of two distribution function model alternatives for a random material property. Measurements are taken of the material property. The engineer uses the two model alternatives to calculate the Chi-square sample statistics and the corresponding sample likelihoods. The results are given in the following table:

Model	**Degrees of freedom**	**Chi-square sample statistic**	**Sample likelihood**
1	2	0.410	0.815
2	1	0.407	0.524

Which of the following statement(s) is(are) correct?

The engineer may accept model 1 at the 5% significance level. ☐

The engineer may accept model 2 at the 5% significance level. ☐

Model 1 is more suitable than model 2. ☐

None of the above. ☐

MODULE F – METHODS OF STRUCTURAL RELIABILITY

12th Lecture

Aim of the present lecture

The aim of the present lecture is to introduce the most basic theory and tools facilitating the representation of events in terms of random variables and to calculate the probability that such events take place. The methods introduced include the classical error accumulation law and set this in perspective to more modern and more general tools to assess probabilities.

On the basis of the lecture it is expected that the students should acquire knowledge on the following issues:

- How may events be represented in terms of basic random variables?
- What is a limit state function and what is a safety margin?
- What is the meaning of a reliability index and how does it relate to a failure probability?
- How to calculate the reliability index for a linear safety margin when all basic random variables are Normal distributed?
- How to calculate the reliability index in the case of nonlinear safety margins?
- What is the idea behind the Monte Carlo Method?
- Which are the steps in the Monte Carlo Method and how are they executed?

F.1 Introduction

The first developments of *First Order Reliability Methods*, also known as FORM methods took place almost 30 years ago. Since then the methods have been refined and extended significantly and by now they form one of the most important methods for reliability evaluations in structural reliability theory. Several commercial computer codes have been developed for FORM analysis and the methods are widely used in practical engineering problems and for code calibration purposes.

In the present chapter first the basic idea behind FORM methods is highlighted and thereafter the individual steps of the methods are explained in detail.

Finally the basic concepts of *Monte Carlo Methods* in structural reliability will be outlined.

F.2 Failure Events and Basic Random Variables

In *reliability analysis* of technical systems and components the main problem is to evaluate the probability of failure corresponding to a specified reference period. However, also other non-failure states of the considered component or system may be of interest, such as excessive damage, unavailability, etc.

In general any state, which may be associated with consequences in terms of costs, loss of lives and impact to the environment are of interest. In the following it will not be differentiated between these different types of states but for simplicity refer to all these as being *failure events*, however, bearing in mind that also *non-failure states* may be considered in the same manner.

It is convenient to describe failure events in terms of functional relations, which if they are fulfilled define that the considered event will occur. A *failure event* may be described by a functional relation, the *limit state function* $g(\mathbf{x})$, in the following way:

$$F = \{ g(\mathbf{x}) \leq 0 \} \tag{F.1}$$

where the components of the vector $\mathbf{x}$ are realisations of the so-called *basic random variables* $\mathbf{X}$ representing all the relevant uncertainties influencing the probability of failure. In Equation (F.1) the failure event F is simply defined as the set of realisations of the function $g(\mathbf{x})$, which is zero or negative.

As already mentioned, other events than failure may also be of interest. In e.g. reliability updating problems events of the following form are highly relevant:

$$I = \{ h(\mathbf{x}) = 0 \} \tag{F.2}$$

Having defined the failure event the *probability of failure* P_F may be determined by the following integral:

$$P_F = \int_{g(\mathbf{x})\le 0} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \tag{F.3}$$

where $f_{\mathbf{X}}(\mathbf{x})$ is the joint probability density function of the random variables $\mathbf{X}$. This integral is, however, non-trivial to solve and numerical approximations are expedient. Various methods for the solution of the integral in Equation (F.3) have been proposed including numerical integration techniques, Monte Carlo simulation and asymptotic *Laplace expansions*. Numerical integration techniques very rapidly become inefficient for increasing dimension of the vector $\mathbf{X}$ and are in general irrelevant. In the following the focus is directed on the widely applied and quite efficient FORM methods, which furthermore can be shown to be consistent with the solutions obtained by asymptotic Laplace integral expansions.

F.3 Linear Limit State Functions and Normal Distributed Variables

For illustrative purposes first the case where the limit state function $g(\mathbf{x})$ is a linear function of the basic random variables $\mathbf{X}$ is considered. Then the limit state function may be written as:

$$g(x) = a_0 + \sum_{i=1}^{n} a_i x_i \tag{F.4}$$

If the basic random variables are Normal distributed the linear *safety margin* M defined through

$$M = a_0 + \sum_{i=1}^{n} a_i X_i \tag{F.5}$$

is also Normal distributed with mean value and variance:

$$\begin{aligned} \mu_M &= a_0 + \sum_{i=1}^{n} a_i \mu_{X_i} \\ \sigma_M^2 &= \sum_{i=1}^{n} a_i^2 \sigma_{X_i}^2 + \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \rho_{ij} a_i a_j \sigma_i \sigma_j \end{aligned} \tag{F.6}$$

where ρ_{ij} are the correlation coefficients between the variables X_i and X_j.

Defining the failure event by Equation (F.1) the probability of failure can be written as:

$$P_F = P(g(\mathbf{X}) \le 0) = P(M \le 0) \tag{F.7}$$

which in this simple case reduces to the evaluation of the standard Normal distribution function:

$$P_F = \Phi(-\beta) \tag{F.8}$$

where β, the so-called *reliability index* due to Cornell (1969) and Basler (1961) is given as:

$$\beta = \frac{\mu_M}{\sigma_M} \tag{F.9}$$

The reliability index β as defined in Equation (F.9) has a geometrical interpretation as illustrated in Figure F.1 where a two dimensional case is considered.

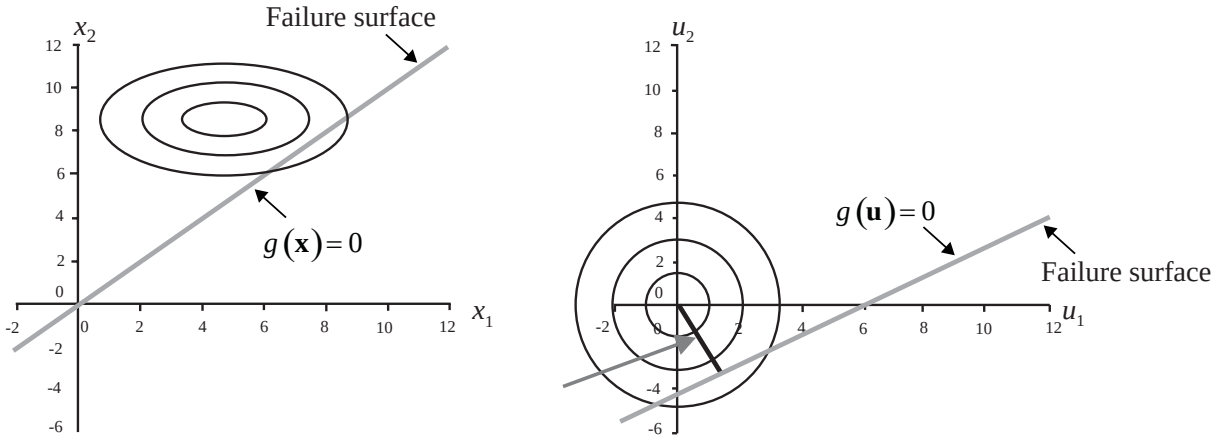


Figure F.1 Illustration of the two-dimensional case of a linear limit state function and Normal distributed variables $\mathbf{X}$.

In Figure F.1 the limit state function $g(\mathbf{x})$ has been transformed into the limit state function $g(\mathbf{u})$ by standardisation of the random variables as:

$$U_i = \frac{X_i - \mu_{X_i}}{\sigma_{X_i}} \tag{F.10}$$

such that the random variables U_i have zero means and unit standard deviations.

The reliability index β has the simple geometrical interpretation as the smallest distance from the line (or generally the hyper-plane) forming the boundary between the *safe domain* and the *failure domain*, i.e. the domain defined by the failure event. It should be noted that this definition of the reliability index due to Hasofer and Lind (1974) does not depend on the limit state function but rather the boundary between the safe domain and the failure domain. The point on the *failure surface* with the smallest distance to the origin is commonly denoted the design point or most likely the failure point.

It is seen that the evaluation of the probability of failure in this simple case reduces to some simple evaluations in terms of mean values and standard deviations of the basic random variables, i.e. the first and second order information.

F.4 The Error Propagation Law

The results given in Equation (F.6) have been applied to study the statistical characteristics of errors ε accumulating in accordance with some differentiable function $h(\mathbf{x})$, i.e.:

$$\varepsilon = h(\mathbf{x}) \tag{F.11}$$

where $\mathbf{x} = (x_1, x_2, .., x_n)^T$ is a vector of realizations of the basic random variables $\mathbf{X}$ representing measurement uncertainties with mean values $\boldsymbol{\mu}_{\mathbf{X}} = (\mu_{X_1}, \mu_{X_2}, .., \mu_{X_n})^T$ and covariances $Cov\left[X_i, X_j\right] = \rho_{ij}\sigma_{X_i}\sigma_{X_j}$ where σ_{X_i} are the standard deviations and ρ_{ij} the correlation coefficients. The idea is to approximate the function $h(\mathbf{x})$ by its Taylor expansion including only the linear terms, i.e.:

$$\varepsilon \cong h(\mathbf{x}_0) + \sum_{i=1}^{n} (x_i - x_{i,0}) \left.\frac{\partial h(\mathbf{x})}{\partial x_i}\right|_{\mathbf{x}=\mathbf{x}_0} \tag{F.12}$$

where $\mathbf{x}_0 = (x_{1,0}, x_{2,0}, .., x_{n,0})^T$ is the point in which the linearization is performed, normally chosen as the mean value point.

$\left.\frac{\partial h(\mathbf{x})}{\partial x_i}\right|_{\mathbf{x}=\mathbf{x}_0}, i = 1, 2, ..., n$ are the first order partial derivatives of $h(\mathbf{x})$ taken in $\mathbf{x} = \mathbf{x}_0$.

From Equation (F.12) and Equation (F.6) it is seen that the expected value of the error $E[\varepsilon]$ can be assessed by:

$$E[\varepsilon] = h(\boldsymbol{\mu}_{\mathbf{X}}) \tag{F.13}$$

and its variance $Var[\varepsilon]$ can be determined by:

$$Var[\varepsilon] = \sum_{i=1}^{n} \left(\left.\frac{\partial h(\mathbf{x})}{\partial x_i}\right|_{\mathbf{x}=\mathbf{x}_0} \right)^2 \sigma_{X_i}^2 + \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \left(\left.\frac{\partial h(\mathbf{x})}{\partial x_i}\right|_{\mathbf{x}=\mathbf{x}_0} \right) \left(\left.\frac{\partial h(\mathbf{x})}{\partial x_j}\right|_{\mathbf{x}=\mathbf{x}_0} \right) \rho_{ij}\sigma_{X_i}\sigma_{X_j} \tag{F.14}$$

It is important to notice that the variance of the error as given by Equation (F.14) depends on the linearization point, i.e. $\mathbf{x}_0 = (x_{1,0}, x_{2,0}, .., x_{n,0})^T$.

Example F.1 – Reliability index – linear safety margin

Consider a steel rod under pure tension loading. The rod will fail if the applied stresses on the rod cross-sectional area ($a = 10\ mm^2$) exceed the steel yield strength. The yield strength R of the rod and the annual maximum stress in the rod S are assumed to be uncertain, modelled by uncorrelated Normal distributed variables. The mean values and the standard deviations of the yield strength and the loading force are given as $\mu_R = 350$ MPa, $\sigma_R = 35$ MPa and $\mu_S = 1500$ N, $\sigma_S = 300$ N respectively.

The limit state function describing the event of failure may be written as:

$$g(\mathbf{x}) = ar - s$$

whereby the safety margin M may be written as:

$$M = aR - S$$

The mean value and standard deviation of the safety margin M are thus:

$$\mu_M = 10 \cdot 350 - 1500 = 2000\,\text{N}$$

$$\sigma_M = \sqrt{10^2 \cdot 35^2 + 300^2} = 461\text{N}$$

whereby the reliability index may be calculated as:

$$\beta = \frac{2000}{461} = 4.33$$

Finally the annual *failure probability* is determined as:

$$P_F = \Phi(-4.33) = 7.5 \cdot 10^{-6}$$

□

Example F.2 – Error propagation law

As an example of the use of the error propagation law consider a right angle triangle *ABC*, where *B* is the right angle. The lengths of the opposite side *b* and adjacent side *a* are measured. Due to measurement uncertainty the length of the sides a and b are modelled as independent Normal distributed random variables with expected values μ_a = 12.2, μ_b = 5.1 and standard deviations σ_a = 0.4 and σ_b = 0.3, respectively. It is assumed that a critical condition will occur if the hypotenuse c is larger than 13.5 and the probability that this condition should happen is to be assessed.

Based on the probabilistic model of a and b the statistical characteristics of the hypotenuse c given by:

$$c = \sqrt{a^2 + b^2}$$

may be assessed through the error propagation model given by Equations (F.13)-(F.14), yielding:

$$E[c] = \sqrt{\mu_a^2 + \mu_b^2} \text{ and } Var[c] = \sum_{i=1}^{n} \left(\frac{\partial h(\mathbf{x})}{\partial x_i}\bigg|_{\mathbf{x}=\mathbf{x}_0} \right)^2 \sigma_{X_i}^2 = \frac{\mu_a^2}{\mu_a^2 + \mu_b^2}\sigma_a^2 + \frac{\mu_b^2}{\mu_a^2 + \mu_b^2}\sigma_b^2$$

which by inserting for a and b their expected values yield:

$$E[c] = \sqrt{12.2^2 + 5.1^2} = 13.22 \text{ and } Var[c] = \frac{12.2^2}{12.2^2 + 5.1^2} 0.4^2 + \frac{5.1^2}{12.2^2 + 5.1^2} 0.3^2 = 0.15$$

As seen from the above the variance of the hypotenuse c depends on the chosen linearization point. If instead of the mean value point a value corresponding to the mean value plus two standard deviations was chosen, the variance of c would have been:

$$Var[c] = \frac{13^2}{13^2 + 5.7^2} 0.4^2 + \frac{5.7^2}{13^2 + 5.7^2} 0.3^2 = 0.149$$

which can be shown to imply a 0.3% reduction of the probability that the hypotenuse is larger than 13.5. Even though such a change seems small it could be of importance in a practical important situation where the consequences of errors can be significant.

□

F.5 Non-linear Limit State Functions

When the limit state function is non-linear in the basic random variables $\mathbf{X}$ the situation is not as simple as outlined in the previous. An obvious approach is, however, considering the error propagation law explained in the foregoing, to represent the failure domain in terms of a linearization of the boundary between the *safe domain* and the *failure domain*, i.e. the failure surface, but the question remains how to do this appropriately.

Hasofer and Lind (1974) suggested performing this *linearization* in the *design point* of the failure surface represented in normalised space. The situation is illustrated in the 2-dimensional space in Figure F.2.

In Figure F.2 a principal sketch is given, illustrating that the failure surface is linearized in the *design point* $\mathbf{u}^*$ by the line $g'(\mathbf{u}) = 0$. The $\boldsymbol{\alpha}$-vector is the out ward directed Normal vector to the failure surface in the design point $\mathbf{u}^*$ i.e. the point on the linearized failure surface with the shortest distance - β - to the origin.

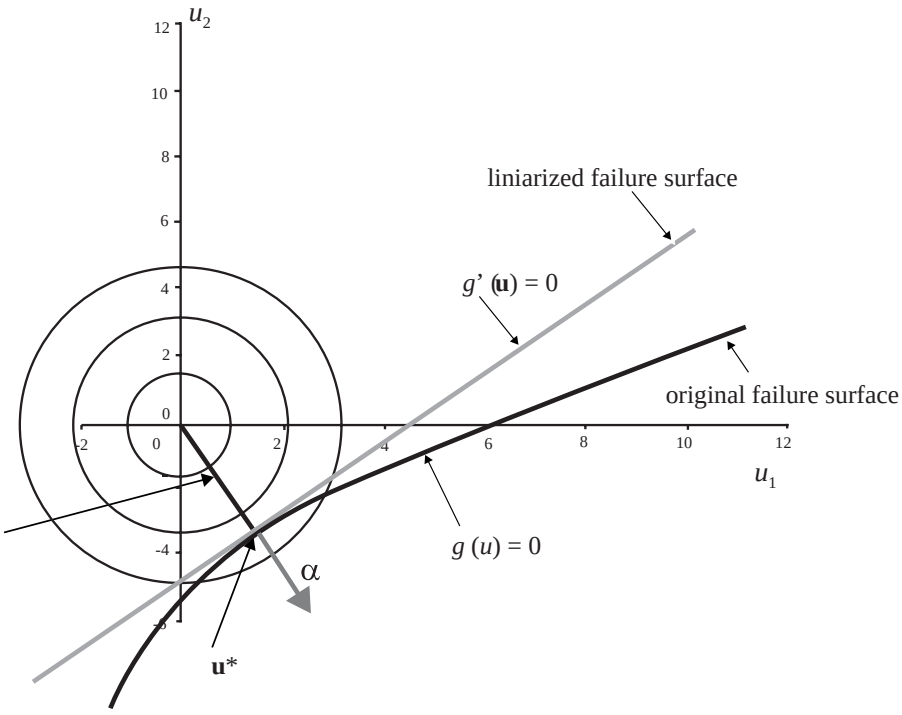


Figure F.2 Illustration of the linearization proposed by Hasofer and Lind (1974) in standard Normal space.

As the limit state function is in general non-linear one does not know the design point in advance and this has to be found iteratively e.g. by solving the following optimisation problem:

$$\beta = \min_{\mathbf{u}\in\{g(\mathbf{u})=0\}} \sqrt{\sum_{i=1}^{n} u_i^2} \tag{F.15}$$

This problem may be solved in a number of different ways. Provided that the limit state function is differentiable the following simple iteration scheme may be followed:

$$\alpha_i = \frac{-\dfrac{\partial g}{\partial u_i}(\beta\boldsymbol{\alpha})}{\left[\displaystyle\sum_{i=1}^{n}\left(\frac{\partial g}{\partial u_i}(\beta\boldsymbol{\alpha})\right)^2\right]^{1/2}}, \quad i = 1,2,...,n \tag{F.16}$$

$$g(\beta\alpha_1, \beta\alpha_2, ..., \beta\alpha_n) = 0 \tag{F.17}$$

First a design point is guessed $\mathbf{u}^* = \beta\boldsymbol{\alpha}$ and inserted into Equation (F.16) whereby a new Normal vector $\boldsymbol{\alpha}$ to the failure surface is achieved. Then this $\boldsymbol{\alpha}$-vector is inserted into Equation (F.17) from which a new β-value is calculated.

The iteration scheme will converge in a few, say normally 6-10 iterations and provides the design point $\mathbf{u}^*$ as well as the reliability index β and the outward normal to the failure surface in the design point $\boldsymbol{\alpha}$. As already mentioned the reliability index β may be related directly to

the probability of failure. The components of the $\boldsymbol{\alpha}$-vector may be interpreted as sensitivity factors giving the relative importance of the individual random variables for the reliability index β.

Second Order Reliability Methods (SORM) follow the same principles as FORM, however, as a logical extension of FORM the failure surface is expanded to the second order in the design point. The result of a SORM analysis may be given as the FORM β multiplied with a correction factor evaluated on the basis of the second order partial derivatives of the failure surface in the design point. The SORM analysis becomes exact for failure surfaces given as a second order polynomial of the basic random variables. However, in general the result of a SORM analysis can be shown to be asymptotically exact for any shape of the failure surface as β approaches infinity. The interested reader is referred to the literature for the details of SORM analyses, e.g. Madsen et al. (1986).

Example F.3 – FORM – non linear limit state function

Consider again the steel rod from example F.1. However, now it is assumed that the cross sectional areas of the steel rod A is also uncertain.

The steel yield strength R is Normal distributed with mean values and standard deviation $\mu_R = 350$ MPa $\sigma_R = 35$ MPa and the loading S is Normal distributed with mean value and standard deviation $\mu_S = 1500$ N, $\sigma_S = 300$ N. Finally the cross sectional area A is assumed Normal distributed with mean value and standard deviation $\mu_A = 10$ mm^2 $\sigma_A = 1$ mm^2.

The limit state function may be written as:

$$g(\mathbf{x}) = r\ a - s$$

Now the first step is to transform the Normal distributed random variables R, A and S into standardized Normal distributed random variables, i.e.:

$$U_R = \frac{R - \mu_R}{\sigma_R}$$

$$U_A = \frac{A - \mu_A}{\sigma_A}$$

$$U_S = \frac{S - \mu_S}{\sigma_S}$$

The limit state function may now be written in the space of the standardized Normal distributed random variables as:

$$g(u) = (u_R\sigma_R + \mu_R)(u_A\sigma_A + \mu_A) - (u_S\sigma_S + \mu_S)$$

$$= (35u_R + 350)(1u_A + 10) - (300u_S + 1500)$$
$$= 350\mathrm{u}_\mathrm{R} + 350u_A - 300u_S + 35u_Ru_A + 2000$$

The reliability index and the design point may be determined in accordance with Equation (F.16) and (F.17) as:

$$\beta = \frac{-2000}{350\alpha_R + 350\alpha_A - 300\alpha_S + 35\beta\alpha_R\alpha_A}$$

$$\alpha_R = -\frac{1}{k}(350 + 35\beta\alpha_A)$$

$$\alpha_A = -\frac{1}{k}(350 + 35\beta\alpha_R)$$

$$\alpha_S = \frac{300}{k}$$

with:

$$k = \sqrt{(350 + 35\beta\alpha_A)^2 + (350 + 35\beta\alpha_R)^2 + (300)^2}$$

which by calculation gives the iteration history shown in Table F.1.

Iteration	Start	1	2	3	4	5
β	3.0000	3.6719	3.7399	3.7444	3.7448	3.7448
α_R	-0.5800	-0.5701	-0.5612	-0.5611	-0.5610	-0.5610
α_A	-0.5800	-0.5701	-0.5612	-0.5611	-0.5610	-0.5610
α_S	0.5800	0.5916	0.6084	0.6086	0.6087	0.6087

Table F.1 Iteration history for the non-linear limit state example.

From Table F.1 it is seen that the basic random variable *S* modelling the load on the steel rod is slightly dominating with an α-value equal to 0.6087. Furthermore it is seen that both the variables *R* and *A* are acting as resistance variables as their α-values are negative. The annual failure probability for the steel rod is determined as $P_F = \Phi(-3.7448) = 9.02 \cdot 10^{-5}$.

□

F.6 Simulation Methods

The probability integral considered in Equation (F.3), i.e.:

$$P_f = \int_{g(\mathbf{x}) \le 0} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \tag{F.18}$$

for the estimation of which is has been seen that FORM methods, may successfully be applied may also be estimated by so-called *simulation techniques*. In the literature a large variety of simulation techniques may be found and a treatment of these will not be given in the present

text. Here it is just noted that simulation techniques have proven their value especially for problems where the representation of the limit state function is associated with difficulties. Such cases are e.g. when the limit state function is not differentiable or when several design points contribute to the failure probability.

However, as all simulation techniques have their origin in the so-called *Monte Carlo Method.* The principles of this very crude simulation technique will be shortly outlined in the following.

The basis for simulation techniques is well illustrated by rewriting the probability integral in Equation (F.18) by means of an *indicator function* as shown in Equation (F.19):

$$P_F = \int_{g(\mathbf{x})\le 0} f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} = \int I\left[g(\mathbf{x})\le 0\right] f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} \tag{F.19}$$

where the integration domain is changed from the part of the sample space of the vector $\mathbf{X} = \left(X_1, X_2, \ldots, X_n\right)^T$ for which $g(\mathbf{x}) \le 0$ to the entire sample space of $\mathbf{X}$ and where $\mathrm{I}\left[g(\mathbf{x}) \le 0\right]$ is an indicator function equal to 1 if $g(\mathbf{x}) \le 0$ and otherwise equal to zero. Equation (F.19) is in this way seen to yield the expected value of the indicator function $I\left[g(\mathbf{x}) \le 0\right]$. Therefore, if now N realisations of the vector $\mathbf{X}$, i.e. $\hat{\mathbf{x}}_j, j = 1, 2\ldots, N$ are sampled it follows from sample statistics that:

$$P_F = \frac{1}{N}\sum_{j=1}^{N} I\left[g(\mathbf{x}) \le 0\right] \tag{F.20}$$

is an unbiased estimator of the failure probability P_F.

The principle of the crude *Monte Carlo simulation* technique rests directly on the application of Equation (F.20). A large number of realisations of the basic random variables **X**, i.e. $\hat{\mathbf{x}}_j, j = 1, 2\ldots, N$ are generated (or simulated) and for each of the outcomes $\hat{\mathbf{x}}_j$ it is checked whether or not the limit state function taken in $\hat{\mathbf{x}}_j$ is positive. All the simulations for which this is not the case are counted (n_F) and after *N* simulations the failure probability P_F may be estimated through:

$$P_F = \frac{n_F}{N} \tag{F.21}$$

which then may be considered a sample expected value of the probability of failure. In fact for $N \to \infty$ the estimate of the failure probability becomes exact. However, simulations are often costly in computation time and the uncertainty of the estimate is thus of interest. It is easily realised that the coefficient of variation of the estimate is proportional to $1/\sqrt{n_F}$ meaning that if Monte Carlo simulation is pursued to estimate a probability in the order of 10^{-6} it must be expected that approximately 10^8 simulations are necessary to achieve an estimate with a coefficient of variance in the order of 10%. A large number of simulations are thus required

using crude Monte Carlo simulation and all refinements of this crude technique have the purpose of reducing the variance of the estimate. Such methods are for this reason often referred to as *variance reduction methods*.

The simulation of the *N* outcomes of the joint density function in Equation (F.21) is in principle simple and may be seen as consisting of two steps. Here the steps will be illustrated assuming that the *n* components of the random vector **X** are independent.

In the first step a "*pseudo random*" number with a uniform distribution between 0 and 1 is generated for each of the components in $\hat{\mathbf{x}}_j$ i.e. $\hat{x}_{ji}$, $i = 1, 2, 3, ..., n$. The generation of such numbers may be facilitated by build-in functions of basically all programming languages and spreadsheet software.

In the second step the outcomes of the "pseudo random" numbers z_{ji} are transformed to outcomes of $\hat{x}_{ji}$ by:

$$x_{ji} = F_{X_i}^{-1}(z_{ji}) \tag{F.22}$$

where $F_{X_i}(\cdot)$ is the cumulative distribution function for the random variable X_i.

The principle is also illustrated in Figure F.3.

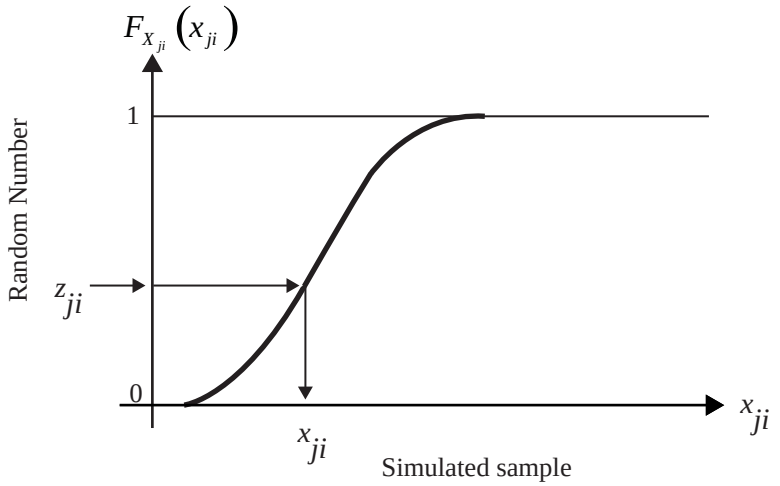


Figure F.3 Principle for simulation of a random variable.

This process is continued until all components of the vector $\hat{\mathbf{x}}_j$ have been generated.

Self Assessment Questions/Exercises

F.1 How may failure events be represented in terms of basic random variables, in the context of the structural reliability theory?

F.2 What is the geometrical interpretation of the reliability index and how does it relate to the failure probability?

F.3 Using the Monte Carlo Simulation method, a sample expected value of the probability of failure is estimated. How may the accuracy in the estimation of the probability of failure be increased?

F.4 Consider a timber beam subjected to an annual maximum bending moment L. The bending strength of the beam R is modelled by a Normal distributed random variable with mean $\mu_R = 30kNm$ and standard deviation $\sigma_R = 5kNm$ and the annual maximum bending moment is modelled by a Normal distributed random variable with mean $\mu_L = 9kNm$ and standard deviation $\sigma_L = 2kNm$. It is assumed that R and L are independent. The timber beam fails when the applied moment exceeds the bending strength. Calculate the reliability index β and the probability of failure of the timber beam.

F.5 Consider a steel rod that carries a deterministic load, S =35 KN. The resistance, R, of the rod is given by the following product: $R = A \cdot f_y$, where A is the area of the rod, equal to 100 mm^2 and f_y is the yield stress modelled as a Normal distributed random variable with mean $\mu_{f_y} = 425 \cdot 10^{-3} KN/mm^2$ and standard deviation $\sigma_{f_y} = 25 \cdot 10^{-3} KN/mm^2$. Formulate a proper safety margin, M, for the steel rod and estimate the rod's reliability. Draw the probability density function of the safety margin and indicate the safe and failure regions.

F.6 The position of a ship is measured by two fixed points A and B located at the coast, see Figure F.4.

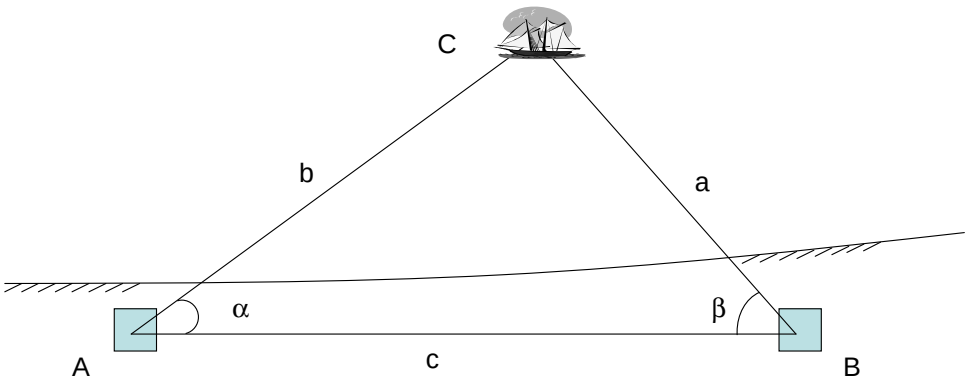


Figure F.4: Position determination of a ship.

Angles α and β have been measured from the basis line AB at the same time. Determine the error in b if the following information is provided:

$$c = 6\ km \pm 0.005\ km$$
$$\alpha = 0.813\ rad \pm 0.011\ rad$$
$$\beta = 1.225\ rad \pm 0.011\ rad$$

where, for instance, $c = 6\ km \pm 0.005\ km$ means that the mean value of c is 6km and the standard deviation of c is 0.005km.

MODULE G – BAYESIAN DECISION ANALYSIS

13th Lecture

Aim of the present lecture

The aim of the present lecture is to illustrate how the basic knowledge acquired through the present course provides a strong basis for engineering decision making. By establishing probabilistic engineering models that are consistent with the available knowledge it is shown how risk or simply expected consequences can be utilized to identify and rank different engineering decision alternatives. To this end on the basis of a simple example, the three principally different types of decision analysis are introduced, namely the prior- posterior- and the pre-posterior decision analysis. Whereas the prior and the posterior decision analyses only differ in the available information at hand at the time of decision making and may serve as direct basis for the planning of engineering activities involving changes of the state of nature, the pure-posterior analysis form a strong basis for the planning of collection of information through e.g. experiments in the laboratory or in the field.

On the basis of the lecture, it is expected that the students should acquire knowledge and skills in regard to:

- What must be identified before a decision analysis can be performed?
- What is a *utility function* and what role does it play in decision making?
- How does risk and utility in principle relate?
- How is a decision event tree constructed?
- How may expected utility be calculated based on branching probabilities and consequences?
- How may the uncertainty associated with information be accounted for in decision analysis?
- What is the difference between prior and posterior decision analysis?
- What is the idea behind the pre-posterior decision analysis?
- How can the value of information be assessed?
- What role does decision making have in engineering risk assessment?

G.1 Introduction

The ultimate task for the engineer is to establish a consistent *decision basis* for the planning, design, manufacturing construction, operation and management of engineering facilities such that the overall life cycle benefit of the facilities are maximized and such that the given requirements to the safety of personnel and environment specified by legislation or society are fulfilled.

As the available information (regarding, e.g., soil properties, loading, material properties, future operational conditions and deterioration processes in general) is incomplete or uncertain, the decision problem is a decision problem subject to uncertain information.

The present chapter introduces some fundamental issues of *decision making* subject to uncertain information. The presentation in turn considers general aspects of decision theory and illustrates these using a simple example. Finally the *risk analysis* decision problem is defined in general terms within the context of *decision theory*.

G.2 The Decision / Event Tree

In practical decision problems such as feasibility studies, reassessment of existing structures or decommissioning of facilities that have become obsolete, the number of alternative actions can be extremely large and a framework for the systematic analysis of the corresponding consequences is therefore expedient.

A *decision/event tree* as illustrated in Figure G.1 may conveniently represent the decision problems.

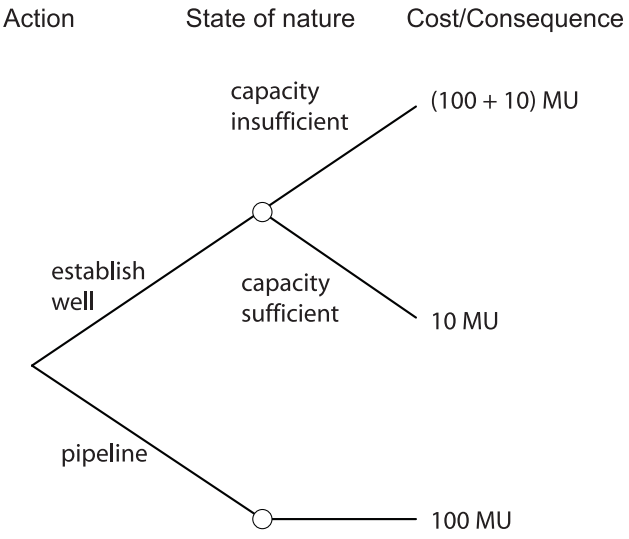


Figure G.1: Decision/event tree.

For the purpose of illustration the *decision/event tree* in Figure G.1 considers the following very simple decision problem. In the specifications for the construction of a production facility using large amounts of fresh water in the production it is specified that a water source capable of producing at least 100 units of water per day must be available. As it is known that

the underground at the location of the planned production facility actually contains a water reservoir, one option is to develop a well locally at the site of the production facility. However, as the capacity of the local water reservoir is not known with certainty another option is to get the water from another location where a suitable well already exists.

The different options are associated with different costs and different potential consequences. The costs of establishing a well locally is assumed to be equal to 10 monetary units (MU). If the already existing well is used it is necessary to construct a pipeline. As the existing well is located far away from the planned production facility the associated costs are assumed to be equal to 100 monetary units.

Based on experience from similar geological conditions it is judged that the probability that a local well will be able to produce the required amount of water is 0.4. Correspondingly the probability that the well will not be able to fulfill the given requirements is 0.6.

The consequence of establishing a well locally which turns out not to be able to produce the required amount of water is that a pipeline to the existing - but distant - well must be constructed. It is assumed that in this case all the water for the production facility is supplied from this well.

The task is now to analyse such decision problems in a way making consistent use of all the information available to the engineer, including her *degree of belief* in the possible states, her subsequent observed data and her *preferences* among the various possible action/state pairs.

To this end use will be made of the fact that decisions shall be based on expected values of the corresponding consequences. This issue is addressed further in the following.

G.3 Decisions Based on Expected Values

Consider the simple case where the engineer must choose between actions a_1 and a_2. The consequence of action a_2 is C with certainty whereas the consequence of action a_1 is uncertain. The state of nature may be θ_1, in which case the consequence is A and the state of nature may be θ_2 in which case the consequence is B. The *decision/event tree* is illustrated in Figure G.2.

Before the true state of nature is known the optimal decision depends upon the likelihood of the various states of the nature θ and the seriousness of the consequences A, B and C.

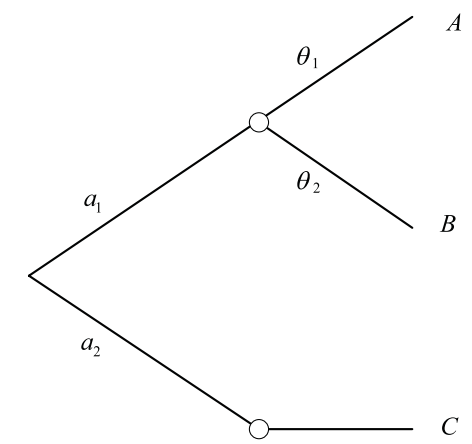


Figure G.2: Decision/event tree illustrating a basic decision problem.

A further analysis of the decision problem requires the numerical assessment of the *preferences* of the decision maker. It is assumed that the decision maker prefers B to A, C to A, and B to C. This statement of *preferences* may be expressed by any function u such that:

$$u(B) > u(C) > u(A) \tag{G.1}$$

The task is to find a particular function u namely the *utility function* such that it is logically consistent to decide between a_1 and a_2 by comparing $u(C)$ with the expected value of the utility of the action a_1, namely:

$$pu(A) + (1-p)u(B) \tag{G.2}$$

where p is the probability that the state of nature is θ_1.

Assuming that $u(A)$ and $u(B)$ have been given appropriate values the question is - what value should $u(C)$ have in order to make the expected value a valid decision criterion? If the probability of θ_1 being the state of nature p is equal to 0 the decision maker would choose a_1 over a_2 because she prefers B to C. On the other hand if the probability of θ_1 being the state of nature is equal to 1 she would choose a_2 over a_1. For a value of p somewhere between 0 and 1 the decision maker will be indifferent to choosing a_1 over a_2. This value p^* may be determined and $u(C)$ is assigned as:

$$u(C) = p^* u(A) + (1-p^*)u(B) \tag{G.3}$$

From Equation (G.3) it is seen that $u(C)$ will lie between $u(A)$ and $u(B)$ for all choices of p^* and therefore the *utility function* is consistent with the stated *preferences*. Furthermore it is seen that the decision maker should choose the action a_1 to a_2 only if the expected utility

given this action $E\left[u\middle|a_1\right]$ is greater than $E\left[u\middle|a_2\right]$. This is realized by noting that for all p greater than p^* and with $u(C)$ given by Equation (G.3). There is:

$$
\begin{aligned}
& u(C) > p\,u(A) + (1-p)\,u(B) \\
& \Updownarrow \\
& p^*\,u(A) + (1-p^*)\,u(B) > p\,u(A) + (1-p)\,u(B) \\
& \Updownarrow \\
& u(B) + (u(A)-u(B))\,p^* > u(B) + (u(A)-u(B))\,p
\end{aligned}
\qquad \text{(G.4)}
$$

This means that if $u(C)$ is properly assigned in consistency with the decision makers stated *preferences* i.e. *B* preferred to *C* preferred to *A* and the indifference probability p^* the ranking of the expected values of the utility determines the ranking of actions.

G.5 Decision Making Subject to Uncertainty

Having formulated the decision problem in terms of a *decision/event tree*, with proper assignment of utility and probability structure, the numerical evaluation of decision alternatives may be performed.

Depending on the state of information at the time of the decision analysis, three different analysis types are distinguished, namely *prior analysis*, *posterior analysis* and *pre-posterior analysis*. Each of these are important in practical applications of decision analysis and are therefore discussed briefly in the following.

G.6 Decision Analysis with Given Information - Prior Analysis

When the *utility function* has been defined and the probabilities of the various state of nature corresponding to different consequences have been estimated the analysis reduces to the calculation of the expected utilities corresponding to the different action alternatives. In the following the utility is represented in a simplified manner through the costs, whereby the optimal decisions now should be identified as the decisions minimizing expected costs, which then is equivalent to maximizing expected utility.

At this stage the probabilistic description $P[\theta]$ of the state of nature θ is usually called a prior description and denoted $P'[\theta]$.

To illustrate the prior decision analysis the decision problem from section G.2 is considered again. The decision problem is stated as follows. The decision maker has a choice between two actions:

a_1: Establish a new well.

a_2 : Establish a pipeline from an existing well.

The possible states of nature are the following two:

θ_1 : Capacity insufficient

θ_2 : Capacity sufficient

The prior probabilities are:

$P'[\theta_1] = 0.60$

$P'[\theta_2] = 0.40$

Based on the prior information alone it is easily seen that the expected cost $E'[C]$ amounts to:

$E'[C] = \min\{P'[\theta_1] \cdot (100+10) + P'[\theta_2] \cdot 10;\ 100\} = \min\{70;100\} = 70\,MU$.

The *decision/event tree* is illustrated in Figure G.3 together with the expected costs (in boxes).

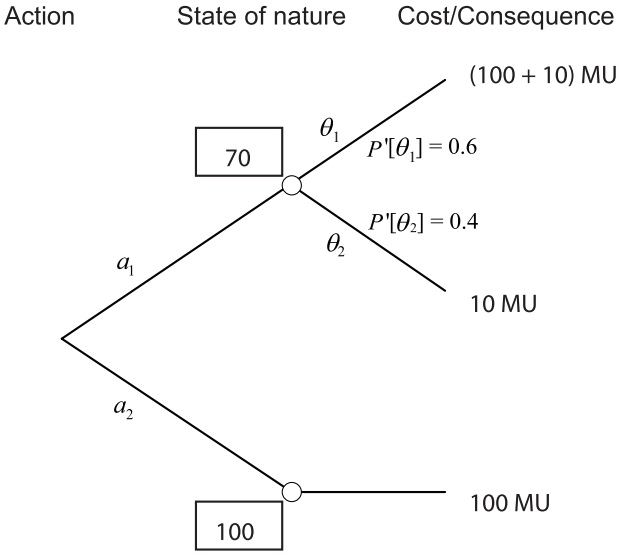


Figure G.3: Simple decision problem with assigned prior probabilities and utility.

It is seen that action alternative a_1 yields the smallest expense (largest expected utility) so this action alternative is the optimal decision.

G.7 Decision Analysis with Additional Information - Posterior Analysis

When additional information becomes available, the probability structure in the decision problem may be updated. Having updated the probability structure the decision analysis is unchanged in comparison to the situation with given - prior information.

Given the result of an experiment z_k the updated probability structure (or just the posterior probability) is denoted $P''[\theta]$ and may be evaluated by use of the Bayes' rule:

$$P''[\theta_i] = \frac{P\left[z_k \middle| \theta_i\right] P'[\theta_i]}{\sum_j P\left[z_k \middle| \theta_j\right] P'\left[\theta_j\right]} \tag{G.5}$$

which may be explained as:

$$\begin{pmatrix} \text{Posterior probability of } \theta_i \\ \text{with given sample outcome} \end{pmatrix} = \begin{pmatrix} \text{Normalising} \\ \text{constant} \end{pmatrix} \cdot \begin{pmatrix} \text{Sample likelihood} \\ \text{given } \theta_i \end{pmatrix} \cdot \begin{pmatrix} \text{prior probability} \\ \text{of } \theta_i \end{pmatrix} \tag{G.6}$$

The normalizing factor is to ensure that $P''[\theta_i]$ forms a proper probability. The mixing of new and old information appears through the sample likelihood $P\left[z_k \middle| \theta_i\right]$ and the prior probability $P'[\theta_i]$. The likelihood is the probability of obtaining the observation z_k given the true state of nature θ_i.

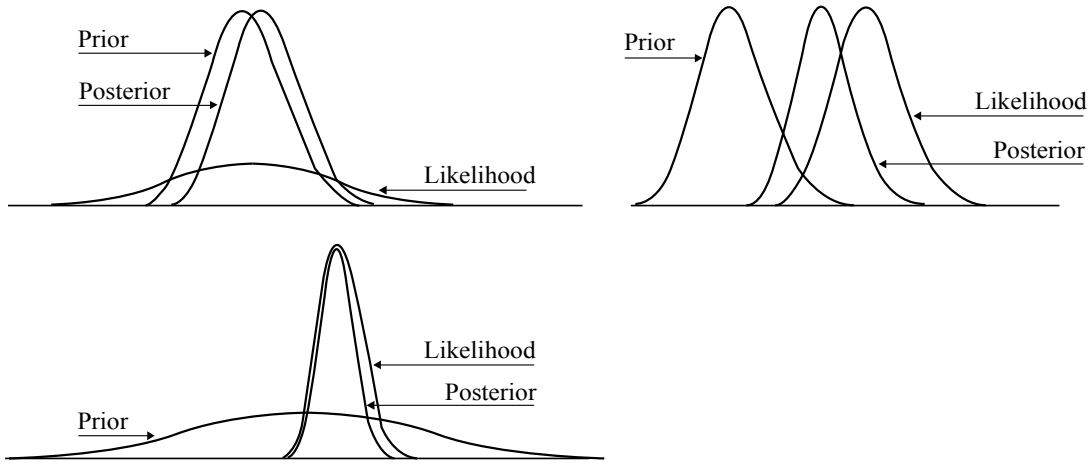


Figure G.4: Illustration of updating of probability structures.

In Figure G.4 an illustration is given of corresponding prior and posterior probability density functions together with likelihood functions. In the first case the prior information is strong and the likelihood is weak (small sample size). In the second case the prior information and the likelihood are of comparable strength. In the last case the prior information is relatively weak in comparison to the likelihood.

To illustrate the posterior decision analysis the water supply decision problem is considered again.

It is assumed that information about the capacity of the local reservoir can be estimated by the implementation of a less expensive test well and subsequent pump test. It is assumed that the cost of establishing a test well is equal to 1 monetary unit. However, the information obtained

from the pump test is only indicative as the result of the difference in scale from the test well to the planned local well.

It is assumed that the pump test can provide the following different information – i.e. indicators regarding the capacity of the local reservoir.

The capacity of the reservoir is:

- larger than the given production requirements by 5% i.e. larger than 105 water volume units per day,
- less than 95% of the required water production, i.e. less than 95 water volume units,
- between 95 and 105 water units.

The information from the pump test is subject to uncertainty and the likelihood of the actual capacity of the local reservoir given the three different indications described above are given in Table G.1.

	True capacity of the reservoir	
Indicators	θ_1 : Less than 100	θ_2 : Larger than 100
I_1 : Capacity >105	0.1	0.8
I_2 : Capacity < 95	0.7	0.1
I_3 : 95≤Capacity≤105	0.2	0.1

Table G.1: Likelihood of the true capacity of the reservoir given the trial pump test results.

Given that a test well is established and a trial pump test conducted with the result that a capacity is indicated smaller than 95 water volume units a posterior decision analysis can be performed to identify whether the optimal decision is to establish a well locally or if it is more optimal to construct a pipeline to the existing well.

Therefore, the posterior probabilities given the new information $P''\left[\theta \middle| z\right]$ can be given as:

$$P''[\theta_1 \mid I_2] = \frac{P[I_2 \mid \theta_1]P'[\theta_1]}{P[I_2 \mid \theta_1]P'[\theta_1] + P[I_2 \mid \theta_2]P'[\theta_2]} = \frac{0.7 \cdot 0.6}{0.7 \cdot 0.6 + 0.1 \cdot 0.4} = \frac{0.42}{0.46} = 0.913$$

$$P''[\theta_2 \mid I_2] = \frac{P[I_2 \mid \theta_2]P'[\theta_2]}{P[I_2 \mid \theta_1]P'[\theta_1] + P[I_2 \mid \theta_2]P'[\theta_2]} = \frac{0.1 \cdot 0.4}{0.7 \cdot 0.6 + 0.1 \cdot 0.4} = \frac{0.04}{0.46} = 0.087$$

which are also shown in Figure G.5 Having determined the updated probabilities the posterior expected values $E''\left[C \middle| I_2\right]$ of the utility corresponding to the optimal action alternative is readily obtained as:

$$E''[C \mid I_2] = \min\left\{P''[\theta_1 \mid I_2] \cdot (100+10) + P''[\theta_2 \mid I_2] \cdot 10;\ 100\right\}$$
$$= \min\{101.3; 100\} = 100\ MU$$

and indicated in boxes in Figure G.5.

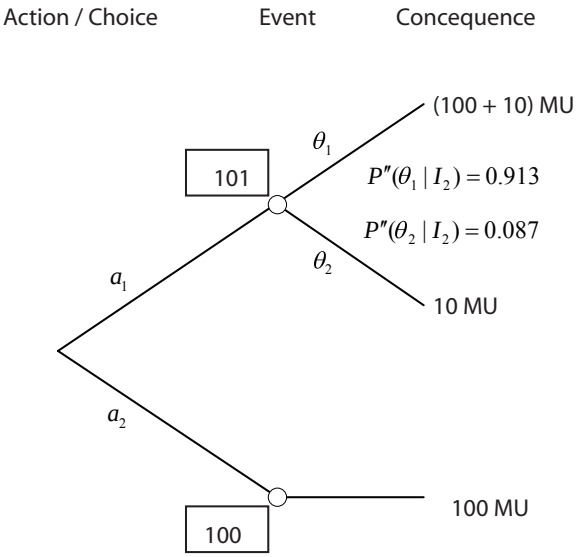


Figure G.5: Illustration of decision/event tree for water supply decision problem.

Considering the additional information the optimal decision has been switched to a_2.

G.8 Decision Analysis with 'Unknown' Information - Pre-posterior Analysis

Often the decision maker has the possibility to 'buy' additional information through an experiment before actually making her choice of action. If the cost of this information is small in comparison to the potential value of the information, the decision maker should perform the experiment. If several different types of experiments are possible the decision maker must choose the experiment yielding the overall largest utility.

If the example from the previous sections is considered again, the decision problem could be formulated as a decision to decide whether or not to perform the trial pump tests.

The situation prior to performing the experiment has already been considered in Section G.6. There it was found that the expected cost based entirely on the prior information $E'[C]$ is 70 monetary units.

In this situation the experiment is planned but the result is still unknown. In this situation the expected cost disregarding the experiment cost can be found as:

$$E[C] = \sum_{i=1}^{n} P'[I_i] E''[C|I_i] = \sum_{i=1}^{n} P'[I_i] \min_{j=1,\ldots,m} \{E''[C(a_j)|I_i]\} \qquad \text{(G.7)}$$

where *n* is the number of different possible experiment findings and *m* is the number of different decision alternatives. In Equation (G.7) the only new term in comparison to the previous section is $P'[I_i]$ which may be calculated by:

$$P'[I_i] = P\left[I_i|\theta_1\right]P'[\theta_1] + P\left[I_i|\theta_2\right]P'[\theta_2] \tag{G.8}$$

With reference to Section G.6 and G.7 the prior probabilities of obtaining the different indications by the tests are $P'[I_1]$, $P'[I_2]$ and $P'[I_3]$ given by:

$$P'[I_1] = P\left[I_1|\theta_1\right]P'[\theta_1] + P\left[I_1|\theta_2\right]P'[\theta_2] = 0.1\cdot 0.6 + 0.8\cdot 0.4 = 0.38$$

$$P'[I_2] = P\left[I_2|\theta_1\right]P'[\theta_1] + P\left[I_2|\theta_2\right]P'[\theta_2] = 0.7\cdot 0.6 + 0.1\cdot 0.4 = 0.46$$

$$P'[I_3] = P\left[I_3|\theta_1\right]P'[\theta_1] + P\left[I_3|\theta_2\right]P'[\theta_2] = 0.2\cdot 0.6 + 0.1\cdot 0.4 = 0.16$$

The posterior expected cost in Equation (G.7) are found to be:

$$\begin{aligned} E''[C|I_1] &= \min\{P''[\theta_1|I_1]\cdot(100+10) + P''[\theta_2|I_1]\cdot 10;\ 100\} \\ &= \min\{0.158\cdot 110 + 0.842\cdot 10;\ 100\} \\ &= \min\{25.8;100\} = 25.8\ MU \end{aligned}$$

$$\begin{aligned} E''[C|I_2] &= \min\{P''[\theta_1|I_2]\cdot(100+10) + P''[\theta_2|I_2]\cdot 10;\ 100\} \\ &= \min\{0.913\cdot 110 + 0.087\cdot 10;100\} \\ &= \min\{101.3;100\} = 100\ MU \end{aligned}$$

$$\begin{aligned} E''[C|I_3] &= \min\{P''[\theta_1|I_3]\cdot(100+10) + P''[\theta_2|I_3]\cdot 10;\ 100\} \\ &= \min\{0.75\cdot(100+10) + 0.25\cdot 10;100\} \\ &= \min\{85;100\} = 85\ MU \end{aligned}$$

where the posterior probabilities $P''\left[\theta_i|I_1\right]$ and $P''\left[\theta_i|I_2\right]$ are determined as already shown in section G.7 for $P''\left[\theta_i|I_3\right]$.

The expected cost corresponding to the situation where the experiment with the experiment costs C_P is therefore:

$$\begin{aligned} E[C] &= E''[C|I_1]P'[I_1] + E''[C|I_2]P'[I_2] + E''[C|I_3]P'[I_3] \\ &= (25.8 + C_P)\cdot 0.38 + (100 + C_P)\cdot 0.46 + (85 + C_P)\cdot 0.16 \\ &= (69.4 + C_P)\,MU \end{aligned}$$

By comparison of this result with the expected cost corresponding to the prior information it is seen that the experiment should be performed if the cost of the experiment a less than 0.6:

$$E'[C] - E[C] = 70 - (69.4 + C_P) = 0.6 - C_P$$

G.9 The Risk Treatment Decision Problem

Having introduced the fundamental concepts of decision theory it will now be considered how these carry over to the principally different types of *risk analysis*.

The simplest form of the *risk analysis*, i.e. a simple evaluation of the risks associated with a given activity and/or decision alternative may be related directly to the prior decision analysis. In the *prior analysis* the risk is evaluated on the basis of statistical information and probabilistic modelling available prior to any decision and/or activity. A simple *decision/event tree* in Figure G.6 illustrates the *prior analysis*. In a *prior analysis* the risk for each possible activity/option may e.g. be evaluated as:

$$R = E[U] = \sum_{i=1}^{n} P_i C_i \tag{G.9}$$

where R is the risk, U the utility, P_i is the i^{th} branching probability and C_i the consequence of the event of branch i.

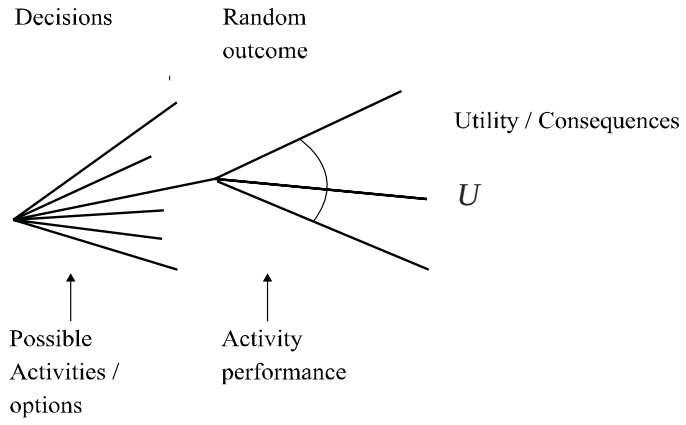


Figure G.6: Decision/event tree for prior and posterior decision analysis.

A *prior analysis* in fact corresponds closely to the assessment of the risk associated with a known activity. A *prior analysis* thus forms the basis for the comparison of risks between different activities.

A *posterior analysis* is in principle of the same form as the *prior analysis*, however, changes in the branching probabilities and/or the consequences in the *decision/event tree* reflect that the considered problem has been changed as an effect of risk reducing measures, risk mitigating measures and/or collection of additional information.

A *posterior analysis* may thus be used to evaluate the effect of activities, which factually have been performed. For example, for assessment of existing facilities the testing and inspection of the "as built" facility would be expected to reveal many gross design and construction errors, leading to a more accurate reliability analysis.

A pre-posterior analysis may be illustrated by the decision/event tree shown in Figure G.7. Using pre-posterior analysis optimal decisions in regard to activities that may be performed in

the future, e.g. the planning of risk reducing activities and/or collection of information may be identified. An important prerequisite for pre-posterior analysis is that decision rules need to be formulated for specifying the future actions that will be taken on the basis of the results of the planned activities.

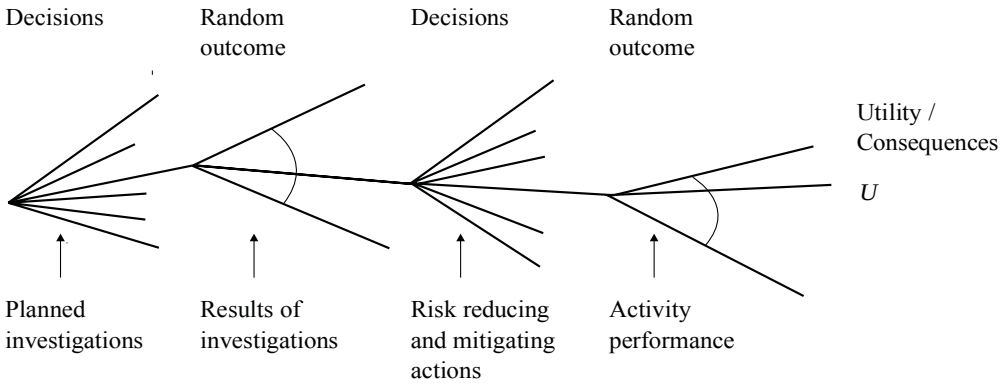


Figure G.7: Decision/event tree for pre-posterior decision analysis.

In a pre-*posterior analysis* the optimal investigation a^* is identified through:

$$\min_a E'_Z\left[E''_Z[C(a(z),z)]\right] = \min_a E'_Z\left[\sum_{i=1}^{n} P_i''(a(z),z)C_i(a(z))\right] \tag{G.10}$$

where $a(z)$ are the different possible actions that can be taken on the basis of the result of the considered investigation z, $E[\cdot]$ is the expected value operator. ´ and ´´ refer to the probabilistic description of the events of interest based on prior and posterior information respectively. In Equation (G.10) the expected utility has been associated only with expected costs why the optimal decision is identified through a minimization. If utility more generally is associated with expected benefits the optimization should be performed through maximization.

Pre-posterior analyses form a strong decision support tool and have been intensively used for the purpose of risk based inspection planning. However, so far pre-posterior decision analysis has been grossly overlooked in risk assessments.

It is important to note that the probabilities for the different events represented in the prior or posterior decision analyses may be assessed by logic tree analysis, classical reliability analysis and structural reliability analysis or any combination of these. The *risk analysis* thus in effect includes all these aspects of systems and component modelling in addition to providing the framework for the decision making.

Self Assessment Questions/ Exercises

G.1 What must be identified before a decision analysis can be performed?

G.2 What is a utility function?

G.3 What is the difference between prior and posterior decision analysis?

G.4 What is the idea behind the pre-posterior decision analysis?

G.5 After heavy snowfall, you need to decide whether to clean up a roof from the snow or not. In the following some information is provided to enable in the decision making.

The clean up of the roof can be made from the local fire department. This option is associated with a cost equal to 4000 CHF. In the case of collapse of the roof due to the snow load the associated cost is equal to 1.000.000 CHF.

The probability of collapse of the roof has been estimated using First Order Reliability Methods (FORM). If the snow is dry, SD , the probability of collapse is: $P_f(SD) = 10^{-3}$. If the snow is wet, SW , the probability of collapse is: $P_f(SW) = 6.2 \cdot 10^{-3}$. In case where there is no snow, SN , on the roof the probability of collapse is equal to: $P_f(SN) = 5 \cdot 10^{-4}$.

Built up an appropriate event tree and use it to find out which decision is the most beneficial one in terms of cost?

REFERENCES

Alsalam, N., et al. Economic Effects of Federal Spending on Infrastructure and Other Investments. (1998) [cited; Available from: http://www.cbo.gov

Basler, E. (1961) Untersuchungen über den Sicherheitsbegriff von Bauwerken. Schweizer Archiv für angewandte Wissenschaft und Technik.

Benjamin, J.R. & Cornell, C.A. (1971). Probability, Statistics and Decision for Civil Engineers. McGraw-Hill, New York.

Brundtland, G.H. and International Metalworkers' Federation. (1987), We have only one earth : metalworkers, economic growth, and the environment. 1987, International Metalworkers' Federation: Geneva, Switzerland. p. 27 p.

Cornell, C.A. (1969). A Probability Based Structural Code. ACI-Journal, Vol.66, pp. 974-985.

Hasofer, A.M. and Lind, N.C. (1974). An Exact and Invariant First Order Reliability Format. *Journal of Engineering Mechanics*. Div. Proc. ASCE.100(EMl), pp. 111-121.

JCSS (2001). Probabilistic Model Code. The Joint Committee on Structural Safety.

Madsen, H.O, Krenk, S. and Lind, N.C. (1986). Methods of Structural Safety. Prentice-Hall, Inc., New Jersey.

Melchers, R.E. (1987). Structural Reliability Analysis and Prediction. Ellis Horwood Limited. ISBN 0 85312 930 4.

Schneider, J. (1994). Sicherheit und Zuverlässigkeit im Bauwesen, Grundwissen für Ingenieure. VDF, Hochschulverlag AG an der ETH Zürich. (In German).

INDEX

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

ANNEX A

ANSWERS/SOLUTIONS TO SELF ASSESSMENT QUESTIONS/EXERCISES

Module A

A.1 According to the so-called Brundtland Commission (1987), a sustainable development is defined as a development "that meets the needs of the present without compromising the ability of future generations to meet their own needs". Consideration of a sustainable development leads to sustainable decision making which may be understood as based on a joint consideration of society, economy and environment. *(For more see section A.1)*

A.2 A beneficial engineering facility is understood as: being economically efficient in serving a specific purpose, fulfilling given requirements in regard to the safety of the personnel, and fulfilling given requirements to limit the adverse effects of the facility on the environment. *(For more see section A.2)*

A.3 As discussed in *section A.3* when considering an activity with only one event with potential consequences C, the risk R is the probability P that this event will occur multiplied with the consequences given the event occurs i.e.:

$$R = P\ C$$

A.4 The term “acceptable risks” points out to “what is one prepared to invest and/or pay for the purpose of getting a potential benefit”. *(For more see section A.2)*

A.5 As discussed in *section A.3* the risk of an event is calculated by Equation (A.1) such as: $R = P\ C$. Hence the given table can easily be completed and it can be seen that event 3 is associated with the higher risk.

Event	1	2	3
Event probability	10%	1%	20%
Consequences	100 SFr	500 SFr	100 SFr
Risk	10	5	20

Module B

B.1 The estimation is based on the so called frequentistic interpretation of probability. In the frequentistic interpretation the probability $P(A)$ is simply the relative frequency of occurrence of the event A as observed in an experiment with n trials. It is mathematically defined as:

$$P(A)=\lim \frac{N_A}{n_{\text{exp}}} \qquad \text{for} \qquad n_{\text{exp}} \to \infty$$

(For more see section B.2)

B.2 Following the rule of Bayes' (*see section B.5*) the conditional probability of the event E_1 given that the event E_2 has occurred is written as: $P(E_1 \mid E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}$.

B.3 In probability theory the probability, $P(A)$, of an event A can take any value within the following boundaries:

$0 \le P(A) \le 1$ ☑

$-1 \le P(A) \le 1$ ☐

$-\infty \le P(A) \le \infty$ ☐

B.4 If the intersection of two events, A and B corresponds to the empty set $\varnothing$, i.e. $A \cap B = \varnothing$, the two events are:

Mutually exclusive. ☑

Independent. ☐

Empty events. ☐

B.5 Which one(s) of the following expressions is(are) correct?

The probability of the union of two events A and B is equal to the sum of the probability of event A and the probability of event B, given that the two events are mutually exclusive. ☑

The probability of the union of two events A and B is equal to the probability of the sum of event A and event B, given that the two events are mutually exclusive. ☐

The probability of the intersection of two events A and B is equal to the product of the probability of event A and the probability of event B, given that the two events are mutually exclusive. ☐

The probability of the intersection of two events A and B is equal to the product of the probability of event A and the probability of event B, given that the two events are independent. ☑

B.6 The probability of the intersection of two mutually exclusive events is equal to:

The product of the probabilities of the individual events. ☐

The sum of the probabilities of the individual events. ☐

The difference between the probabilities of the individual events. ☐

One (1). ☐

Zero (0). ☑

B.7 Within the theory of sample spaces and events, which one(s) of the following statements is(are) correct?

An event A is defined as a subset of a sample space Ω. ☑

A sample space Ω is defined as a subset of an event A. ☐

B.8 The probability of the union of two not mutually exclusive events A and B is given as: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. It is provided that the probability of event A is equal to 0.1, the probability of event B is 0.1 and the probability of event B given event A, i.e. $P(B|A)$ is 0.8. Which result is correct?

$P(A \cup B) = -0.6$ ☐

$P(A \cup B) = 0.12$ ☑

$P(A \cup B) = 0.04$ ☐

B.9 For an event A in the sample space Ω, event $\overline{A}$ represents the complementary event of event A. Which one(s) of the following hold?

$A \cup \overline{A} = \Omega$ ☑

$A \cap \overline{A} = \Omega$ ☐

$A \cup \overline{A} = \varnothing$ ☐

B.10 The commutative, associative and distributive laws describe how to:

Operate with probabilities. ☑

Operate with intersections of sets. ☑

Operate with unions of sets. ☑

None of the above. ☐

B.11 Following the principles explained in *section B.5* it is:

a. The table is completed as follows:

SNF final decision D_i	**Dr. Beispiel's indicative assessment,** I_j		
	$I_j = D_1$	$I_j = D_2$	$I_j = D_3$
D_1	**0.86**	**0.1**	**0.04**
D_2	**0.2**	**0.74**	**0.06**
D_3	**0**	**0.1**	**0.9**

b. Using the Bayes' Theorem the probability that the final decision made by SNF is the same with the indicative assessment of Dr. Beispiel is:

$$P\left(D_2 \middle| I = D_2\right) = \frac{P(I = D_2 \middle| D_2)P(D_2)}{\sum_{i=1}^{3} P(I = D_2 \middle| D_i)P(D_i)} =$$

$$\frac{P(I = D_2 \middle| D_2)P(D_2)}{P(I = D_2 \middle| D_1)P(D_1) + P(I = D_2 \middle| D_2)P(D_2) + P(I = D_2 \middle| D_3)P(D_3)} =$$

$$\frac{0.74 \cdot 0.35}{(0.1 \cdot 0.45) + (0.74 \cdot 0.35) + (0.1 \cdot 0.2)} = 0.799$$

Module C

C.1 The main purpose of the use of descriptive statistics is to assess the characteristics and the level of uncertainty of a given quantity of interest without assuming anything in terms of the degree or nature of the randomness underlying the data analysed, (*see also section C.1*)

C.2 By definition the sample coefficient of correlation may lie in the interval $[-1;1]$.

In both extreme cases, there are linear relationships between two data sets, (*see also section C.2*).

C.3 The interval width plays a role for the resolution of the representation of the observations. If the interval width is too large, the histogram tells little about relative occurrences of individual phenomena. If the width is too small, the relative occurrences in each interval fluctuate due to the random nature of the phenomena, (*see also section C.3*).

C.4 As discussed in *section C.3* five characteristics of a data set are normally presented in a Tukey box plot: the lower adjacent value, the lower quartile, the median, the upper quartile and the upper adjacent value. Outside values can also be shown on a Tukey box plot.

C.5 Q-Q plots provide an efficient means of comparison of observations of two different data sets, (*see also section C.3*).

C.6 Provide an estimate of the correlation coefficient of the data sets plotted in the following figure.

A $r_{XY} \approx$ +0.8

B $r_{XY} \approx$ +0.4

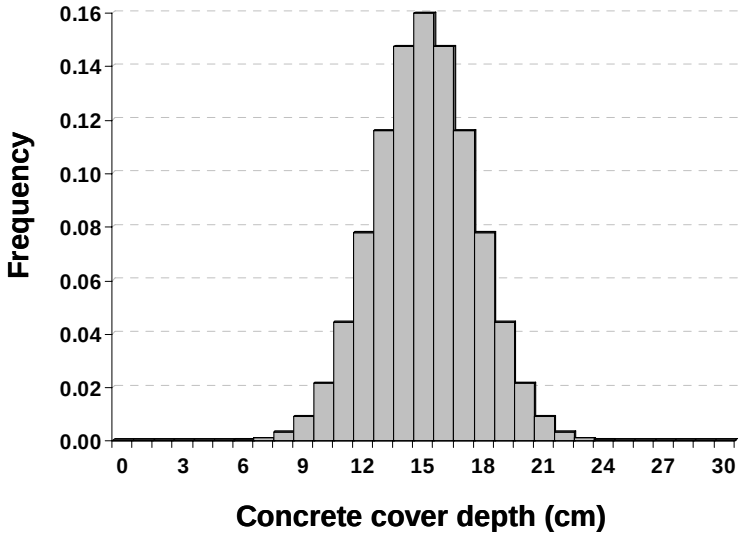
C $r_{XY} \approx$

D $r_{XY} \approx$ -0.4

C.7 A number of statistical terms are shown in the following table. Check if the terms have something to do with (a) location parameter, (b) dispersion parameter or (c) none of the above.

	a	b	c
Mean	☑	☐	☐
Quartile	☑	☐	☐
Sample size	☐	☐	☑
Median	☑	☐	☐
Standard deviation	☐	☑	☐
Coeffiçient of variation	☐	☑	☐

C.8 Measurements were taken of the concrete cover depth of a bridge column. The histogram of the measured values has been plotted.



If X represents the random variable for the concrete cover depth which one(s) of the following statement(s) is(are) correct?

The sample mean, $\bar{x}$, is equal to 0.16 cm. ☐

The sample mean, $\bar{x}$, is equal to 15 cm. ☑

The mode of the data set is equal to 15 cm. ☑

C.9 Which one(s) of the following are features of a symmetrical probability density function?

The variance is equal to the coefficient of variation. ☐

The mode is equal to the median. ☑

The skewness is equal to zero. ☑

None of the above. ☐

Module D

D.1 Inherent natural variability may be interpreted simply as the uncertainty which cannot be reduced by means of collection of additional information. This definition implies that the amount of uncertainty due to inherent natural variability depends on the models applied in the formulation of the engineering problem. Presuming that a refinement of models corresponds to looking more detailed at the problem at hand, one could say that the uncertainty structure influencing a problem is scale dependent. The type of uncertainty associated with the state of knowledge has a time dependency. In principle, if the observation is perfect without any errors the knowledge about the phenomenon is perfect. The modelling of the same phenomenon in the future, however, is uncertain as this involves models subject to natural variability, model uncertainty and statistical uncertainty. The above discussion shows another interesting effect that the uncertainty associated with a model concerning the future transforms from a mixture of aleatory and epistemic uncertainty to a purely epistemic uncertainty when the modelled phenomenon is observed, (*see also section D.2*).

D.2 Epistemic uncertainty involves statistical uncertainty and model uncertainty. Epistemic uncertainty may be reduced by e.g. collecting additional information. On the other hand, aleatory uncertainty is related to the random nature of phenomena, and thus cannot be reduced by collecting information, (*see also section D.2*).

D.3 A continuous random variable is a random variable which can take on any value, (*see also section D.3*)

D.4 With the help of Equations (D.16) and (D.18) it is:

a. $E\left[a+bX\right]=a+bE\left[X\right]$

b. $Var\left[a+bX\right]=b^2\cdot Var\left[X\right]$

D.5 The required characteristics of the random variable are shown in the following illustrations.

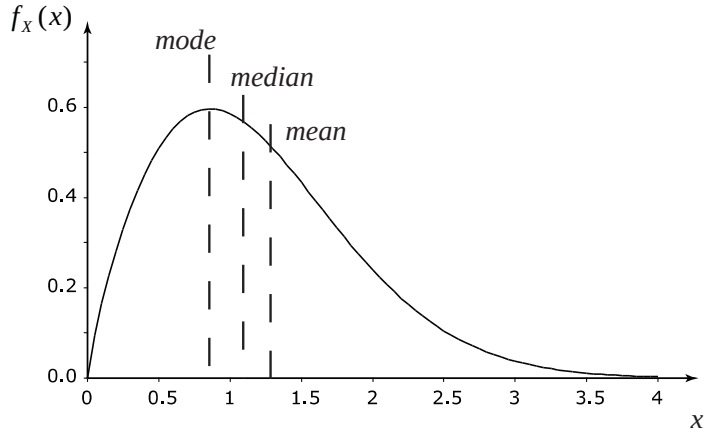


Figure D.12: Illustration of a probability density function.

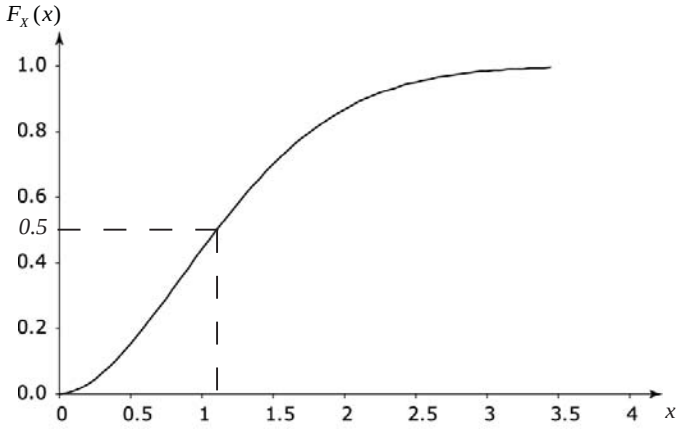


Figure D.13: Illustration of a cumulative distribution function.

D.6 According to the central limit theorem "the probability distribution for the sum of a number of random variables approaches the Normal distribution as the number becomes large".

D.7 The standard Normal distribution is a special version of the Normal distribution. A standardized random variable is a random variable that has been transformed such as its expected value is equal to zero and its variance is equal to one, (*see also Equation D.48*).

D.8 A sequence of experiments with only two possible mutually exclusive outcomes is called a sequence of Bernoulli trials. Typically the two possible events of a Bernoulli trial are referred to as a success or a failure, (*see also section D.4*).

D.9 Poisson process is a family of discrete processes, which may be used for modeling the number of occurrences of events, (*see more in section D.3*).

D.10 The probability of exceeding the value of 5 is calculated as:

$$P(X > 5) = \int_5^{10} \frac{3}{1000} x^2 dx = \frac{x^3}{1000}\Big|_5^{10} = 0.875$$

D.11 The probability that the engine breaks down within 2 years after placed in operation is calculated as:

$$P(T \le 2 years) = F_T(2) = 1 - e^{\left(-\frac{2}{10}\right)} = 0.181$$

D.12 The probability of no snowfall in the next year is equal to 0.067. The probability of exactly 5 snowfalls in the next year is equal to 0.176. The probabilities are calculated as:

$$P(X = 0) = \frac{(5 \cdot 1)^0}{0!} e^{-5 \cdot 1} = 0.0067 \text{ and } P(X = 5) = \frac{(5 \cdot 1)^5}{5!} e^{-5 \cdot 1} = 0.176$$

Module E

E.1 The procedure of establishing a probabilistic model, as described in *section E.1*, consists of five steps:

1) Assessment and statistical quantification of the available data

2) Selection of distribution function

3) Estimation of distribution parameter

4) Model verification and

5) Model updating.

E.2 The probability that the sample average of the steel yield stress will lie within an interval of ±9.8 MPa of the true mean value μ_X is 0.95, (*see also section E.3, Equation (E.22)*).

E.3 The hypothesis testing procedure, as described also in *section E.4*, consists of the following steps/actions:

1) Formulate a null hypothesis, H_0

2) Formulate an operating rule, H_1

3) Select a significance level, α

4) Identify the value resulting in a probability α of performing a Type I error

5) Perform the testing, obtain the sample statistic

6) Judge the null hypothesis

E.4 An engineer tests the null hypothesis that the mean value of the concrete cover depth of a concrete structure corresponds to design assumptions. In a preliminary assessment a limited number of measurements of the concrete cover depth are made, and after performing the hypothesis test the engineer accepts the null hypothesis. After a few years, a comprehensive survey of the concrete cover depth is carried out, i.e. many measurements are made. The survey shows that the mean value of the concrete cover depth does not fulfill the design assumptions. Which of the following statement(s) is(are) correct?

In the preliminary survey the engineer has performed a Type I error. ☐

In the preliminary survey the engineer has performed a Type II error. ☑

In the preliminary survey the engineer has performed a Type I and a Type II error. ☐

E.5 Probability papers are useful for checking the plausibility of a selected distribution family. A probability paper for a given distribution family is constructed such that the cumulative probability distribution function (or the complement) for that distribution family will have the shape of a straight line when plotted on the paper. A probability paper is thus constructed by a non-linear transformation of the y-axis, (*see also section E.5*).

E.6 The data seem to fit well on a straight line and hence the assumption of a Gumbel distribution can be accepted by the engineer.

E.7 The Maximum Likelihood Method (MLM) enables engineers to calculate the distribution parameters of a random variable on the basis of data. Which of the following statement(s) is(are) correct?

The MLM provides point estimates of the distribution parameters. ☑

The MLM provides information about the uncertainty associated with the estimated parameters. ☑

The MLM provides no information about the uncertainty associated with the estimated parameters. ☐

E.8 From past experience it is known that the shear strength of soil can be described by a Lognormal distribution. 15 samples of soil are taken from a site and an engineer wants to use the data in order to estimate the parameters of the Log-normal distribution. The engineer:

may use a probability paper to estimate the parameters of the Lognormal distribution. ☑

may use the maximum likelihood method to estimate the parameters of the Log-normal distribution. ☑

may use the method of moments to estimate the parameters of the Lognormal distribution. ☑

None of the above. ☐

E.9 It is suggested that the data are lumped in a way that each interval contains about 5 or more observations. If the data are realizations from a continuous distribution function, then they must be descritized, (*see also section E.7*).

E.10 Both tests are used to assess the goodness of fit of the assumed model with data. The Chi-square test is used basically for discrete distribution functions, while the Kolmogorov-Smirnov test is used for continuous distribution functions. However,

by descritizing the support of a continuous distribution function, the Chi-square test can be used also for the continuous distribution function. Another difference is that whereas the Chi-square test can be applied for the cases where the distribution parameters are already estimated from data, the Kolmogorov-Smirnov test cannot be applied when the distribution parameters are estimated from data, (*see also section E.7*).

E.11 Following the introduction of the χ^2 goodness of fit test in *section E.7,* the number of degrees of freedom of the Chi-square sample statistic ε_m^2 are $k-1=3-1=2$, where k the number of intervals into which the samples were divided. The null hypothesis H_o that X follows a Normal distribution with the given parameters can be tested using the following operating rule: $P(\varepsilon_m^2 \geq \Delta) = \alpha$, where Δ is the critical value with the sample statistic shall be compared. Using Table T.3 and for a significance level of 5% it is observed that $\Delta = 5.9915$, a value that is larger than $\varepsilon_m^2 = 0.41$. Hence the null hypothesis cannot be rejected at the 5% significance level.

E.12 An engineer wants to examine and compare the suitability of two distribution function model alternatives for a random material property. Measurements are taken of the material property. The engineer uses the two model alternatives to calculate the Chi-square sample statistics and the corresponding sample likelihoods. The results are given in the following table:

Model	Degrees of freedom	Chi-square sample statistic	Sample likelihood
1	2	0.410	0.815
2	1	0.407	0.524

Which of the following statement(s) is(are) correct?

The engineer may accept model 1 at the 5% significance level. ☑

The engineer may accept model 2 at the 5% significance level. ☑

Model 1 is more suitable than model 2. ☑

None of the above. ☐

Module F

F.1 It is convenient to describe failure events in terms of functional relations, which if they are fulfilled define that the considered event will occur. A *failure event* may be described by a functional relation, the limit state function $g(\mathbf{x})$, such as: $F = \{g(\mathbf{x}) \leq 0\}$, where the components of the vector $\mathbf{x}$ are realisations of the so-called basic random variables $\mathbf{X}$ representing all the relevant uncertainties influencing the probability of failure, (*see also section F.2*).

F.2 The reliability index may be defined as the shortest distance between the curve represented by the limit state function and the origin. The reliability index β is related to the probability of failure P_F as: $P_F = \Phi(-\beta)$, (*see also section F.3*).

F.3 The estimate of the failure probability becomes exact as the number of simulation approaches infinity, (*see also section F.6*).

F.4 The reliability index β can be calculated by Equation (F.9), *section F.3*: $\beta = \frac{\mu_M}{\sigma_M}$.

The mean μ_M and standard deviation σ_M of the safety margin $M = R - L$ can be calculated by applying the properties of the expectation operator (*see section D.3*) on the safety margin expression. This gives:

$\mu_M = \mu_R - \mu_L = 30 - 9 = 21 kNm$ *and*

$\sigma_M = \sqrt{\sigma_R^2 + \sigma_L^2} = \sqrt{5^2 + 2^2} = 5.39 kNm$

Hence the reliability index is equal to:

$$\beta = \frac{\mu_M}{\sigma_M} = \frac{21}{5.39} = 3.9 .$$

The annual probability of failure of the timber beam is:

$$P_F = P(M \leq 0) = \Phi(-\beta) = \Phi(-3.9) = 4.8 \cdot 10^{-5}$$

Where $\Phi(-3.9) = 4.8 \cdot 10^{-5}$, can be found from Table T.2.

F.5 The Safety margin can be written as:

$$M = R - S = A \cdot f_y - S = 100 \cdot f_y - 35$$

Since the yield stress f_y is Normal distributed, M is also Normal distributed and its mean and standard deviation can be calculated as follows:

$$\mu_M = E[M] = E[100 \cdot f_y - 35] = 100 \cdot \mu_{f_y} - 35 = 100 \cdot 425 \cdot 10^{-3} - 35 = 7.5KN$$

And the variance is calculated as:

$$\sigma_M^2 = VAR[M] = VAR[100 \cdot f_y - 35] = VAR[100 \cdot f_y] - VAR[35] =$$
$$= 100^2 \cdot \sigma_{f_y}^2 - 0 = 100^2 \cdot (25 \cdot 10^{-3})^2 = 6.25KN^2$$

And the standard deviation is then:

$$\sigma_M = \sqrt{\sigma_M^2} = \sqrt{6.25} = 2.5KN$$

The probability of failure of the rod is then (following Equation (F.8), *section F.3*) :

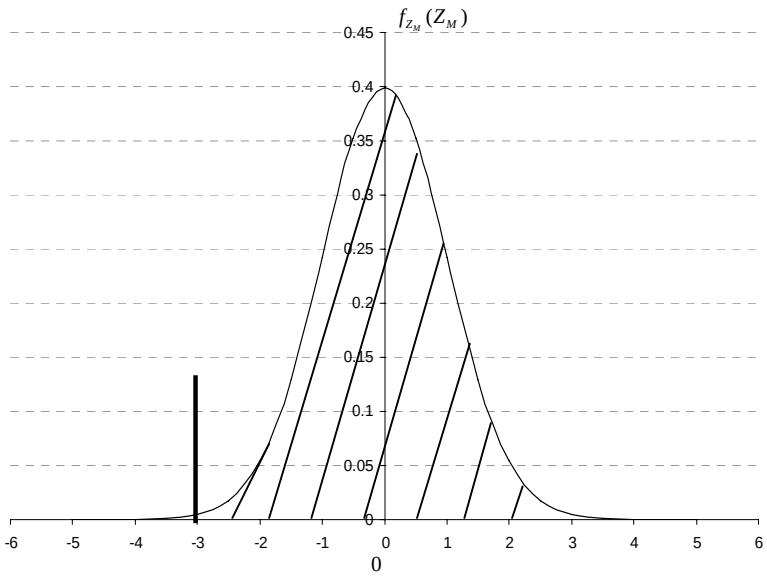
$$P_f = P(M \le 0) = P(Z_M \le \frac{0 - \mu_M}{\sigma_M}) = \Phi(\frac{0 - \mu_M}{\sigma_M}) = \Phi(\frac{0 - 7.5}{2.5}) =$$
$$= \Phi(-3) = 0.00135$$

Whereas the reliability of the rod is simply:

$$\text{Reliability} = 1 - P_f = 1 - 0.00135 = 0.99865$$

(**Note:** The standard Normal distribution value corresponding to -3 is taken from Table T.1)

It is easier to draw the probability density function of the standardized safety margin i.e. of Z_M. The area under the density function to the right of -3 in the x-axis represents the safe region.



F.6 Using the provided Figure and basic principles of geometry it is:

$$\frac{b}{\sin(\beta)} = \frac{c}{\sin(\pi - \alpha - \beta)} \qquad \Rightarrow \qquad b = f(c, \alpha, \beta) = c \cdot \frac{\sin(\beta)}{\sin(\alpha + \beta)}$$

Using the properties of the expectation operator (*see section D.3*) it is:

$$E[b] = E\left[c \cdot \frac{\sin(\beta)}{\sin(\alpha + \beta)} \right] = E[c] \cdot \frac{\sin(\beta)}{\sin(\alpha + \beta)} = 6 \cdot \frac{\sin(1.225)}{\sin(1.225 + 0.813)} = 6.32 \ km$$

While the estimation of the error associated with the measurement of side b is represented by the standard deviation $\sigma[b]$ and is estimated as in the following :

$$V[b] = \left[\frac{\partial f}{\partial c} \right]^2 \cdot \sigma_c{}^2 + \left[\frac{\partial f}{\partial \alpha} \right]^2 \cdot \sigma_\alpha{}^2 + \left[\frac{\partial f}{\partial \beta} \right]^2 \cdot \sigma_\beta{}^2$$

$$\frac{\partial f}{\partial c} = \frac{\sin(\beta)}{\sin(\alpha + \beta)}$$

$$\frac{\partial f}{\partial \alpha} = \left(c \cdot \frac{\sin(\beta)}{\sin(\alpha + \beta)} \right) \frac{\partial}{\partial \alpha} = \left(c \cdot \sin(\beta) \cdot \sin(\alpha + \beta)^{-1} \right) \frac{\partial}{\partial \alpha} = -c \cdot \frac{\sin(\beta) \cdot \cos(\alpha + \beta)}{\left(\sin(\alpha + \beta) \right)^2}$$

$$\frac{\partial f}{\partial \beta}=\left(c\cdot\frac{\sin(\beta)}{\sin(\alpha+\beta)}\right)\frac{\partial}{\partial \beta}=\left(c\cdot\sin(\beta)\cdot\left(\sin(\alpha+\beta)\right)^{-1}\right)\frac{\partial}{\partial \beta}$$

$$=\left(c\cdot\cos(\beta)\cdot\left(\sin(\alpha+\beta)\right)^{-1}+(-1)\cdot\left(\sin(\alpha+\beta)\right)^{-2}\cdot\cos(\alpha+\beta)\cdot(1)\cdot c\cdot\sin(\beta)\right)$$

$$=c\cdot\left(\frac{\cos(\beta)}{\sin(\alpha+\beta)}-\frac{\sin(\beta)\cdot\cos(\alpha+\beta)}{\left(\sin(\alpha+\beta)\right)^{2}}\right)=c\cdot\left(\frac{\cos(\beta)\cdot\sin(\alpha+\beta)-\sin(\beta)\cdot\cos(\alpha+\beta)}{\left(\sin(\alpha+\beta)\right)^{2}}\right)$$

$$=c\cdot\left(\frac{\sin(\alpha+\beta-\beta)}{\left(\sin(\alpha+\beta)\right)^{2}}\right)=c\cdot\left(\frac{\sin(\alpha)}{\left(\sin(\alpha+\beta)\right)^{2}}\right)$$

And eventually it is:

$$V[b]=\left[\frac{\partial f}{\partial c}\right]^{2}\cdot\sigma_{c}{}^{2}+\left[\frac{\partial f}{\partial \alpha}\right]^{2}\cdot\sigma_{\alpha}{}^{2}+\left[\frac{\partial f}{\partial \beta}\right]^{2}\cdot\sigma_{\beta}{}^{2}$$

$$=\left[\frac{\sin\beta}{\sin(\alpha+\beta)}\right]^{2}\cdot\sigma_{c}{}^{2}+\left[\frac{c\cdot\sin\beta\cdot\cos(\alpha+\beta)}{(\sin(\alpha+\beta))^{2}}\right]^{2}\cdot\sigma_{\alpha}{}^{2}+\left[\frac{c\cdot\sin\alpha}{(\sin(\alpha+\beta))^{2}}\right]^{2}\sigma_{\beta}{}^{2}$$

$$=1.0537^{2}\cdot0.005^{2}+3.1894^{2}\cdot0.011^{2}+5.4671^{2}\cdot0.011^{2}=0.004875km^{2}$$

The error in *b* is calculated by:

$$\sigma[b]=\sqrt{0.004875}=0.0698km$$

Module G

G.1 The probabilistic models concerning events of interest and the consequences for each event and action.

G.2 Utility function is a numerical assessment of the preferences of the decision maker, (*see also section G.3*).

G.3 Prior decision analysis is based on existing information and experience for a first estimate of the probability of the considered events. In posterior decision analysis new information is used to update the above probabilities and carry out a reassessment of the decision problem, (*see also sections G.36 and G.7*).

G.4 In pre-posterior decision analysis the decision maker can evaluate whether it is useful or not to "buy" new information that will enable to make her final decision, (*see also section G.8*).

G.6 Using the information provided it is:

$P(SW) = 0.6$
$P(SD) = 0.4$
$P(I_{SD} \mid SD) = 0.75$
$P(I_{SW} \mid SW) = 0.75$
$P(I_{SD} \mid SW) = 1 - P(I_{SD} \mid SD) = 1 - 0.75 = 0.25$
$P(I_{SW} \mid SD) = 1 - P(I_{SW} \mid SW) = 1 - 0.75 = 0.25$

Using the Bayes' Theorem it is:

$$P(SD \mid I_{SD}) = 0.6667 = \frac{P(I_{SD} \mid SD) \cdot P(SD)}{P(I_{SD} \mid SD) \cdot P(SD) + P(I_{SD} \mid SW) \cdot P(SW)} = \frac{0.75 \cdot 0.4}{0.75 \cdot 0.4 + 0.25 \cdot 0.6} = 0.6667$$

$$P(SD \mid I_{sw}) = \frac{P(I_{SW} \mid SD) \cdot P(SD)}{P(I_{SW} \mid SD) \cdot P(SD) + P(I_{SW} \mid SW) \cdot P(SW)} = \frac{0.25 \cdot 0.4}{0.25 \cdot 0.4 + 0.75 \cdot 0.6} = \frac{2}{11} = 0.18182$$

$$P(SW \mid I_{SD}) = \frac{P(I_{SD} \mid SW) \cdot P(SW)}{P(I_{SD} \mid SW) \cdot P(SW) + P(I_{SD} \mid SD) \cdot P(SD)} = \frac{0.25 \cdot 0.6}{0.25 \cdot 0.6 + 0.75 \cdot 0.4} = \frac{1}{3} = 0.3333$$

$$P(SW \mid I_{SW}) = \frac{P(I_{SW} \mid SW) \cdot P(SW)}{P(I_{SW} \mid SW) \cdot P(SW) + P(I_{SW} \mid SD) \cdot P(SD)} = \frac{0.75 \cdot 0.6}{0.75 \cdot 0.6 + 0.25 \cdot 0.4} = \frac{9}{11} = 0.8181$$

And:

$$P(I_{SW}) = P(I_{SW} \mid SW) \cdot P(SW) + P(I_{SW} \mid SW) \cdot P(SW) = 0.75 \cdot 0.6 + 0.25 \cdot 0.4 = 0.55$$

$$P(I_{SD}) = P(I_{SD} \mid SW) \cdot P(SW) + P(I_{SD} \mid SD) \cdot P(SD) = 0.25 \cdot 0.6 + 0.75 \cdot 0.4 = 0.45$$

The event tree can now be filled in. An example of calculation is provided in the following.

Consider the branch associated with the activity "clean up the roof". If the roof is cleaned up there are two events that may occur according to our problem:

- the roof may collapse (due to various reasons)
- the roof will not collapse (survival of the roof)

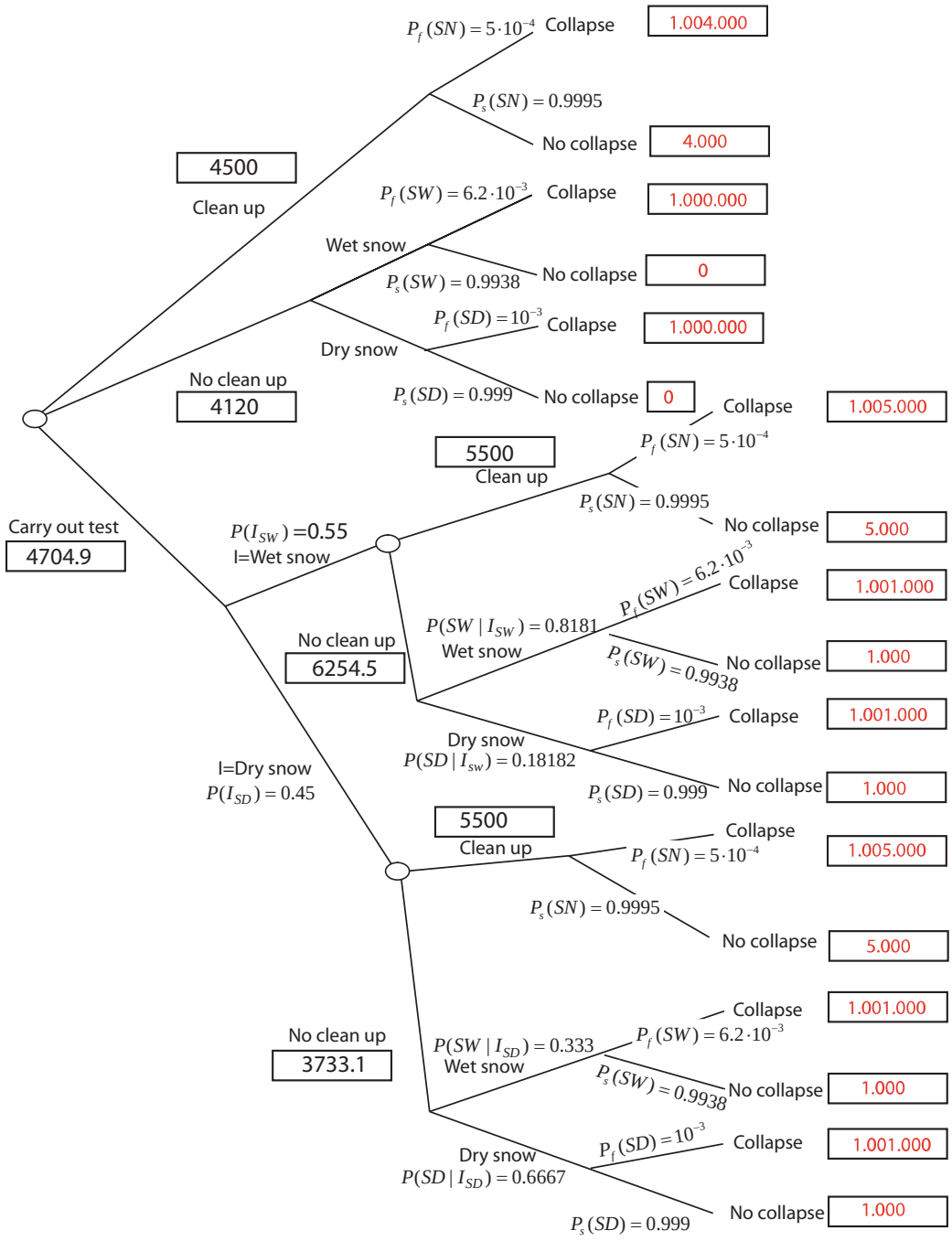
These events are associated with some probability as shown in the event tree branches:

a. $P_f(SN) = 5 \cdot 10^{-4}$ and b. $P_s(SN) = 1 - 5 \cdot 10^{-4} = 0.9995$.

Hence the expected cost of this action is:

$$E[C_{clean\ up}] = P_f(SN) \cdot 1000004 + P_s(SN) \cdot 4000 = 5 \cdot 10^{-4} \cdot 1004000 + 0.9995 \cdot 4000$$
$$= 4500\ CHF$$

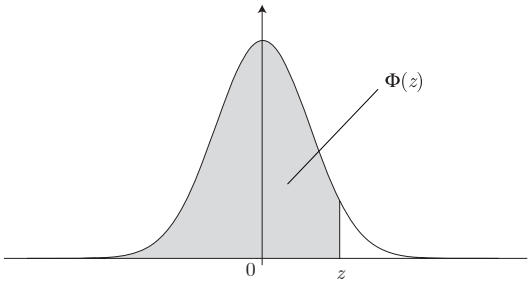
In a similar way the rest of the event tree may be completed.



It can be seen that the action associated with the smaller cost is not to clean up the roof.

MODULE T – TABLES

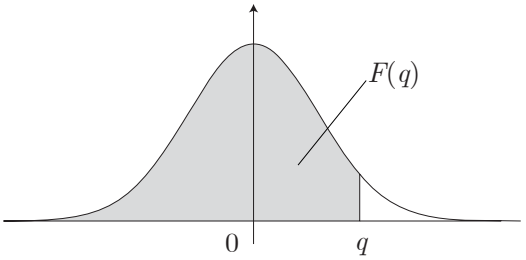
Table T.1: Cumulative distribution function of the standard Normal distribution $\Phi(z)$.



Probability density function of the standard normal random variable.

z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$
0.00	0.5000	0.50	0.6915	1.00	0.8413	1.50	0.9332	2.00	0.9772
0.01	0.5040	0.51	0.6950	1.01	0.8438	1.51	0.9345	2.10	0.9821356
0.02	0.5080	0.52	0.6985	1.02	0.8461	1.52	0.9357	2.20	0.9860966
0.03	0.5120	0.53	0.7019	1.03	0.8485	1.53	0.9370	2.30	0.9892759
0.04	0.5160	0.54	0.7054	1.04	0.8508	1.54	0.9382	2.40	0.9918025
0.05	0.5199	0.55	0.7088	1.05	0.8531	1.55	0.9394	2.50	0.9937903
0.06	0.5239	0.56	0.7123	1.06	0.8554	1.56	0.9406	2.60	0.9953388
0.07	0.5279	0.57	0.7157	1.07	0.8577	1.57	0.9418	2.70	0.9965330
0.08	0.5319	0.58	0.7190	1.08	0.8599	1.58	0.9429	2.80	0.9974449
0.09	0.5359	0.59	0.7224	1.09	0.8621	1.59	0.9441	2.90	0.9981342
0.10	0.5398	0.60	0.7257	1.10	0.8643	1.60	0.9452	3.00	0.9986501
0.11	0.5438	0.61	0.7291	1.11	0.8665	1.61	0.9463	3.10	0.9990324
0.12	0.5478	0.62	0.7324	1.12	0.8686	1.62	0.9474	3.20	0.9993129
0.13	0.5517	0.63	0.7357	1.13	0.8708	1.63	0.9484	3.30	0.9995166
0.14	0.5557	0.64	0.7389	1.14	0.8729	1.64	0.9495	3.40	0.9996631
0.15	0.5596	0.65	0.7422	1.15	0.8749	1.65	0.9505	3.50	0.9997674
0.16	0.5636	0.66	0.7454	1.16	0.8770	1.66	0.9515	3.60	0.9998409
0.17	0.5675	0.67	0.7486	1.17	0.8790	1.67	0.9525	3.70	0.9998922
0.18	0.5714	0.68	0.7517	1.18	0.8810	1.68	0.9535	3.80	0.9999277
0.19	0.5753	0.69	0.7549	1.19	0.8830	1.69	0.9545	3.90	0.9999519
0.20	0.5793	0.70	0.7580	1.20	0.8849	1.70	0.9554	4.00	0.9999683
0.21	0.5832	0.71	0.7611	1.21	0.8869	1.71	0.9564	4.10	0.9999793
0.22	0.5871	0.72	0.7642	1.22	0.8888	1.72	0.9573	4.20	0.9999867
0.23	0.5910	0.73	0.7673	1.23	0.8907	1.73	0.9582	4.30	0.9999915
0.24	0.5948	0.74	0.7704	1.24	0.8925	1.74	0.9591	4.40	0.9999946
0.25	0.5987	0.75	0.7734	1.25	0.8944	1.75	0.9599	4.50	0.9999966
0.26	0.6026	0.76	0.7764	1.26	0.8962	1.76	0.9608	4.60	0.9999979
0.27	0.6064	0.77	0.7794	1.27	0.8980	1.77	0.9616	4.70	0.9999987
0.28	0.6103	0.78	0.7823	1.28	0.8997	1.78	0.9625	4.80	0.9999992
0.29	0.6141	0.79	0.7852	1.29	0.9015	1.79	0.9633	4.90	0.9999995
0.30	0.6179	0.80	0.7881	1.30	0.9032	1.80	0.9641	5.00	0.9999997
0.31	0.6217	0.81	0.7910	1.31	0.9049	1.81	0.9649		
0.32	0.6255	0.82	0.7939	1.32	0.9066	1.82	0.9656		
0.33	0.6293	0.83	0.7967	1.33	0.9082	1.83	0.9664		
0.34	0.6331	0.84	0.7995	1.34	0.9099	1.84	0.9671		
0.35	0.6368	0.85	0.8023	1.35	0.9115	1.85	0.9678		
0.36	0.6406	0.86	0.8051	1.36	0.9131	1.86	0.9686		
0.37	0.6443	0.87	0.8078	1.37	0.9147	1.87	0.9693		
0.38	0.6480	0.88	0.8106	1.38	0.9162	1.88	0.9699		
0.39	0.6517	0.89	0.8133	1.39	0.9177	1.89	0.9706		
0.40	0.6554	0.90	0.8159	1.40	0.9192	1.90	0.9713		
0.41	0.6591	0.91	0.8186	1.41	0.9207	1.91	0.9719		
0.42	0.6628	0.92	0.8212	1.42	0.9222	1.92	0.9726		
0.43	0.6664	0.93	0.8238	1.43	0.9236	1.93	0.9732		
0.44	0.6700	0.94	0.8264	1.44	0.9251	1.94	0.9738		
0.45	0.6736	0.95	0.8289	1.45	0.9265	1.95	0.9744		
0.46	0.6772	0.96	0.8315	1.46	0.9279	1.96	0.9750		
0.47	0.6808	0.97	0.8340	1.47	0.9292	1.97	0.9756		
0.48	0.6844	0.98	0.8365	1.48	0.9306	1.98	0.9761		
0.49	0.6879	0.99	0.8389	1.49	0.9319	1.99	0.9767		

Table T.2: Quantile values of the t-distribution q.

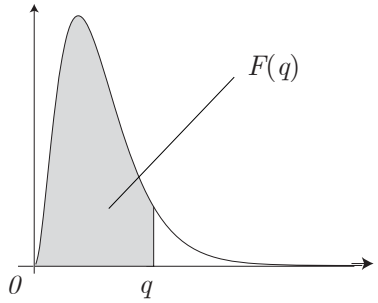


Probability density function of t-distribution.

ν	F(q)=0.75	0.8	0.85	0.9	0.95	0.975	0.99	0.995
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704
50	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678
60	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660
70	0.678	0.847	1.044	1.294	1.667	1.994	2.381	2.648
80	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639
90	0.677	0.846	1.042	1.291	1.662	1.987	2.368	2.632
100	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576

ν : Degrees of freedom.

Table T.3: Quantile values of the Chi-square distribution q.



Probability density function of Chi-square distribution.

ν	F(q)=0.75	0.90	0.95	0.98	0.99	0.995	0.999
1	1.3233	2.7055	3.8415	5.4119	6.6349	7.8794	10.8276
2	2.7726	4.6052	5.9915	7.8240	9.2103	10.5966	13.8155
3	4.1083	6.2514	7.8147	9.8374	11.3449	12.8382	16.2662
4	5.3853	7.7794	9.4877	11.6678	13.2767	14.8603	18.4668
5	6.6257	9.2364	11.0705	13.3882	15.0863	16.7496	20.5150
6	7.8408	10.6446	12.5916	15.0332	16.8119	18.5476	22.4577
7	9.0371	12.0170	14.0671	16.6224	18.4753	20.2777	24.3219
8	10.2189	13.3616	15.5073	18.1682	20.0902	21.9550	26.1245
9	11.3888	14.6837	16.9190	19.6790	21.6660	23.5894	27.8772
10	12.5489	15.9872	18.3070	21.1608	23.2093	25.1882	29.5883
11	13.7007	17.2750	19.6751	22.6179	24.7250	26.7568	31.2641
12	14.8454	18.5493	21.0261	24.0540	26.2170	28.2995	32.9095
13	15.9839	19.8119	22.3620	25.4715	27.6882	29.8195	34.5282
14	17.1169	21.0641	23.6848	26.8728	29.1412	31.3193	36.1233
15	18.2451	22.3071	24.9958	28.2595	30.5779	32.8013	37.6973
16	19.3689	23.5418	26.2962	29.6332	31.9999	34.2672	39.2524
17	20.4887	24.7690	27.5871	30.9950	33.4087	35.7185	40.7902
18	21.6049	25.9894	28.8693	32.3462	34.8053	37.1565	42.3124
19	22.7178	27.2036	30.1435	33.6874	36.1909	38.5823	43.8202
20	23.8277	28.4120	31.4104	35.0196	37.5662	39.9968	45.3147
21	24.9348	29.6151	32.6706	36.3434	38.9322	41.4011	46.7970
22	26.0393	30.8133	33.9244	37.6595	40.2894	42.7957	48.2679
23	27.1413	32.0069	35.1725	38.9683	41.6384	44.1813	49.7282
24	28.2412	33.1962	36.4150	40.2704	42.9798	45.5585	51.1786
25	29.3389	34.3816	37.6525	41.5661	44.3141	46.9279	52.6197
26	30.4346	35.5632	38.8851	42.8558	45.6417	48.2899	54.0520
27	31.5284	36.7412	40.1133	44.1400	46.9629	49.6449	55.4760
28	32.6205	37.9159	41.3371	45.4188	48.2782	50.9934	56.8923
29	33.7109	39.0875	42.5570	46.6927	49.5879	52.3356	58.3012
30	34.7997	40.2560	43.7730	47.9618	50.8922	53.6720	59.7031

ν : Degrees of freedom.

Table T.4: Critical values of the Kolmogorov-Smirnov test.

n	α=0.01	0.02	0.05	0.1	0.2
1	0.995	0.990	0.975	0.950	0.900
2	0.929	0.900	0.842	0.776	0.684
3	0.829	0.785	0.708	0.636	0.565
4	0.734	0.689	0.624	0.565	0.493
5	0.669	0.627	0.563	0.509	0.447
6	0.617	0.577	0.519	0.468	0.410
7	0.576	0.538	0.483	0.436	0.381
8	0.542	0.507	0.454	0.410	0.358
9	0.513	0.480	0.430	0.387	0.339
10	0.489	0.457	0.409	0.369	0.323
11	0.468	0.437	0.391	0.352	0.308
12	0.449	0.419	0.375	0.338	0.296
13	0.432	0.404	0.361	0.325	0.285
14	0.418	0.390	0.349	0.314	0.275
15	0.404	0.377	0.338	0.304	0.266
16	0.392	0.366	0.327	0.295	0.258
17	0.381	0.355	0.318	0.286	0.250
18	0.371	0.346	0.309	0.279	0.244
19	0.361	0.337	0.301	0.271	0.237
20	0.352	0.329	0.294	0.265	0.232
21	0.344	0.321	0.287	0.259	0.226
22	0.337	0.314	0.281	0.253	0.221
23	0.330	0.307	0.275	0.248	0.217
24	0.323	0.301	0.269	0.242	0.212
25	0.317	0.295	0.264	0.238	0.208
26	0.311	0.290	0.259	0.233	0.204
27	0.305	0.284	0.254	0.229	0.200
28	0.300	0.279	0.250	0.225	0.197
29	0.295	0.275	0.246	0.221	0.194
30	0.290	0.270	0.242	0.218	0.190
31	0.285	0.266	0.238	0.214	0.187
32	0.281	0.262	0.234	0.211	0.185
33	0.277	0.258	0.231	0.208	0.182
34	0.273	0.254	0.227	0.205	0.179
35	0.269	0.251	0.224	0.202	0.177
36	0.265	0.247	0.221	0.199	0.174
37	0.262	0.244	0.218	0.196	0.172
38	0.258	0.241	0.215	0.194	0.170
39	0.255	0.238	0.213	0.192	0.168
40	0.252	0.235	0.210	0.189	0.166
$n > 40$	$1.63/\sqrt{n}$	$1.52/\sqrt{n}$	$1.36/\sqrt{n}$	$1.22/\sqrt{n}$	$1.07/\sqrt{n}$

α : Significance level.

n : Sample size.

Table T.5: Gamma function.

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt$$

x	$\Gamma(x)$	x	$\Gamma(x)$	x	$\Gamma(x)$	x	$\Gamma(x)$	x	$\Gamma(x)$	x	$\Gamma(x)$
0.00010	9999.4	0.0010	999.42	0.010	99.43	0.10	9.514	1.0	1.000	5.5	52.3
0.00012	8332.8	0.0012	832.76	0.012	82.77	0.12	7.863	1.1	0.951	5.6	61.6
0.00014	7142.3	0.0014	713.71	0.014	70.87	0.14	6.689	1.2	0.918	5.7	72.5
0.00016	6249.4	0.0016	624.42	0.016	61.94	0.16	5.811	1.3	0.897	5.8	85.6
0.00018	5555.0	0.0018	554.98	0.018	55.00	0.18	5.132	1.4	0.887	5.9	101.3
0.00020	4999.4	0.0020	499.42	0.020	49.44	0.20	4.591	1.5	0.886	6.0	120.0
0.00022	4544.9	0.0022	453.97	0.022	44.90	0.22	4.150	1.6	0.894	6.1	142.5
0.00024	4166.1	0.0024	416.09	0.024	41.11	0.24	3.786	1.7	0.909	6.2	169.4
0.00026	3845.6	0.0026	384.04	0.026	37.91	0.26	3.478	1.8	0.931	6.3	201.8
0.00028	3570.9	0.0028	356.57	0.028	35.16	0.28	3.217	1.9	0.962	6.4	240.8
0.00030	3332.8	0.0030	332.76	0.030	32.78	0.30	2.992	2.0	1.000	6.5	287.9
0.00032	3124.4	0.0032	311.93	0.032	30.70	0.32	2.796	2.1	1.046	6.6	344.7
0.00034	2940.6	0.0034	293.54	0.034	28.87	0.34	2.624	2.2	1.102	6.7	413.4
0.00036	2777.2	0.0036	277.20	0.036	27.24	0.36	2.473	2.3	1.167	6.8	496.6
0.00038	2631.0	0.0038	262.58	0.038	25.77	0.38	2.338	2.4	1.242	6.9	597.5
0.00040	2499.4	0.0040	249.43	0.040	24.46	0.40	2.218	2.5	1.329	7.0	720.0
0.00042	2380.4	0.0042	237.52	0.042	23.27	0.42	2.110	2.6	1.430	7.1	869.0
0.00044	2272.2	0.0044	226.70	0.044	22.19	0.44	2.013	2.7	1.545	7.2	1050
0.00046	2173.3	0.0046	216.82	0.046	21.21	0.46	1.925	2.8	1.676	7.3	1271
0.00048	2082.8	0.0048	207.76	0.048	20.30	0.48	1.845	2.9	1.827	7.4	1541
0.00050	1999.4	0.0050	199.43	0.050	19.47	0.50	1.772	3.0	2.000	7.5	1871
0.00052	1922.5	0.0052	191.74	0.052	18.70	0.52	1.706	3.1	2.198	7.6	2275
0.00054	1851.3	0.0054	184.61	0.054	17.99	0.54	1.645	3.2	2.424	7.7	2770
0.00056	1785.1	0.0056	178.00	0.056	17.33	0.56	1.589	3.3	2.683	7.8	3377
0.00058	1723.6	0.0058	171.84	0.058	16.72	0.58	1.537	3.4	2.981	7.9	4123
0.00060	1666.1	0.0060	166.10	0.060	16.15	0.60	1.489	3.5	3.323	8.0	5040
0.00062	1612.3	0.0062	160.72	0.062	15.61	0.62	1.445	3.6	3.717	8.1	6170
0.00064	1561.9	0.0064	155.68	0.064	15.11	0.64	1.404	3.7	4.171	8.2	7562
0.00066	1514.6	0.0066	150.94	0.066	14.64	0.66	1.366	3.8	4.694	8.3	9281
0.00068	1470.0	0.0068	146.49	0.068	14.19	0.68	1.331	3.9	5.299	8.4	11406
0.00070	1428.0	0.0070	142.29	0.070	13.77	0.70	1.298	4.0	6.000	8.5	14034
0.00072	1388.3	0.0072	138.32	0.072	13.38	0.72	1.267	4.1	6.813	8.6	17290
0.00074	1350.8	0.0074	134.57	0.074	13.00	0.74	1.239	4.2	7.757	8.7	21328
0.00076	1315.2	0.0076	131.01	0.076	12.65	0.76	1.212	4.3	8.855	8.8	26340
0.00078	1281.5	0.0078	127.64	0.078	12.32	0.78	1.187	4.4	10.136	8.9	32569
0.00080	1249.4	0.0080	124.43	0.080	12.00	0.80	1.164	4.5	11.632	9.0	40320
0.00082	1218.9	0.0082	121.38	0.082	11.69	0.82	1.142	4.6	13.381	9.1	49974
0.00084	1189.9	0.0084	118.48	0.084	11.40	0.84	1.122	4.7	15.431	9.2	62011
0.00086	1162.2	0.0086	115.71	0.086	11.13	0.86	1.103	4.8	17.838	9.3	77036
0.00088	1135.8	0.0088	113.07	0.088	10.87	0.88	1.085	4.9	20.667	9.4	95809
0.00090	1110.5	0.0090	110.54	0.090	10.62	0.90	1.069	5.0	24.000	9.5	119292
0.00092	1086.4	0.0092	108.13	0.092	10.38	0.92	1.053	5.1	27.932	9.6	148696
0.00094	1063.3	0.0094	105.81	0.094	10.15	0.94	1.038	5.2	32.578	9.7	185551
0.00096	1041.1	0.0096	103.60	0.096	9.93	0.96	1.025	5.3	38.078	9.8	231792
0.00098	1019.8	0.0098	101.47	0.098	9.72	0.98	1.012	5.4	44.599	9.9	289868