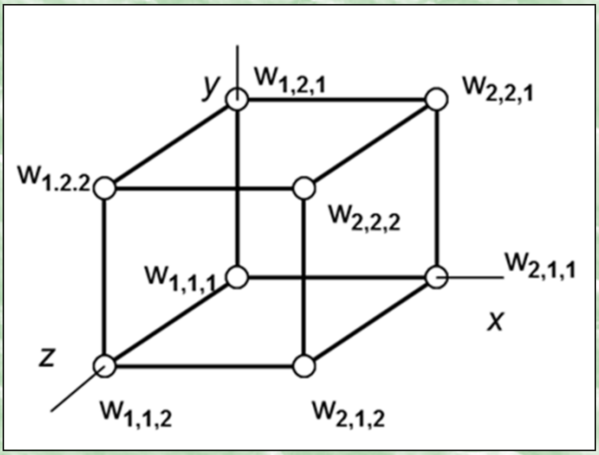
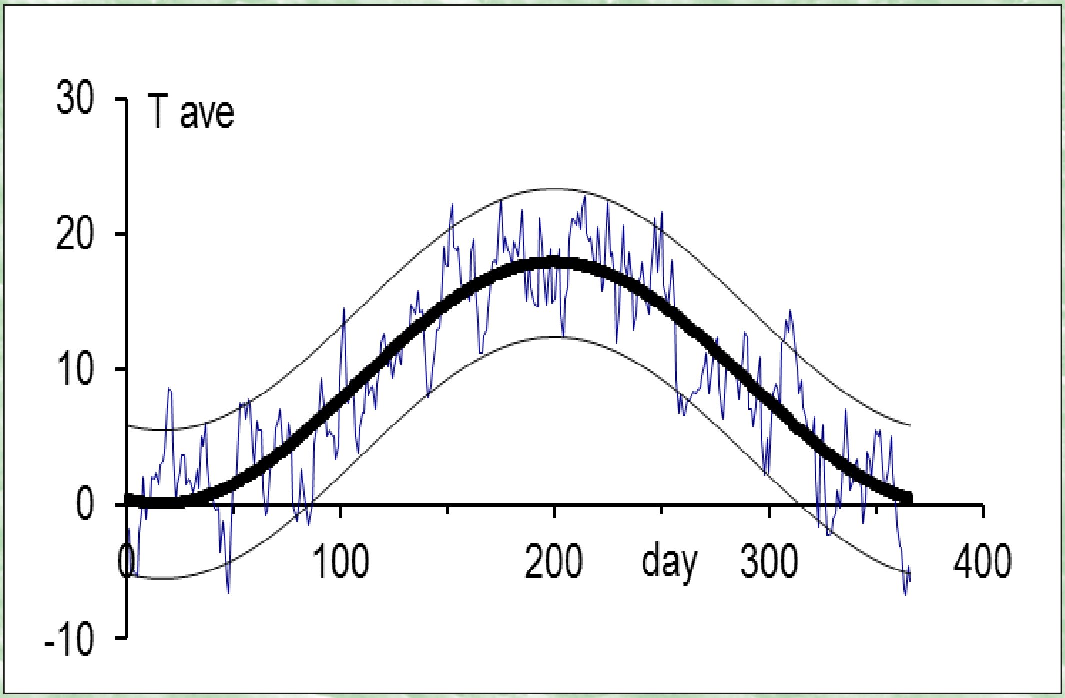


Jaroslav Menčík

Introduction to Experimental Analysis



University of Pardubice, 2017

Reviewers:

Prof. Dr.Rer.Nat. Dietrich Munz
Doc. Ing. Jiří Němeček, PhD.
Prof. Michael V. Swain

ISBN 978-80-7560-068-4 (print)
ISBN 978-80-7560-069-1 (pdf)

The book is freely accessible on http://hdl.handle.net/10195/66961.

Introduction to Experimental Analysis

This book brings a review of basic methods for design and evaluation of experiments. It shows the principal steps and explains the causes of dispersion and errors of the measured values, as well as statistical methods for their evaluation. It is shown how the measured values can be described by simple characteristics or probability distributions. Then, characterisation of relations between investigated quantities is described, including fitting of the data by regression functions. It is explained how confidence intervals can be created, which contain the true values of the parameters, and how many tests are necessary for obtaining the results with the demanded accuracy or for the verification of a certain hypothesis. One chapter is devoted to the theory of similarity and dimensional analysis, which help one to reduce the extent of experiments and make the results more general. Other chapters explain the analysis of variance, design of experiments, and experimental finding of a maximum or minimum. These procedures facilitate finding of the most important factors and optimum parameters. Sensitivity analysis shows how variations of the input quantities cause deviations of the investigated quantity from the optimum or nominal value. The last chapter is devoted to the efficient tools for the study of random influences, such as the Monte Carlo simulation technique. The use of the described methods is illustrated on examples and the individual chapters are supplemented by references.

The pdf version is freely accessible on http://hdl.handle.net/10195/66961.

Acknowledgment

During preparation of this book, I used the experience gained during my work and teaching at various institutions. I am grateful to my colleagues for their advice, as well as to the students for their questions. Special thanks belong to Prof. Milan Lánský and Prof. Miroslav Steiner of the University of Pardubice, Prof. Michael Swain of the University of Sydney, Prof. Dietrich Munz of the Research Centre of Karlsruhe (KIT), Dr. Andy Bushby of the University of London, Prof. Hitoshi Tokura of the Tokyo Institute of Technology, Dr. Gaston Rauchs of the Luxembourg Institute of Science and Technology, Prof. Eckhard Quandt of the University Kiel, and Prof. Jiří Němeček from the Czech Technical University, Prague. Thanks also belong to my former coworkers Ing. Radovan Drahorád and Ing. Miroslav Matušek, CSc., from the research institute VÚSAB Jablonec n.N., where I started my research activities.

Pardubice, Czech Republic, 2017 Jaroslav Menčík

CONTENTS

There is nothing more practical in the world than good theory.

Ludwig Boltzmann

Measure everything measurable, and the not measurable make to be measurable.

Galileo Galilei

1. Introduction

The main purpose of research is to obtain information on the investigated object. Often, also, it is necessary to find optimum composition of a compound or optimum parameters of a process, a component or another object. The demanded information can be obtained by theoretical activity, by observation and by experiment. **Theoretical activity** means work with abstract models. In **observation**, our attitude is passive; we just observe what is happening and record characteristic phenomena and values of variables, without intervening into the investigated object or changing the conditions during the observation. In some cases, observation is the only possibility for obtaining information. Examples are the study of properties of stars or human society, or the development of new kinds of medicine that could cure, but also kill. **Experiment** is a purposeful activity, which should bring deeper insight and more information. It is a series of activities enabling systematic observation with controlled action on the investigated real object or a model of the real object. The **model** can be physical (built from real materials) or computer (simulation). Observation and experiment enable collecting of input data and verification of a hypothesis on the investigated object.

Experiments are very important for providing information. Historically, various methods have been developed, which make experimenting and the evaluation of results more systematic and efficient. Every year, many students and other people become engaged in experimental research and must learn how to do it. And still valid is the experience made by Pascal: “Only at the end of work we know how we should have started”. This book wants to mitigate this problem by presenting a brief review of the basic methods and approaches.

The author has spent many years in applied research. Later, at the University of Pardubice, he gave lectures on research methods for students from various countries and branches of science and engineering. With this extensive experience he decided to prepare a concise book for students and other people who wish to have some insight as to how experiments can be effectively organised and evaluated – regardless their professional orientation. Therefore, he has placed emphasis on explaining the basic terms and universal procedures useful for various branches of research. In order to enable practicing the described methods for readers without special tools, the examples shown in the book can be solved with the help of “omnipresent” Excel.

This book on experimental analysis is divided into fourteen chapters. This first chapter introduces the topic and the arrangement of the book. The second chapter outlines the principal steps of experimental research. Chapter 3 identifies what kinds of errors can appear in an experiment or measurement, and shows also ways for their avoidance or reduction. The following chapter explains the basic terms of probability and statistical methods usual in experimental analysis – just to help those who are not familiar with these important tools. Chapter 5 indicates the determination of important characteristics of an investigated quantity, such as the average, a histogram, or the parameters of probability distribution. Chapter 6 is devoted to various characteristics of relationships between two or more quantities. Chapter 7 explains the fitting of empirical data by regression functions and determination of their parameters. The applications are shown on specific examples. Chapter 8 is devoted to the determination of the repetition of experiments and measurements needed for obtaining results with demanded accuracy. It explains the confidence intervals and also the statistical tests for proving whether the difference between two procedures is significant or not. Chapter 9 presents the principles of the similarity theory and dimensional analysis. They both are very powerful tools that can reduce the extent of necessary experiments and make the results more general. Chapter 10 explains the principles of the analysis of variance, which can reveal the significance of various factors. The following chapter 11 is devoted to the design of experiments (DOE), which aims at fast and efficient revelation of the most influential factors and finding the best parameters of a structure or conditions of a process. Chapter 12 shows experimental procedures for finding a maximum or minimum of a certain function. Chapter 13 (Sensitivity analysis) shows how the variations of input variables contribute to the deviations of the output quantity from its nominal or optimum value. The last chapter (14) is devoted to very efficient tools for the investigation of the behaviour of random quantities, namely the simulation method Monte Carlo and Latin Hypercube Sampling. The individual chapters are complemented with references.

2. Principal Steps of Experimental Research

Every experimental research consists of the following stages: preparation, realisation of experiments and measurements, evaluation of results, formulation of conclusions, and publication of results. The individual stages will be described here in detail.

1) Preparation

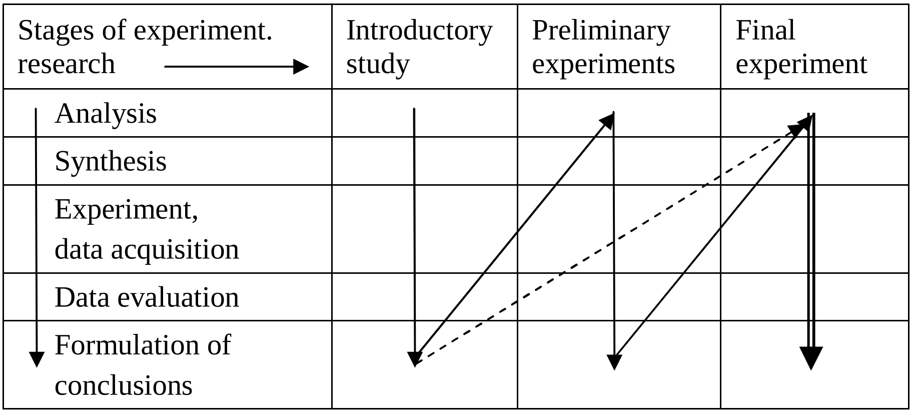
The preparation of experiments consists of the following steps:

- Familiarisation with the problem and its general analysis, formulation of the object and task of the investigation.
- Selection of suitable kinds of experiments and measuring methods with respect to available possibilities, including equipment, finance and time.
- Choice of characteristic physical and other quantities describing the behaviour and important properties of the investigated object or phenomenon. Sometimes it is obvious from the beginning, what quantities will be used. Sometimes not, especially when studying something quite new. In some cases, we must even create a new quantity. Do not worry; remember Galileo Galilei: “Measure everything measurable, and the unmeasurable make to be measurable!” For example, who would imagine (a hundred years ago) that the quantity of information could be measured!
- Preparation of the plan of the experiments and measurements, including their extent (choice of the range and number of the levels of the measured quantities, and the number of experiments and tests).
- Preparation of devices and necessary equipment, obtaining of specimens.

The detailed plan of all experiments, including the time schedule, must be recorded in advance.

2) Realisation of experiments, observations and measurements

The experiments are carried out according to the prepared plan. However, research is always accompanied by some uncertainty, and often the initial plan and time schedule must be modified with respect to the achieved results. Sometimes, therefore, the experiments are done in two or even more stages [1], as depicted in Fig. 2.1. The experiments in the first stage are of limited extent and serve for better determination of the extent and conditions of the remaining experiments.



Stages of experiment. research →	Introductory study	Preliminary experiments	Final experiment
Analysis			
Synthesis			
Experiment, data acquisition			
Data evaluation			
Formulation of conclusions			

Figure 2.1. *Stages of experimental research.*

All experiments should be described in a research or operational log-book. Such records contain the following information: date and time of the tests, the list of used devices (including their types, series numbers and arrangement), the list of participating persons (including their role), and the description of the experiments, results and any associated comments or remarks regarding the observations. If possible, the records are done using a computer or laptop, but they can also be written by hand. Sometimes, pre-printed forms are used. Photos and sketches are useful. The experimenter can also use a video- or tape recorder, and comment there on his or her observations. All this should be done with care, because sometimes it is not possible to repeat the measurements or observations if the original data were lost. In general, the experiments and the results should be described in sufficiently explanatory manner to enable them to be repeated.

3) Evaluation of results

The processing of the obtained data includes their sorting, creation of tables and diagrams, analysis of the results and proposal and verification of the relationships among the investigated quantities, including the determination of constants in regression functions, construction of confidence intervals and testing of various hypotheses about the investigated phenomena.

4) Formulation of conclusions

The evaluated results serve for the formulation of conclusions and for the preparation of a plan of further works. Then, a report can be written. Often, the results are published in a form of a presentation or poster at a conference or a paper in a journal.

5) Publication of results

Publishing is very important especially in the scientific community (universities, research institutions), but the reports, prepared in a readable form, are important for development departments as well. When preparing the information on our research for publication, it is useful to adhere to certain well-proven rules. A scientific paper is usually arranged in the following manner:

1. Introduction, a brief formulation of the task of the work.
2. A review of the state of knowledge on the topic, e.g. a review of relevant publications (books, papers in journals and conference proceedings, research reports).
3. A detailed description of the used methods, devices and procedures.
4. Description of the experiments and measurements.
5. Evaluation and analysis of the results.
6. Conclusions.
7. List of references.

Sometimes, acknowledgment is placed at the end of paper (before the references). Here, the author can express thanks for help, for the support from a grant project or other sources, and also for the permission to publish some results or parts of other authors´ works; in such cases one must always cite properly the source. If we want to take over a full figure or data from another paper or a book, we should also get the permission of the copyright owner (and mention it in the acknowledgment). And – our text must be free of any plagiarism.

As regards the details of preparation of a paper for publication in a scientific journal, the author should always consult the guidelines for the arrangement of papers in that journal, including the style of references, page or word limitation, form of the abstract and number of figures allowed; various journals use various styles. Often, Instructions for Contributors are given at the pertinent web pages of the relevant journal.

An example – Proposal of experiments

In any investigation several possibilities usually exist as how to obtain the required information. The investigator has to decide which method to use with respect to his or her experience, the money and time available for this research, the equipment that is available "in house" or can be purchased or hired, and the requirements on the results with respect to their importance and accuracy. The diversity of possibilities of experimental research can be demonstrated on a simple problem – ascertaining the technical condition of a combustion engine. The relevant information can be obtained from: 1) Power characteristics (power and torque as functions of RPM measured by motor brake), or 2) Compression pressure and the tightness of combustion chamber, or 3) Noise and vibrations of the engine or its parts, 4) Consumption of fuel and lubricants, 5) Condition of lubricants (chemical composition, content of metallic particles), 6) Composition of exhaust gasses or other exhalations (CO, NOx…), 7) The power necessary to rotate the idle engine, which characterises the mechanical losses, 8) Wear of the cylinders (measured directly or from the metal particles in the oil)... Any of these possibilities can yield less or more relevant information, and the choice is the matter of the investigator.

The reader is encouraged to propose further methods.

Models and simulation

A frequent task in research is creation of a suitable model of the studied process or object, or a model of the influence of the important factors on a certain property or phenomenon. On the other hand, the properties or phenomena are often investigated by means of an appropriate model or by simulation. For better understanding, some terms from this area will be explained here.

A **model** is a simplified object or a system, which can help in the analysis of a problem, usually at lower cost and in shorter time. It represents a system or its part and can be created in physical form (e.g. from metal or wood) or in mathematical form suitable for demonstrating its behaviour. **Simulation** involves subjecting the model to various inputs or conditions and observing how it behaves. It can deal with physical models subjected to the actual environment, or with mathematical models subjected to mathematical disturbance functions that simulate the expected conditions.

A model can be descriptive or predictive. A **descriptive model** helps to understand a real-world object, system or phenomenon (e.g. a cutaway model of an engine). A **predictive model** helps to understand and predict its performance.

Models can be classified as static or dynamic, deterministic or probabilistic, and iconic or analogue or symbolic. Properties of **static models** do not change with time, while **dynamic models** consider time-varying effects. **Deterministic models** are used if the outcome of the investigated event occurs with certainty. **Probabilistic models** are necessary if these events or values occur with some probability. **Iconic model** looks like a real thing (for example a scale model of an aircraft for wind tunnel tests). **Analogue models** are those that behave like real systems; however, such a model does not need to look like the real system it represents. There are many analogies between physical phenomena; well-known is the membrane analogy for study of the twist of bars via the response of inflated membrane of similar shape. **Symbolic models** are abstractions of the important quantifiable components of a certain system. A mathematical equation expressing the dependence of the output parameters on the input parameters is a symbolic model. One can distinguish between **theoretical models**, which are based on universally accepted laws of nature, and **empirical models**, which are the approximate mathematical representations based on experimental data. Both kinds of models are often denoted as **mathematical models**.

In mathematical modelling the parts of the system are represented by idealised elements, which have the essential characteristics of the real components and whose behaviour can be described by mathematical equations. Only the simplest models can be studied by classical analytic methods. Computers have greatly expanded the use of **mathematical modelling**. The numerical methods and the ease with which they can test many specific states of the model have firmly established **computer modelling** and **simulation** as powerful tools in research or design. The ability to simulate the operation of a system via a mathematical model is a great advantage in providing information, at lower cost and in shorter time than if experimentation with real objects were used. Moreover, there are situations in which experimentation is impossible because of cost, safety, or time. For example, airline pilots train on flight simulators and nuclear power plant operators learn from reactor simulators.

More about organisation of experiments can be found in literature, for example [1 – 4].

Before we start explaining the individual methods of experimental analysis, let us make this serious topic less serious: Tibor Dévényi [5] likened scientific activity to the work of a four-stroke engine: 1. Intake (= study of the literature), 2. Compression (= making experiments, measurements and analysis of the results), 3. Ignition and combustion (= getting an idea, evaluation of the results), 4. Exhaust (= publication of the results). The similarity is obvious. However, this does not mean that our publications might be as harmful as exhaust gasses.

References to Chapter 2

1. Bernard, J.: Technical experiment (In Czech: Technický experiment). ČVUT, Praha, 1999. 74 p.
2. Montgomery, D. C.: Design and analysis of experiments. Wiley, New York, 2012 (8th edition). 730 p.
3. Kropáč, O.: Methods of experimental research. (In Czech: Metody experimentálního výzkumu.) ČVUT, Praha, 1979. 139 p.
4. Dieter, G. E.: Engineering design. 2nd Edition. McGraw-Hill, New York, 1991. 721 p.
5. Dévényi, T.: Career of Dr. Géza Ezésez or scientists and rodents. (In Hungarian: Dr. Ezésez Géza karrierje avagy Tudósok és rágcsálók.) Gondolat, Budapest, 1975. 206 p.

3. Errors and Variance of Measured Values

Accurate values of the investigated quantities are known only seldom. Generally, three kinds of values can be distinguished [1, 2]:

- actual,
- measured,
- used in the following calculations.

The measured values often differ from the accurate ones due to various errors appearing in the measurement. The reasons for the errors will be discussed in the next paragraphs. However, even if the measurement is accurate, the measured values can less or more vary even if the tests or measurements are repeated under the same conditions. One reason is inherent variability of the measured quantity or phenomenon. For example, the strength of a brittle material varies from one tested piece to another, the reason being different size of material defects responsible for the strength and fracture of the individual samples. Also the conditions of the individual tests can slightly vary, for example the temperature or humidity of the environment; sometimes the measurement is influenced by vibrations or other factors. And, finally, the values used in various subsequent calculations can differ from the measured ones because usually the average value or a certain quantile is used instead of the individual values; examples are coefficient of thermal expansion or nominal strength of a material.

Let us now look at the errors in measurements. Three kinds of errors can be distinguished: gross, systematic and random [3 – 5].

A **gross error** appears as a value obviously out of the common range of other values (Fig. 3.1). Gross errors arise due to inattention in reading the measured values, by using a wrong range of the measuring device, or by a technical fault. They can be revealed by repeating the measurement, by visual check of the plotted series of data, or by statistical tests for extreme values, so-called outliers [6].

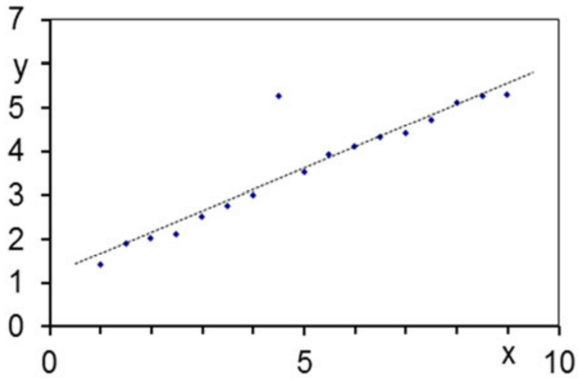


***Figure 3.1.** Example of an outlier.*

Systematic errors arise due to permanent influence of some hidden factors (for example higher temperature or inaccuracy of the measuring device. They cause permanent shift of the measured values, either positive or negative. They cannot be revealed by repeating the experiment, but by the use of another method or conditions of the experiment.

The main causes of systematic errors are:

1) imperfection of our senses (vision, hearing), bad mental condition of the personnel (e.g. work under stress or in a hurry, tiredness, exhaustion),
2) inaccuracy of measuring devices and methods,
3) impossibility to arrange suitable conditions (temperature, pressure, humidity, no parasitic vibrations),
4) the measurement itself can influence the measured quantity (examples: a relatively heavy sensor attached to a light component changes its dynamic characteristics, electric current can increase the temperature of a strain gauge and thus also its resistance).
5) inappropriate method or approximation used in the data processing (e.g. the regression function is used in a wider interval than from which its constants were determined). Low numbers of the digits in calculations (errors due to rounding can sum up in chained calculations) – see at the end of this chapter.

Systematic errors can be avoided in the following ways:

- Mental well-being of the personnel, without stresses during the experiments and measurements.
- Use of sufficiently accurate devices. A rule of thumb says that the common error of the measuring device should be at least ten times smaller than the acceptable error in the determination of the measured quantity. For example, if the thickness of a certain component should be determined with accuracy 0.01 mm, a gauge with accuracy not worse than 0.001 mm must be used. Important devices must be calibrated from time to time.
- The individual members in the measuring chain “sensor - connecting cables - amplifier - measuring device…” are arranged in series and their errors and inaccuracies sum up. The most efficient way for improvement is to replace the “weakest” member by a more accurate. The researcher thus should know their accuracies. In dynamic problems, devices with appropriate dynamic characteristics should be used.
- Exclusion of the undesired influence by suitable arrangement of the test (for example, making all measurements at constant temperature).
- Elimination of the undesired influence by recalculation of the measured data using correction factors (for temperature, e.g.).
- Permanent balancing of the experiment (e.g. the use of Wheatstone bridge circuit).
- Randomisation of the experiments, i.e. the use of random combinations of the values of individual input variables in the sequential series of tests.
- Use of sufficiently high number of digits, especially if the measured values are processed further (see below).

Random errors. These errors are caused by random influences that cannot be controlled. Their magnitude varies from one test to another. They can be revealed by repetition of the tests, and the repetition is also used to reduce their influence. This improvement can be achieved by the methods of mathematical statistics, for example by determining confidence interval that contains the pertinent value with high probability. For more, see the next chapter and Chapters 7 and 8).

REMARK. Several words can be said here on the **optimal number of digits used in the processing of measured values**. We say that a number has *n* ***significant digits*** if its absolute error does not exceed half of the order of the *n*-th digit. If the input has *n* significant digits, not more than *n* digits will be significant in the final result. More digits do not increase the accuracy of the result. If the result should have *n* significant digits, all intermediate calculations must have (at least) $n + 1$ digits; the result is then rounded to *n* digits. In multiplying and dividing, the individual factors are rounded so as to have (at least) one digit more than the factor with the lowest number of significant digits (i.e. with the largest relative error). For more, see for example [7].

References to Chapter 3

1. Pechoč, V.: Evaluation of measurement and computing methods in chemical engineering. (In Czech: Vyhodnocování měření a početní metody v chemickém inženýrství.) SNTL, Praha, 1981. 226 p.
2. Bernard, J.: Technical experiment (In Czech: Technický experiment). ČVUT, Praha, 1999. 74 p.
3. Handbook of measuring technology for machinery and energetics. (In Czech: Příručka měřicí techniky pro strojírenství a energetiku.) SNTL, Praha, 1965. 928 p.
4. Jenčík, J., and Kuhn, L.: Technical measurements in mechanical engineering. (In Czech: Technická měření ve strojnictví.) SNTL, Praha, 1982. 584 p.
5. Taylor, J. R.: An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements. University Science Books, Herndon, 1997. 327 p.
6. Kupka, K.: Statistical quality control. (In Czech: Statistické řízení jakosti.) Trilobyte, Pardubice, 1997. 191 p.
7. Nekvinda, M., Šrubař, J., and Vild, J.: Introduction to numerical mathematics. (In Czech: Úvod do numerické matematiky.) SNTL, Praha, 1976. 288 p.

4. Basics of Probability and Statistics for Experimental Research

The values of many quantities, as well as occurrence of various events, are accompanied by uncertainty. This is due to factors that we cannot control, and call them therefore random. For better work with them we use the concept of probability and statistical methods. The corresponding procedures can help in solving many problems. As computers can do all laborious work, the only thing a user of probabilistic methods needs is some understanding of the basic concepts. This chapter offers a brief overview of principal terms, such as random quantity, probability, population, sample, average, mean, variance, standard deviation, coefficient of variation, probability density, distribution function, quantile, critical value, confidence interval and testing of hypotheses. Important probability distributions are also shown. Details can be found in statistical literature, for example [1 – 4].

Probability is a quantitative measure of possibility that a random event occurs. The simplest definition of probability P is based on numerous occurrence of an event or repetition of a trial:

$$P \approx n / N \tag{4.1}$$

N is the total number of trials (assumed very high, $N \to \infty$) and n is the number of trials with certain outcome, for example a tossed coin with the eagle on the top, the number of days with the maximum temperature higher than 20°C, or the number of defective components in a batch. Probability is a dimensionless quantity that can attain values between 0 and 1; zero denotes the impossible event and 1 a sure event. **Random variable** is a variable, which can attain various values with certain probabilities. Random quantities are **discrete** or **continuous**. Examples of a **discrete random quantity** are the number of fatalities in traffic accidents or the number of loading cycles of a machine till failure. **Continuous random quantity** can attain any value (in some interval), for example strength of a material, wind velocity, temperature, length, weight…, time to failure, duration of a repair, or probability of failure. Some examples are depicted in Fig. 4.1.

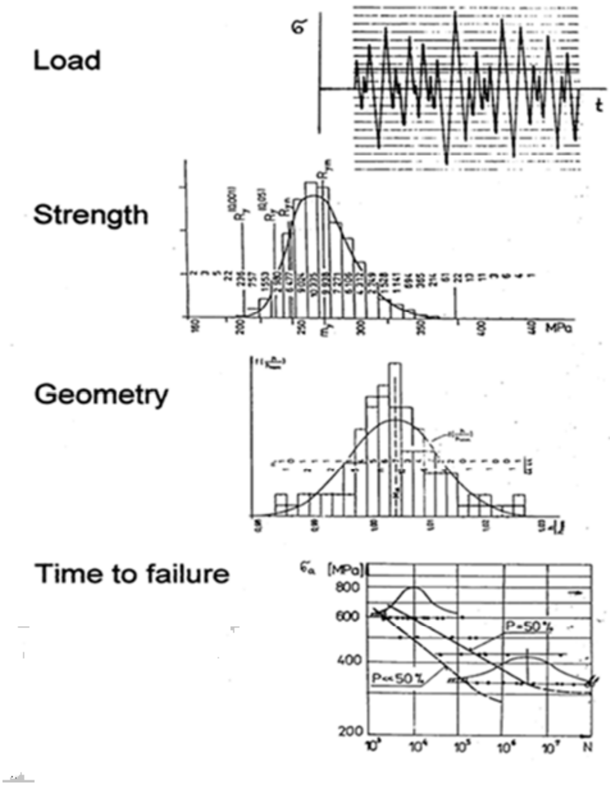


Figure 4.1. *Examples of random quantities [5].*

Random quantities can be described by **probability distribution** or by single numbers, called **parameters**, if they are related to the **population** (= the set of all possible elements or values of the investigated quantity), or **characteristics**, if they are calculated from a **sample** of limited size. Parameters are denoted by Greek letters and characteristics by Latin letters.

Description by parameters

The main parameters (or characteristics) of random quantities are given below, with the formulae for calculation from samples of limited size.

Mean μ (or **average value** $\overline{x}$) characterises the position of the quantity on numerical axis; it corresponds to its centroid,

$$\mu = \int_{-\infty}^{+\infty} x\, f(x) dx\,, \quad \bar{x} = \frac{\sum x_j}{n} \tag{4.2}$$

x_j is *j*-th value; *n* – size of the sample.

Variance σ^2 (or s^2) – characterises the dispersion of the quantity, and is calculated as

$$\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x) dx\,; \quad s^2 = \frac{\sum (x_j - \bar{x})^2}{n-1} \tag{4.3}$$

Standard deviation σ (or s) is defined as the square root of variance,

$$\sigma = \sqrt{\sigma^2}\,, \quad s = \sqrt{\frac{\sum (x_j - \bar{x})^2}{n-1}} \tag{4.4}$$

It has the same dimension as the investigated variable *x* and therefore it is used for the characterization of dispersion more often than variance.

Coefficient of variation *v* characterizes the relative dispersion, compared to the mean value,

$$v = \frac{s}{\bar{x}} \tag{4.5}$$

It can be used for comparison of random variability of various quantities.

A disadvantage of the average value $\bar{x}$ is its sensitivity to extreme values; addition of a very high or low value can cause its significant change. Less sensitive (i.e. robust) characteristic of the "mean" of a series of values is the **median *m***. This is the value in the middle of the series of data ordered from minimum to maximum; for example *m* = 4 for the series 2, 6, 1, 8, 10, 4, 3.

Description by probability distribution

More comprehensive information is obtained from probability distribution, which informs how a random variable is distributed along the numerical axis. For discrete quantities, **probability function *p*(*x*)** is used (Fig. 4.2), which expresses the probabilities that the random variable ***x*** attains the individual values x^*,

$$p(x^*) = P(x = x^*) \tag{4.6}$$

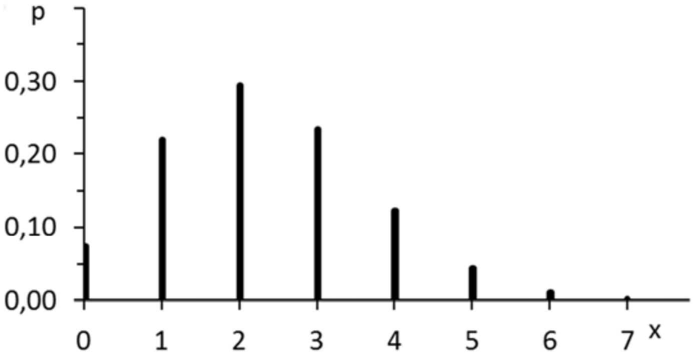


Figure 4.2. *Binomial distribution. An example; the parameter p = 0.23; n = 10 [5].*

Probability density *f*(*x*) is used for continuous quantities and shows where this quantity appears more or less often (Fig. 4.3). Mathematically, it expresses the probability that the variable *x* will lie within an infinitesimally narrow interval between x^* and $x^* + \mathrm{d}x$.

Distribution function *F*(*x*) is used for discrete as well as continuous quantities (Fig. 4.3), and expresses the probability that the random variable ***x*** attains values smaller or equal $\boldsymbol{x^*}$,

$$F(x^*) = P(x \leq x^*) \tag{4.7}$$

Both functions are related mutually as

$$f(x) = dF/dx, \quad F(x) = \int_{-\infty}^{x} f(x)\, dx, \quad \text{or} \quad F(x) = \sum_{i=1}^{n} p(x_i)\,. \tag{4.8}$$

Figure 3 shows two possibilities for depicting these functions: by histograms or by analytical expressions. **Histogram** is obtained by dividing the range of all possible values into several intervals, counting the number of values in each interval and plotting rectangles of heights proportional to these numbers. To make the results more general, the frequencies of occurrence in individual intervals are divided by the total number of all events or values. This gives ***relative frequencies*** and ***relative cumulative frequencies***, which approximately correspond to probability density and distribution function, respectively. Determination of these characteristics from empirical values will be explained in detail in Chapter 5.

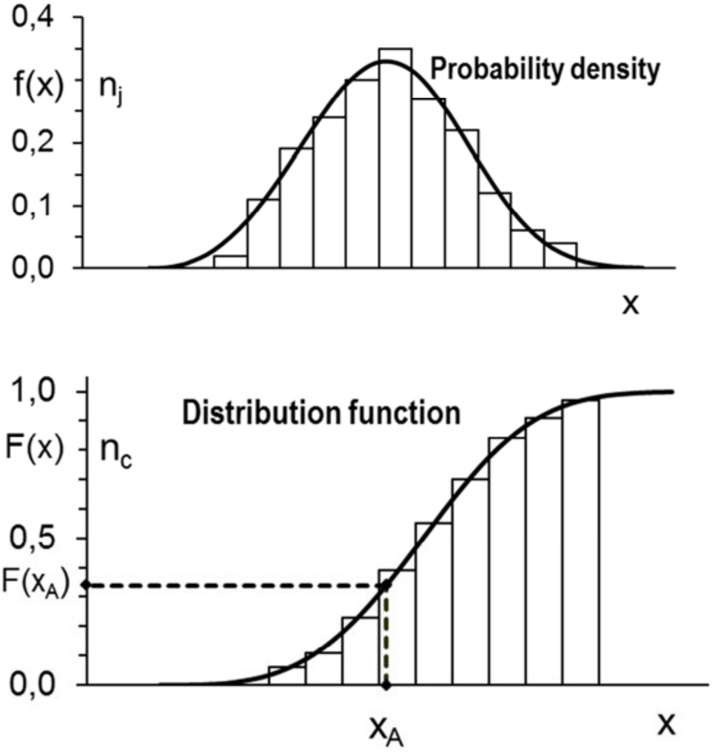


Figure 4.3. *Probability density f(x) and distribution function F(x) of a continuous quantity [5]. The histograms show relative frequency (n_j) and relative cumulative frequency ($n_{c,j}$).*

Probability of some event (e.g. snow height x lower than x_α) can be determined as the corresponding area below the curve $f(x)$ from $-\infty$ to x_α, or – directly – as the value $F(x_\alpha)$ of the distribution function.

NOTE: Probability is non-dimensional, but probability density has dimension, equal the reciprocal of the investigated quantity, such as m^{-1}, MPa^{-1} or K^{-1}. For example, the length L of a component is distributed so that $F(L \leq 2.00\ \mathrm{m}) = 0.45$ and $f(L = 3.00\ \mathrm{m}) = 0.026\ \mathrm{m}^{-1}$.

Shape of a probability distribution is quickly characterised by the following two numbers: skewness and kurtosis. **Skewness** α (coefficient of asymmetry) informs whether the distribution is symmetrical ($\alpha = 0$) or elongated towards right ($\alpha > 0$) or left ($\alpha < 0$). **Kurtosis** ε informs whether the distribution is sharper ($\varepsilon > 0$) or blunter ($\varepsilon < 0$) than normal distribution ($\varepsilon = 0$). Both quantities are shown in Figure 4 and their definitions can be found in statistical textbooks, e.g. [1 – 4].

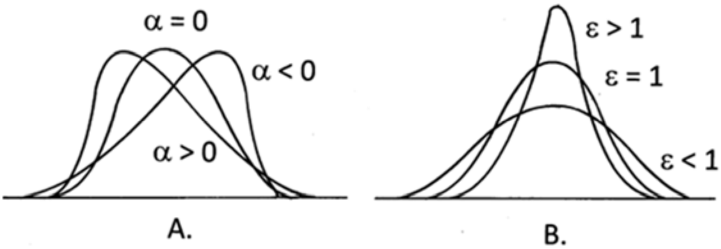


Figure 4.4. *Skewness α and kurtosis ε of probability distribution.*

Very important are also the following two quantities.

Quantile $\boldsymbol{x}_\alpha$ is such value of the random quantity $\boldsymbol{x}$, that the probability of $\boldsymbol{x}$ being smaller or equal to $\boldsymbol{x}_\alpha$ is only α,

$$P(x \leq x_\alpha) = \alpha \tag{4.9}$$

Quantiles are inverse to the values of distribution function. In Fig. 3, x_α is the α-quantile, which corresponds to the probability $F(x_\alpha)$,

$$x_\alpha = F^{-1}(\alpha) \tag{4.10}$$

Quantiles are used for the determination of the “guaranteed” or “safe” minimum value of some quantity, such as the strength or time to failure.

Critical value $\boldsymbol{x}^\beta$ is such value of the random quantity $\boldsymbol{x}$**,** that the probability of it being exceeded is only β,

$$P(x > x^\beta) = \beta \tag{4.11}$$

Note that β in this case does not denote an exponent!

Critical values are used for the determination of the maximum expectable value of some quantity, such as wind velocity or maximum height of snow in some area. They are also used for hypotheses testing, for example whether two samples come from the same population. Probability β is complementary to α, that is $\beta = 1 - \alpha$, and

$$x^{\beta} = x_{1-\alpha}\,, \quad x_{\alpha} = x^{1-\beta} \qquad (4.12)$$

More about the basic probability definitions and rules can be found in [1 – 4].

Important probability distributions

Several distributions are very important. For discontinuous quantities it is binomial and

Poisson´s distributions. The main distributions for continuous quantities are normal, log-normal, Weibull and exponential. For some purposes also uniform distribution, chi-square (χ^2) and Student´s *t*-distribution are used. The brief descriptions follow; more details can be found in comprehensive literature, such as [1 – 4].

Binomial distribution (Fig. 4.2) pertains to the probability of occurrence of ***x*** positive outcomes in ***n*** trials if this probability in each trial equals *p*. An example is the number ***x*** of faulty items in a sample of size *n*, if their proportion in the population is *p*. The probability function is

$$p(x) = \binom{n}{x} p^{x} (1-p)^{n-x} \qquad (4.13)$$

and the mean value is $\mu = np$. This distribution is discrete and has only one parameter *p*, which can be determined from the total number *m* of positive outcomes in *n* trials as *p* = *m*/*n*; *n* should be very high.

Poisson distribution is similar to a binomial distribution, but more suitable for rare events with low probabilities *p*. The probability function, giving the probability of occurrence of *x* positive outcomes in *n* trials is

$$p(x) = \frac{\lambda^{x} e^{-\lambda}}{x!} \qquad (4.14)$$

λ is the distribution parameter. (λ corresponds to the average occurrence of *x* and, in fact, to the product *np* of binomial distribution.)

Normal distribution, called also Gauss distribution, resembles symmetrical bell-shaped curve (Figures 4.3 and 4.5). It is used very often for continuous variables, especially if the variations are caused by many random influences and the variance is not too big. The probability density is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right] \tag{4.15}$$

with the mean μ and standard deviation σ as parameters. There is no closed-form expression for the distribution function $F(x)$; it must be calculated as the integral of the probability density, cf. Eq. (4.8). In practice, various approximate formulae are used for calculation of F; see, for example [6].

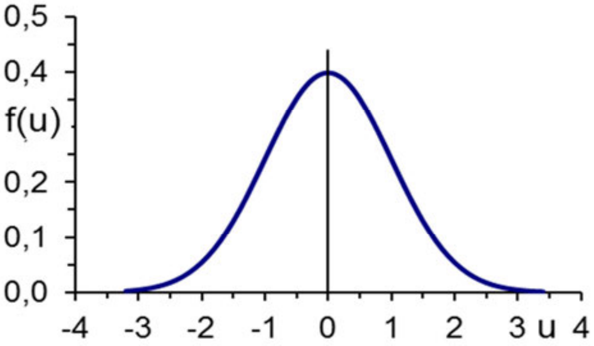


Figure 4.5. *Standard normal distribution (μ = 0, σ = 1).*

Standard normal distribution corresponds to normal distribution with parameters $\mu = 0$ and $\sigma = 1$ (Fig. 4.5). The expression for probability density is usually written as

$$f(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-u^2/2\right) \tag{4.16}$$

u is the standardised variable, which is related to the variable x of the normal distribution as

$$u = (x - \mu)/\sigma \tag{4.17}$$

It expresses the distance of x from the mean μ as the multiple of standard deviation. It is useful to remember that 68.27% of all values of normal distribution lie within the interval $\mu \pm \sigma$, 95.45% within $\mu \pm 2\sigma$, and 99.73% within $\mu \pm 3\sigma$.

Log-normal distribution is asymmetrical (elongated towards right, similar to Weibull distribution with b = 2 in Fig. 4.6) and appears if the logarithm of random variable has normal distribution.

Weibull distribution (Fig. 6) has the distribution function

$$F(x) = 1 - \exp \{ - [(x - x_0)/a]^b \} \tag{4.18}$$

with three parameters: scale parameter a, shape parameter b, and threshold parameter x_0, which corresponds to the minimum possible value of x. The probability density $f(x)$ can be obtained as the derivative of distribution function. Weibull distribution is very flexible thanks to the shape parameter b (Fig. 4.6). It is often used for approximation of strength or time to failure. It belongs to the family of **extreme values distributions** [7], and appears if failure of an object is caused by its weakest part. Determination of parameters of this very important distribution from empirical data will be explained in Chapter 5.

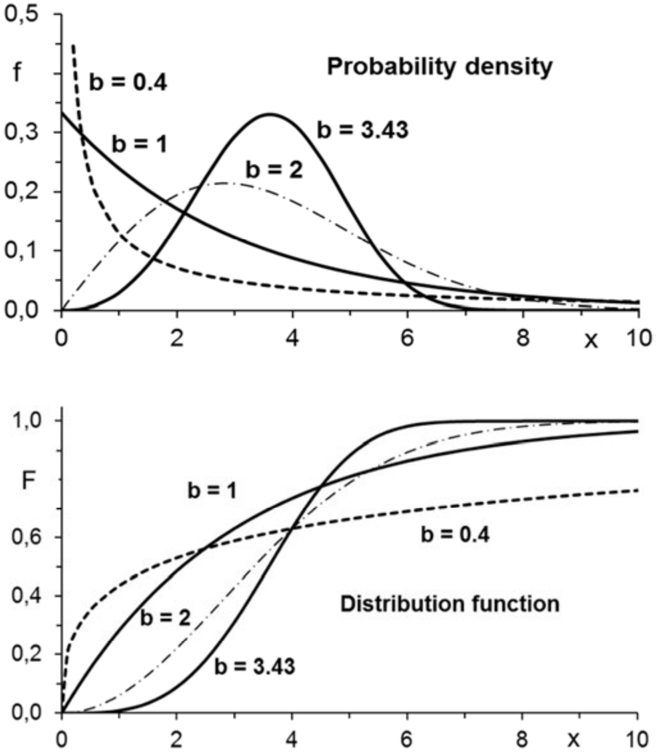


Figure 4.6. *Weibull distribution for various values of shape parameter b [5].*

In addition to flexibility, the Weibull distribution has a special advantage in the analysis of reliability. The shape parameter b in Equation (4.18) is related to the character of failures and informs generally about the period in the life of the object. The values $b < 1$ are typical of the period of early failures, while $b > 1$ pertains to the period of aging. The value $b \approx 1$ corresponds to useful life with failures from many various reasons.

Exponential distribution is a special case of Weibull distribution (4.18) for shape parameter $b = 1$ (see also Fig. 4.6) , with the distribution function

$$F(t) = 1 - \exp[-(t/T_0)] \tag{4.19}$$

It is used, for example, for the times t between failures caused by many various reasons, e.g. in complex systems consisting of many parts. This distribution has only one parameter, T_0, which corresponds to the mean μ and has the same value as the standard deviation σ. (Note: in this case of time, the symbol t was used instead of x; the difference is only formal. Moreover, the minimum possible value x_0 is often assumed 0.)

The following four distributions are important especially for the determination of confidence intervals, for statistical tests and for the Monte Carlo simulations, as it will be shown later.

Uniform distribution has constant probability density, $f = const$, in the definition interval $<a; b>$, so that it looks like a rectangle. The mean value corresponds to the average of both boundaries, $\mu = (a + b)/2$, and the variance is $\sigma^2 = (b - a)^2/12$.

Chi-square distribution (χ^2) is the distribution of the sum of n random quantities, each defined as the square of standard normal variable. An important parameter is the number of degrees of freedom, equal in this case n. For more, see [1 – 4].

***t* – distribution**, called also Student distribution, arises from a ratio of standard normal distribution and chi-square distribution. It looks similar to a standard normal distribution (Fig. 4.5), but it is lower and wider, especially for lower numbers of degrees of freedom; see [1 – 4].

***F*–distribution** corresponds to the ratio of two chi-square distributions, and it is used for comparison of two variances. This distribution depends on the number of degrees of freedom of each variable [1 - 4].

Further information can be obtained from Wikipedia or the quoted references. The values of distribution functions and quantiles of the above distributions can be found via special tables or statistical or universal programs, including Excel.

Now, two important probabilistic concepts will be explained.

Confidence interval. A consequence of random variability of many quantities is that every measurement and the following calculations give a different result depending on the specimen used. Therefore, the average $\bar{x}$ = $\Sigma x_j/n$ is usually determined from several values to obtain a more definite information. This, however, does not say how far it is from the actual mean μ. For this reason, confidence interval is often determined, which contains (with high probability) the actual value μ. For example, the α-confidence interval for the mean is

$$\bar{x} - t_{\alpha,n-1}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha,n-1}\frac{s}{\sqrt{n}} \qquad (4.20)$$

$\bar{x}$ and s are the average and standard deviation of the sample of n values and $t_{\alpha,\, n-1}$ is the α – critical value of two-sided t–distribution for n – 1 degrees of freedom. The probability that the true mean μ will lie inside the interval (4.20), is 1 – α, and α that it will lay outside it. (A practical application will be shown later.) Confidence intervals can also be determined for other quantities. For more see [1 - 4].

REMARK: Also one-sided critical values exist. Such a value (α') corresponds to the probability that μ will be either higher or lower than the pertinent critical value. α' is related to α as $\alpha' = \alpha/2$. When using statistical tables or computer tools one must be aware how the pertinent quantity was defined.

Testing of hypotheses. Often one must decide which of two procedures or products is better, or whether a true difference exists between two groups of measured quantities. Such decision can be based on the comparison of the values of a characteristic parameter, for example the average value. However, the individual values usually vary, so that also a difference can exist between the calculated parameters. If this difference is big, the decision is easy. In the opposite case one must take into account that a part of the variability of individual values is due to random reasons. For a reliable decision, statistical tests are used, which can reveal whether the differences between the characteristics of compared samples are only random, or if they reflect a real difference between both populations (e.g. types of products). These tests consist of several steps. In the first step, the so-called ***null hypothesis*** is formulated: "there is no difference between both populations". (The ***alternative hypothesis*** is "significant difference exists between the populations".) In the second step, a test criterion is calculated from statistical characteristics of both samples; the form of this criterion depends on the kind of the problem and can be found in statistical literature [1 – 4] or computer software. If the null hypothesis is valid, the test criterion has certain distribution. In the third step, the calculated value of the criterion is compared with the *critical value* of this distribution. If the calculated value is higher than the low-probability critical value, an event has happened, which was expected only with very low probability α (e.g. 5% or 1%), and we conclude that the difference is not random – the null hypothesis is rejected on the significance level α. If the calculated value of the criterion is lower than the critical value, we usually conclude that there is no substantial difference between both populations. We also say that the difference between the considered cases is not statistically significant. From this point of view it is important what probability α we consider as significant; this is a matter of our choice.

REMARK: In this test, the probability α exists that our conclusion "Both populations differ", based on the rejection of the null hypothesis, is wrong, and no actual difference between them exists. This is so-called error of the first kind. If the null hypothesis was not rejected, an opposite risk exists that, in fact, both populations differ (= error of the second kind). The probability of this wrong conclusion is β. Higher confidence in correctly rejecting the null hypothesis also means higher risk of accepting the alternative hypothesis, and usually a compromise must be found.

Tests of hypotheses are explained in detail in literature, e. g. [1 – 4], and are available in various statistical or universal computer programs. Also Excel offers several tests: for the difference between the mean values or between the variances of two populations. Applications will be shown in Chapter 8.

Order statistics

A frequent problem in experimental analysis is that we have a series of experimental values (e.g. strength or time to failure) and want to find the parameters of the probability distribution, or a certain quantile or the value of distribution function. In some cases it is simple; for example the parameters of a normal distribution are the mean and standard deviation. Sometimes, it is not so straightforward, e.g. with Weibull three-parameter distribution, or if the histogram of the measured values has a more complex shape. Fortunately, in such cases it is possible to assign the values of the distribution function to the measured values in the following simple way. First, the measured values are rank-ordered from the minimal (j = 1) to the maximal ($j = n$); j is the rank number and n is the total number of measured values. The corresponding values of the distribution function are calculated as

$$F_j = j / (n + 1) \tag{4.21}$$

The explanation of formula (4.21) is simple. If we have, say, 100 values of the time to failure t, and order them from the minimal to maximal, then the probability F that t will be smaller or equal to the lowest of 100 values, t_1, is approximately 1:100. The probability of $t \leq t_2$ is 2/100, etc.; generally $F_j = j/n$. In Equation (4.21), 1 was added to the denominator because of mathematical correctness. The probability F that t will be smaller or equal t_n must be smaller than 1, because if more measurements would be done, also values higher than the above value t_n can appear.

REMARK: Also other formulae exist for the calculation of empirical F_j values, for example $F_j = (j - ½)/n$. However, none can be recommended unequivocally, especially when considering the fact that bigger errors in the determination of distribution parameters can arise due to small amount of randomly varying empirical data included into the sample, than due to the formula used for F_j.

Nonparametric methods

Statistical tests and procedures usually assume a certain probability distribution of the investigated quantity, and work with its parameters. Nevertheless, also nonparametric or distribution-free methods exist [8], which do not require any assumption on the distribution nor the knowledge of its parameters. Distribution-free methods can be used also in cases where any information on the distribution is missing. On the other hand, they usually need a larger size n of the sample to achieve the same power of the information or test.

The most important nonparametric methods applicable in experimental research are: (1) determination of quantiles, (2) tests of goodness-of-fit, used to check whether the sample has a certain distribution, (3) tests to check whether two samples are drawn from the same population, and (4) tests of correlation of two variables. They usually work with rank-ordered values. In the next paragraph the first method will be explained, which is used very often. The other methods will be described later in this book.

Quantiles. In the previous paragraph, the assignment of the F_j values of a distribution function to the individual rank-ordered values y_j was explained. The finding of α–quantile of y is the opposite problem: it is such y_α value of the data series, which corresponds to the value α of distribution function F. If the exact value $F = \alpha$ is not available, it can be found from the neighbouring lower and higher values of F by interpolation. The quantile is found also by interpolation from the neighbouring values of y.

Acknowledgment. Parts of this chapter were previously published in Chapter 2 of Ref. [5].

References to Chapter 4

1. Freund, J. E.: Modern elementary statistics. Prentice-Hall, Inc., Englewood Cliffs, New Jersey,1981(6th edition). 561 p.
2. Freund, J. E., and Perles, B. E.: Modern elementary statistics. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 2006 (12th edition). 576 p.
3. Suhir, E.: Applied Probability for Engineers and Scientists. McGraw-Hill, New York, 1997. 593 p.
4. Montgomery, D. C., and Runger, G. C.: Applied Statistics and Probability for Engineers. John Wiley, New York, 2006 (4th edition). 784 p.
5. Menčík, J.: Concise reliability for engineers. InTech, Rijeka, 2016, 204 p. *Open Access publication,* ISBN 978-953-51-2278-4. *Available* at: http://www.intechopen.com/books/concise-reliability-for-engineers,
6. Abramowitz, M. and Stegun, I.: Handbook of mathematical functions. National Bureau of Standards, Washington. 1972 (tenth printing). 1046 p.
7. Gumbel. J. E.: Statistics of Extremes. Columbia University Press, New York, 1958. 375 p.
8. Conover, W.J.: Practical nonparametric statistics. 3rd edition, 1999. Wiley, 584 p.

5. Determination of Characteristics of Investigated Quantities

Measured values often vary for random reasons and therefore are usually described by some characteristic values or by probability distribution. This chapter explains construction of histograms and finding of parameters of probability distribution. Attention is paid to flexible Weibull distribution, illustrated on a practical example.

Characteristic values

The most important parameters are the mean μ and standard deviation σ. When empirical data are evaluated, they are replaced by the sample average $\bar{x}$ and sample standard deviation *s*, defined by formulae (4.2) and (4.4) in Chapter 4. Additional characteristics are the coefficient of asymmetry (skewness) and kurtosis, also mentioned there. The characteristic values are the only source of information on the position of the random quantity on the numerical axis and on its dispersion if the amount of empirical data is very small, less than about 15. Universal programs can, after a single command, calculate all characteristics and print a table with them. In Excel, for example, it is sufficient to give the commands Data analysis and Descriptive statistics.

REMARK. Programs Data analysis and Solver (for solution of equations and search of extreme values) are installed in every Excel, but not always accessible. If they are not visible at the upper bar of the submenu Data, they must be activated as follows. After pressing the button File, we choose Options and Add-ins. Then, we mark Analysis ToolPak and press the command Go. In the following small menu we mark “Analysis ToolPak” and also “Solver Add-In” and press OK. That is all.

Histogram

The first idea about the probability distribution of the investigated quantity can be obtained from a **histogram** (Fig. 4.3 in Chapter 4 and Fig. 5.1 on the next page).

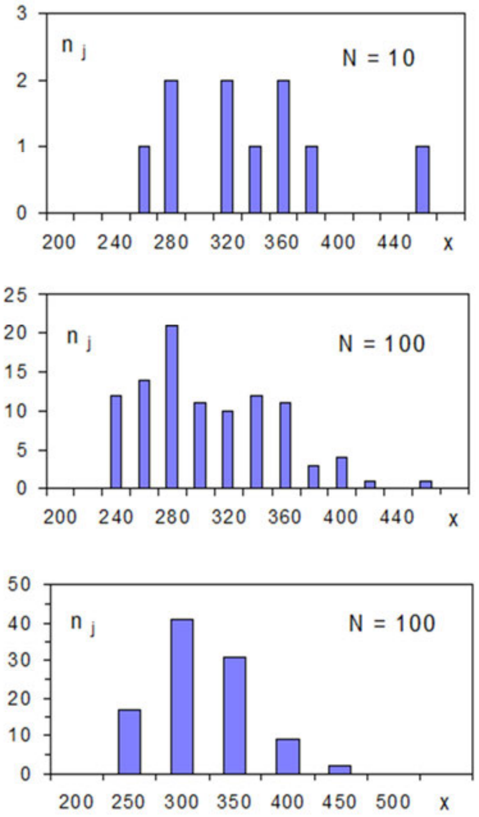


Figure 5.1. *Histograms of a group of experimental data for various N. Examples of poor and good appearance.*

Histogram is constructed from all recorded values by dividing the range of all possible values into several intervals of the same width, counting the number of values (= frequency) in the individual intervals and plotting rectangles of heights proportional to these numbers [1, 2]. Histograms are created easily by universal programs such as Matlab, Mathcad, SPSS, Statistica, or Excel. With the last named it is ensured by the command *Histogram* from the menu *Data Analysis*; in this case also the number of intervals (bins) and their boundaries must be known in advance. Unfortunately, there is no universal formula for the determination of the number *m* of bins. In literature, two following empirical formulae are given most often:

$$m = \mathrm{INT}(2 \ln N) \ , \quad m = \mathrm{INT}(2 \sqrt{N}) \qquad (5.1)$$

N is the total number of all values and INT means the integer part of the expression. However, these formulae are suitable only for several tens of values. As universal computer programs create histograms instantly, it can be recommended (especially if the histogram, constructed from a low number of values, looks strange) to plot several variants of the histogram, with various numbers of bins and various coordinates of their borders, and to choose the best looking one, of a simple shape. Examples of histograms with appealing and poor appearance are shown in Figure 5.1. If a more complicated distribution can be expected (e.g. with two "hills"), at least several hundred values are necessary.

In addition to the histograms that give frequencies n_i in the individual bins (i), it is also possible to construct histograms with **cumulative frequencies**: each bin contains the number of all values from the left-end bin to the investigated j-th one:

$$n_{j,cum} = \sum_{i=1}^{j} n_i \tag{5.2}$$

If the numbers of values in the individual bins are divided by the total number of values *N*, **relative frequencies** $\boldsymbol{f_j}$ and **relative cumulative frequencies** $\boldsymbol{F_{j,cum}}$ are obtained. These two quantities correspond approximately to the probability density *f* and distribution function *F*.

Probability distribution

It is advantageous if empirical data can be fitted by some of the standard probability distributions. Simple situation is with normal, lognormal, or exponential distribution. Normal (Gauss) distribution (Fig. 4.3 or 4.5 in Chapter 4) is described fully by the mean and standard deviation. Its use is therefore very easy. In practice, the parameters μ and σ are replaced by the sample average and standard deviation, defined in Chapter 3. Lognormal distribution works in similar manner with the logarithms of the measured values. Exponential distribution (Fig. 4.6 in Chapter 4, case b = 1) has only one parameter, the mean; the standard deviation has the same value as the mean, so that it is sufficient to calculate the average of the measured values. (The calculated estimates can slightly differ.)

Great flexibility in fitting various shapes of continuous probability distributions is offered by Weibull distribution (Fig. 4.6 in Chapter 4).

Weibull distribution

General form of its distribution function (Fig. 5.2 here and Fig. 4.6 in Chapter 4) is

$$F(t) = 1 - \exp\{-[(t - t_0)/a]^b\} \qquad (5.3)$$

with parameters a, b, and t_0. The ***scale parameter*** a is related to the values of t and ensures that the distribution is independent of the units of t (e.g. minutes or hours).

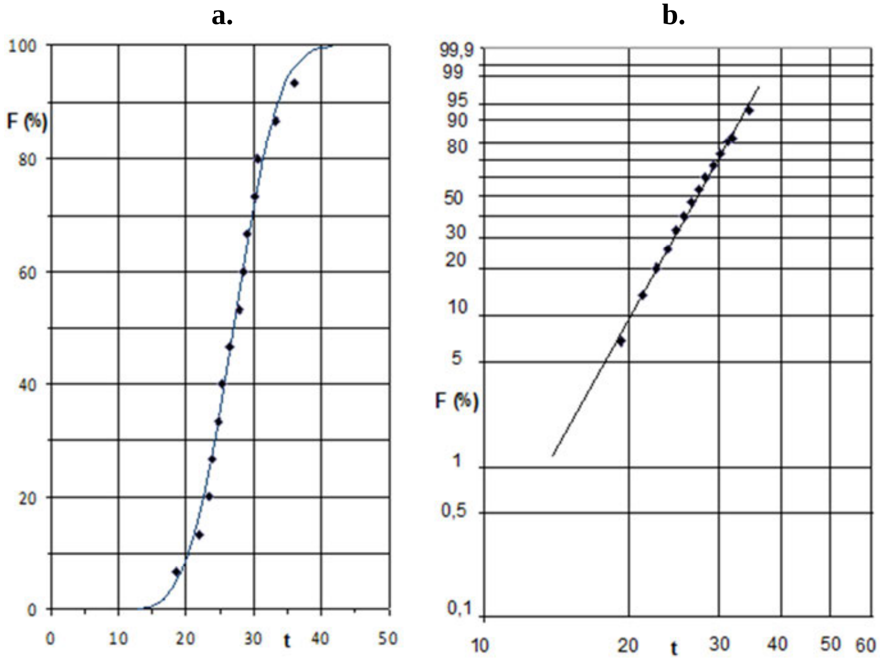


Figure 5.2. *Weibull distribution function F(t): (a) original coordinate system, (b) Weibull probabilistic paper with transformed coordinates [3].*

The constant b is the ***shape parameter***. Depending on its value, Weibull function can approximate various, even very different shapes (Fig. 4.6 in Chapter 4).

Weibull distribution is suitable for the characterisation of time to failure or strength of brittle materials and became popular in reliability assessment. However, it can be used in many other cases as well. The constant t_0 is the ***threshold value*** that corresponds to the minimum possible value and characterises the position of the distribution on the t-axis. (t is the usual symbol for time; for other quantities, other symbols may be used.) In this section, two methods for the determination of the parameters will be explained. Both are based on the minimisation of the distances between the data points and the distribution function. The first method is based on fitting the transformed data by a straight line; the second does it directly via an optimisation program. Their applications will be explained here.

Two-parameter Weibull distribution

The strength or time to failure cannot attain negative values, and the threshold parameter is thus often assumed zero, $t_0 = 0$. The distribution function (5.3) has only two parameters:

$$F(t) = 1 - \exp[\, -(t/a)^b] \qquad (5.4)$$

Parameters a and b can be found easily, as the transformed data may be fitted by a straight line [3]. Double logarithmic transformation and rearrangement change Equation (5.4) to

$$\ln t = \ln a + (1/b) \ln\{\ln[1/(1 - F)]\} \qquad (5.5)$$

This corresponds to the equation of a straight line (Fig. 5.2b)

$$Y = A + BX \qquad (5.6)$$

$$Y = \ln t, X = \ln\{\ln[1/(1 - F)]\}, \quad A = \ln a, B = 1/b \qquad (5.7)$$

The regression constants A, B can be obtained by fitting the empirical data Y, X by a straight line. In the past, the measured values of t and F (see later) were plotted on a special diagram, called Weibull paper (Fig. 5.2b), and fitted by a straight line using a ruler and a guesstimate. In manufacturing it is still sometimes used for the determination of distribution parameters from the operation data. Today, however, many universal computer programs enable easy fitting of curves. For example, Excel, has the command *Insert Trendline*; then only the chart $Y(X)$ is needed. This graph is constructed from the measured (and transformed) values t_j and the corresponding values F_j of the empirical distribution function. The individual values $Y_\mathrm{j} = \ln t_\mathrm{j}$ are obtained by rank ordering of all n transformed values (e.g. times to failure) from the minimal value (j = 1) to maximal ($j = n$). The corresponding values of distribution function are calculated (see Chapter 4) as

$$F_\mathrm{j} = j \,/\, (n + 1) \qquad (5.8)$$

and then transformed to X_j values via Equation (5.7). A plot of the empirical data in the coordinate system $X = \ln\{\ln[1/(1 - F)]\}$, $Y = \ln t$, enables a good visual check.

In the ideal case, if Equation (4) is valid, the data lie along a straight line. The regression constants A and B are then obtained by right-part mouse clicking on any point of the data series, then marking ***Insert Trendline*** in the menu and selecting ***Linear regression***. We must also mark ***Show the equation*** and ***Coefficient of determination R^2***. A straight line and Equation (5.6) with the values of both constants A, B appear in the chart. The constant R^2 characterises the quality of the fit; the closer it is to 1, the better. (Explanation is given in Chapter 7.) Then, the constants in the original distribution function (5.4) are obtained from A and B by inverse transformations:

$$b = 1/B, \ \ a = \exp(A) \qquad (5.9)$$

Three-parameter Weibull distribution

Two-parameter distribution is not always suitable. Sometimes, the transformed data do not lie on a straight line, or it is obvious that the distribution should have a threshold value t_0 significantly higher than zero. In such case, a three-parameter function (5.3) is better.

The parameters in this distribution can be found by the procedure for a two-parameter function if t in Equation (5.4) is replaced by the expression $t - t_0$; the constant t_0 must be defined in advance. For various t_0 values, the shape of empirical distribution varies. The best t_0 value is such, for which the transformed data best resemble a straight line. Often, several trials are necessary. Fortunately, a straightforward procedure exists [3], described further.

Direct determination of parameters

The constants a, b, and t_0 can be obtained in a simple way without any transformation. The solution of Equation (5.3) for t gives the formula for quantiles:

$$t = t_0 + a\{\ln[1/(1 - F)]^{1/b}\} \qquad (5.10)$$

We shall now look for such constants a, b, and t_0, which will minimize the sum of squared differences between the measured and calculated values of t,

$$\Sigma(t_{\mathrm{j,meas}} - t_{\mathrm{j,calc}})^2 = \min ! \qquad (5.11)$$

This is the principle of the so-called **least-squares method**. If a suitable solver is available for such minimization (one is present also in Excel), it is then sufficient to prepare one series of measured data, $t_{\mathrm{j,meas}}$, and another series of the values $t_{\mathrm{j,calc}}$, calculated via Equation (5.10) for the same values of F_j using the parameters *a*, *b*, and t_0. Solver's command to minimize the expression (5.11) by changing *a*, *b*, and t_0 will do the job. No transformation is necessary.

NOTE. The variable *F* is considered independent here, because its values are deterministic, following from the number of values. The variable *t* (e.g. the time to failure or strength) exhibits random variability, and is thus considered as dependent variable.

Example 1.

Strength of an alloy was measured on ten specimens, with the results: S_{meas} = 250 – 201 – 232 – 281 – 297 – 211 – 276 – 302 – 315 – 265 MPa. Find the parameters of three-parameter Weibull probability distribution and determine the "guaranteed" 1% strength!

Solution. All values are given in Table 1 below. The measured strengths were rank-ordered from minimum to maximum ($S_{j,meas}$), and the corresponding values of distribution function were calculated as $F_j = j/(N + 1)$. Then, the strengths $S_{j,calc}$ were calculated for the same values F_j via Equation (5.10) for the constants *a*, *b*, S_0, defined in advance. Also the sum of squared differences (SSD) between the measured and computed strengths was calculated. [The Excel function for this expression is *SUMXMY2*, which means: sum(*x* minus *y*)^2; now $x = S_{j,meas}$ and $y = S_{j,calc}$.] The individual values are plotted in Figure 5.3: the rhombs represent the measured values, while the crosses correspond to the results of strength calculations for the (arbitrarily) chosen initial values *a* = 250 MPa, *b* = 2 and S_0 = 0 MPa. One can see that these crosses do not coincide with the measured strengths. The application of Solver (minimisation of the content of the cell containing *SUMXMY2* by changing the values *a*, *b*, and S_0) has given the following values: *a* = 280.6 MPa, *b* = 6.659 and S_0 = 0 MPa. These constants fit the measured data very well; see the solid thin curve extrapolated to the lower and higher probabilities. The guaranteed 1%-strength, $S_{0.01}$, calculated via Equation (5.10) for *F* = 0.01, is 140.6 MPa. (The reader is encouraged to solve this example for gaining practice.)

In this example it was necessary to limit the threshold strength as $S_0 \geq 0$, because the first trial without any limitation has given negative value of strength S_0, which is impossible. With this limitation Solver has immediately "recommended" the threshold value S_0 = 0, so that the probability distribution has – in fact – only two parameters *a*, *b*. In some cases there can be rather big difference between the low-

Table 1.

j	$S_{j,meas}$	F_j	$S_{j,calc}$
1	201.0	0.0909	197.2
2	211.0	0.1818	216.2
3	233.0	0.2727	232.9
4	250.0	0.3636	246.5
5	265.0	0.4545	258.4
6	276.0	0.5455	269.6
7	281.0	0.6364	280.7
8	297.0	0.7273	292.3
9	302.0	0.8182	305.4
10	315.0	0.9091	322.7

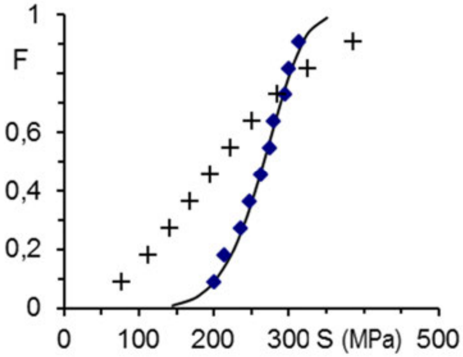


Figure 5.3. *Measured and computed strengths (S) and distribution function F from Ex. 1.*

probability values predicted by two or three parameter Weibull function. The two-parameter distribution with the assumed threshold value S_0 = 0 would give lower allowable stress, which is safer. On the other hand, the size of cross-section of such component must be larger, and therefore more expensive. Sometimes, a compromise must be found between safety and economy.

Estimation of distribution parameters from censored data

In some cases the amount of experimental data is limited and only part of the results is known. For example, the time to fatigue failure or to the failure of complex objects varies, and when a group of such items is tested in order to obtain the characteristics of the lifetime distribution, these times could be impracticably long for some of the tested pieces. Therefore, the lifetime tests are sometimes terminated after some time t_{test} or after failure of a certain fraction of tested parts. We know exactly the times to failure of the failed parts, and know also that the lifetime of the remaining components would be longer (but not know how long they will be). Another case is if the measured quantity has some values beyond the range of the used measuring device; in this case we say that the data are censored. The situation is depicted in Figure 5.4. If the kind of probability distribution is known, its parameters can be estimated from the part of data for which the times to failure are known, if each of these rank-ordered values (t_j) is assigned the corresponding value of distribution function F_j = $j/(N+1)$. These issues are very important in reliability testing, and various test schemes and procedures for the processing of results have been developed; cf. [4].

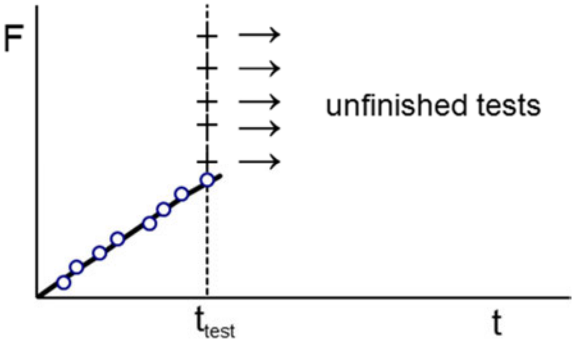


Figure 5.4. *Censored data from lifetime tests (a schematic).*

Q - Q plot

A simple graphical tool for comparing two probability distributions is a **quantile-quantile plot**, or Q–Q plot [5]. In this plot, the corresponding quantiles are plotted against each other. The coordinate x_j of a point P_j in this plot corresponds to j-quantile of one distribution and the coordinate y_j corresponds to the same quantile of the other distribution. If the two compared distributions are similar, the points in the Q–Q diagram lie approximately on the line $y = x$ (Fig. 5.5). The Q–Q plot informs whether the location, shape and skewness of the compared distributions are similar or different. These plots can be used to compare collections of two sets of data, or to compare an empirical distribution with a theoretical one.

A quantile-quantile plot is created as follows. The compared quantities (the first is

x and the second is y) are ordered from minimum to maximum, the corresponding values of distribution function F are calculated via Equation (5.7), and the couples of values of quantiles x_j, y_j for the same value F_j are plotted in coordinates x, y (Fig. 5.5). This is easiest if both samples have the same number of values. Otherwise, the quantiles for the same probabilities must be recalculated by interpolation from the neighboring points.

As an example, Figure 5 shows the Q-Q plot for 20 theoretical values of normal distribution (with the average μ = 5.0 and standard deviation σ = 1.0) compared with 20 “empirical” values, generated in this example for the same parameters using the Monte Carlo method. One can see that the individual points lie (approximately) on a straight line.

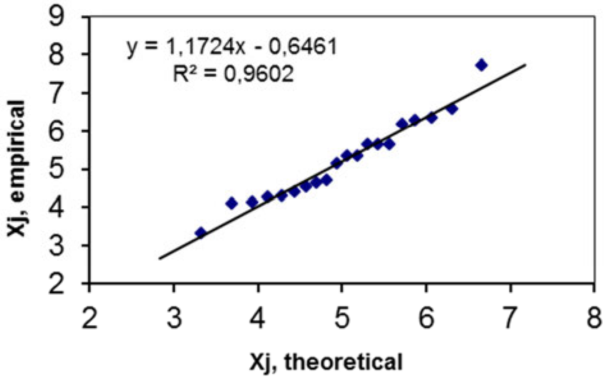


Figure 5.5. *Q – Q plot for normal distribution (parameters: μ = 5, σ = 1, n = 20). Horizontal axis – theoretical values, vertical axis – empirical values.*

Acknowledgment. A part of this chapter was published in Chapter 11 of Ref. [3].

References to Chapter 5

1. Pechoč, V.: Evaluation of measurement and computing methods in chemical engineering. (In Czech: Vyhodnocování měření a početní metody v chemickém inženýrství.) SNTL, Praha, 1981.
2. Felix, M., and Bláha, K.: Statistical methods in chemical industry. (In Czech: Matematickostatistické metody v chemickém průmyslu.) SNTL, Praha, 1962. 336 p.
3. Menčík, J.: Concise reliability for engineers. (Chapter 11). InTech, Rijeka, 2016, *Open Access, available* at: http://www.intechopen.com/books/concise-reliability-for-engineers, 204 p. ISBN 978-953-51-2278-4.
4. Bednařík, J. et al.: Reliability techniques in electronic practice. (in Czech: Technika spolehlivosti v elektronické praxi). Praha, SNTL, 1990. 336 p.
5. Kupka, K.: Statistical quality control (in Czech: Statistické řízení jakosti). Trilobyte, Pardubice, 1997, 2001. 191 p.

6. Relationships of Two or More Quantities

This chapter explains various quantities used for characterisation of relationships between two or more variables, such as covariance, correlation and coefficient of determination. It also shows the use of these quantities for the determination of constants in a regression function and evaluation of the quality of the fit. Autocorrelation shows whether the values in a series of data are correlated among themselves. Finally, obtaining of information by data mining is explained.

Covariance and correlation

One task of experimental research is to reveal whether a relationship exists between two or more quantities, how strong this relationship is, and preferably to describe it by a suitable mathematical expression. The strength of such relationships may range from non-existing over less or more strong to deterministic. Its strength can be characterised by the coefficient of covariance and correlation coefficient. For two variables, *x*, *y*, **covariance** cov(*x*,*y*) is defined as

$$s_{xy} = \frac{\sum (x_j - \bar{x})(y_j - \bar{y})}{n-1} \tag{6.1}$$

$\bar{x}$ and $\bar{y}$ are the average values of both quantities and *n* is the number of pairs x_j, y_j.

Covariance can be positive, if both quantities increase together, and negative, if one quantity grows while the other decreases.

A drawback is that covariance coefficient can attain values from –∞ to +∞, depending also on the values of *x*, *y*. A better measure of the relationship is the **coefficient of correlation** r_{xy}, defined as the covariance divided by standard deviations of both variables:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{6.2}$$

In fact, coefficient of correlation is covariance standardised with respect to the dispersion of both quantities. Its values vary between 0 (no correlation) and 1 (deterministic or functional relationship) for positive correlation. With negative correlation, r_{xy} varies between 0 and –1 (Fig. 6.1).

NOTE: Expressing of relationships by regression functions is given in Chapter 7.

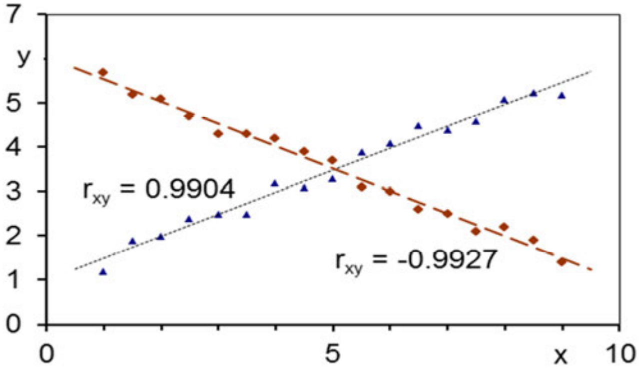


Figure 6.1. *Correlation positive (triangles) and negative (rhombs).*

An analogous nonparametric characteristic is **Spearman´s rank correlation coefficient**:

$$r_S = 1 - \frac{6\sum d_j}{n(n^2 - 1)} \qquad (6.3)$$

n is the number of pairs of values x_j, y_j of the quantities *x*, *y*. Each x_j value is assigned the rank number p_j, and *y* values are assigned the rank numbers q_j (j = 1 corresponds to the smallest values). Then, the differences of rank numbers for the individual pairs x_j, y_j are created as $d_j = p_j – q_j$ and used in Equation (6.3).

Relatively high values of correlation coefficient, *r* = 0.8 and more, indicate functional relationship. Nevertheless, in a case of doubt, statistical test of significance is recommended (see Chapter 8).

Caution: High degree of correlation does not necessarily imply causation. Another factor can exist, which influences both quantities (*x*, *y*) in a similar way. Well known is the following humorous example. An investigation, made in several villages, has shown that there is a high correlation between the number of born babies and the number of storks at the villages. Does it mean that babies are brought by storks? No; the explanation is related to the size of the villages: in larger villages more babies are born, but usually larger villages have more ponds, and thus also more storks.

A very useful quantity is the **coefficient of determination r^2**. It can be used for the characterisation of the quality of linear as well as nonlinear regression functions. Figure 6.2 shows a regression line, one value of *y* measured for certain value *x*, the corresponding value *y*(*x*) on the regression line, and the average values $\bar{x}$, $\bar{y}$ (or x_{ave}, y_{ave}) of a group of x_{j} and y_{j}. The distance of y_{j} from the average $\bar{y}$ is expressed as

$$y_{\mathrm{j}} - \bar{y} = (y_{\mathrm{j}} - y_{\mathrm{j,reg}}) + (y_{\mathrm{j,reg}} - \bar{y}), \text{ or } \quad \Delta_{\mathrm{tot}} = \Delta_{\mathrm{res}} + \Delta_{\mathrm{reg}} \qquad (6.4)$$

$y_{\mathrm{j,reg}}$ is the corresponding value on the regression line; Δ_{tot} means the total difference, Δ_{reg} means the difference of the j-th value on the regression line and the mean, and Δ_{res} is the residual difference, i.e. the distance of the j-th measured value and the corresponding value on the regression line. It is possible to prove that also the following relationship holds:

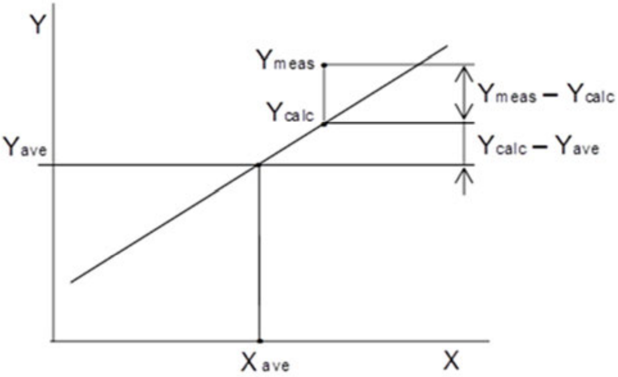


Figure 6.2. *Coefficient of determination r^2; residual component (y_{meas} − y_{calc}) and regression one (y_{calc} – y_{ave}) of the total difference y_{meas} – y_{ave}. Subscripts: meas - measured, calc - calculated, ave - average.*

$$\sum_j (y_j - \bar{y})^2 = \sum_j (y_j - y_{j,reg})^2 + \sum_j (y_{j,reg} - \bar{y})^2 \qquad (6.5)$$

The summation is done for all *n* values. Equation (6.5) can be rewritten as

$$SS_{\mathrm{tot}} = SS_{\mathrm{reg}} + SS_{\mathrm{res}} \qquad (6.6)$$

SS_{tot} is the sum of squared total differences, SS_{reg} is the sum of squared distances between the points on the regression curve and the total average $\bar{y}$, and SS_{res} is the sum of squared distances of the individual measured points from the corresponding points on the regression curve (i.e. the residual distances).

Equation (6.6) can also be rewritten as

$$s_{tot}^2 = s_{reg}^2 + s_{res}^2 \qquad (6.7)$$

This expression was obtained by dividing Eq.(6.6) by $n - 1$. *SS* means sum of squared differences and s^2 means variance; the subscripts have the same meaning as in the previous case; for example *res* corresponds to the residual variance of the individual measured values around the regression function. Coefficient of determination is defined as

$$r^2 = SS_{reg}/SS_{tot} = s_{reg}^2/s_{tot}^2 \qquad (6.8)$$

It expresses what fraction of the total variance is caused by the regression function. For deterministic relationship, $r^2 = 1$. Equation (6.8) can be rewritten to the form

$$r^2 = (SS_{tot} - SS_{res})/SS_{tot} = 1 - (SS_{res}/SS_{tot}), \text{ or } r^2 = 1 - (s_{res}^2/s_{tot}^2) \qquad (6.9)$$

REMARK: Coefficient of determination in Equation (6.8) or (6.9) can be expressed by means of the sums of squared differences or by means of standard deviations, because $s^2 = SS/(n - 1)$.

Coefficient of determination in Eq. (6.8) was derived for linear relationship $y(x)$. However, if it is calculated via Eq. (6.9) by means of residual variance, it can also be used for the characterisation of quality of nonlinear regression functions. From this it follows the importance of residual variance for the determination of parameters in regression functions (see the next chapter).

When studying a relationship between two quantities, one should always make a plot of the measured values and only then evaluate the strength of the relationship. Figure 6.3 shows a group of values, which obviously indicates a nonlinear relationship. If we would – without knowing this fact – determine the coefficient of correlation for a linear relationship, we would obtain a very low value, informing that no linear correlation exists ($r = 0.286$; $r^2 = 0.0616$). In contrast, the coefficient of determination for a quadratic regression function is $r^2 = 0.984$ and $r = 0.992$, which means a very good fit.

Revelation of correlations is the first step in the study of relationships between the investigated quantities. Generally, correlations can exist between two or more

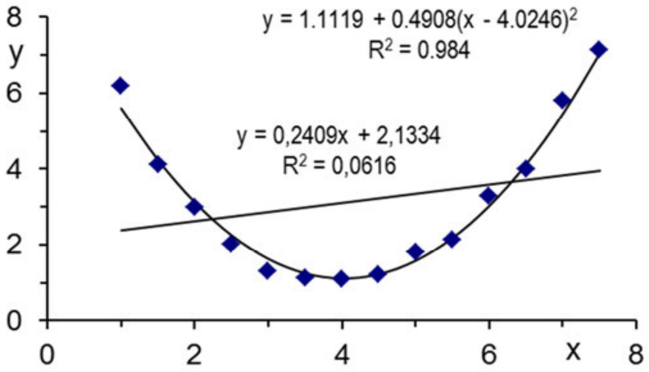


Figure 6.3. *Two approximations of the measured values – a good fit (quadratic function) and a poor fit (linear function). R^2 = coefficient of determination.*

quantities, so-called **multiple correlations**. The correlations in experimental data are usually found by means of a suitable statistical program. In Excel, for example, the command *CORREL*, applied on a group of paired values x_j and y_j, gives the value of (linear) correlation coefficient, r_{xy}. In this case, however, it is more efficient to make a chart of the $y(x)$ data and plot there a regression function using the command *Insert Trendline*, as this gives a very instructive picture. It is also useful to demand (from the menu) that the coefficient of determination r^2 is shown, which characterises the strength of the relationship.

Multiple correlations will be illustrated here on thermal treatment of steel.

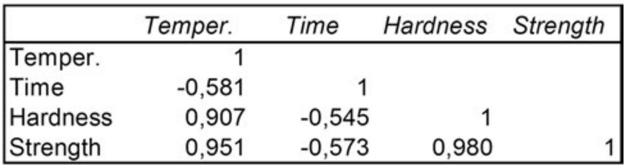
Example.

It was investigated how the hardness and strength of quenched steel are influenced by the temperature of the treatment and the dwell time under high temperature, and whether they are correlated. Eight samples were treated under various conditions, as shown in the upper part of Table 1. The table of multiple correlations below was created in Excel. The keys *Data Analysis* and *Correlation* were used, then the array containing the input data was written into the pertinent cell in the Correlation menu and a cell was marked in the worksheet for positioning the correlation table. After pressing OK, the correlation table appears.

The correlation table indicates clearly which quantities are strongly correlated, and which not. For example, the coefficient of correlation between strength and hardness is very high, 0.980, while that between the dwell and hardness (–0.545) is

TABLE 1. *Input data and the table of multiple correlations (below).*

Sample	Temper.	Time	Hardness	Strength
A	800	10,3	475	212
B	840	9,8	510	246
C	920	9,6	540	285
D	820	10	490	230
E	870	9,7	520	255
F	850	10,5	530	260
G	900	9,4	550	295
H	860	9,1	532	265



	Temper.	*Time*	*Hardness*	*Strength*
Temper.	1			
Time	-0,581	1		
Hardness	0,907	-0,545	1	
Strength	0,951	-0,573	0,980	1

low. The omission of the insignificant quantities means simplification of formulae and later calculations. High correlation of two quantities allows the use of any of them, which can sometimes simplify the work. For example, the determination of tensile strength is more demanding than the measurement of hardness. If the tests with simultaneous measurement of strength and hardness reveal high correlation, as above, it is possible to measure only the hardness and recalculate the strength from it using a suitable transformation formula, as shown in Figure 6.4.

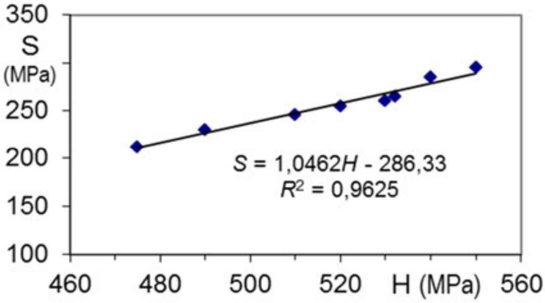


Figure 6.4. *Strength S as a function of hardness H (both in MPa):* $S = 1.046\ H - 286.33$.

Autocorrelation.

Until now, we considered if two quantities are mutually correlated. However, it is also possible to investigate, if the values in one series of data are correlated among themselves. An example is the daily average temperatures. The difference between two temperatures is smaller if it corresponds to two subsequent days than if the interval between them is several months. In this way it is possible to characterise a series of data by autocorrelation. The autocorrelation coefficient is obtained similarly as the correlation coefficient of two variables *x* and *y* (in Excel via the command CORREL). The only difference is that, instead of the quantity *y*, the values of *x* are used again, but shifted by 1, 2,… or *n* positions. This new series is denoted *y´* and we speak about the autocorrelation of the first, second… or *n*-th order. For example, the table below corresponds to the first order autocorrelation.

y :	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	….
$y´$:		y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	….

Autocorrelation can be positive or negative. Both cases are shown in Figure 6.5.

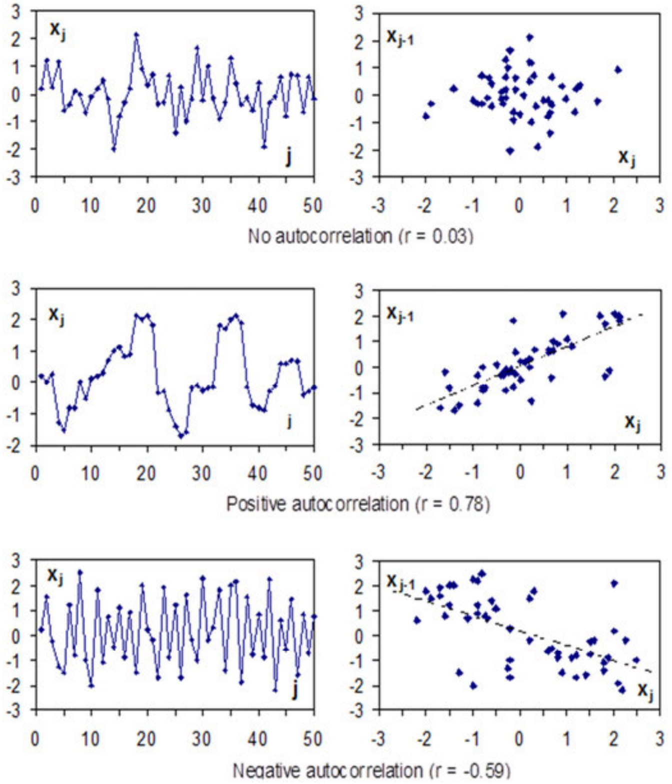


Figure 6.5. *Autocorrelation: a) none (r = 0.03), b) positive (r = 0.78), c) negative (r = –0.59). Diagrams at left: time series; diagrams at right: autocorrelation;* $x_{j-1} = f(x_j)$.

REMARK: Autocorrelation is used especially in the analysis of dynamic processes, time series and signal processing.

More information on covariance and correlation can be found in books [1 – 5].

Data mining. In some branches, such as chemistry, biology, medicine or astronomy, but also in banking and business, big amounts of data exist. Today, powerful computers are able to process them. This has led to the development of a new branch called data mining. In contrast to traditional data analysis, where first a certain hypothesis is formulated, and then it is proved or rejected using data obtained from experiments or observation, data mining goes in the opposite way, It searches through the vast amount of existing data and tries to find some specific patterns in them, which may carry hidden and potentially useful information on some relations yet unknown. As the amount of analyzed data is huge (TB), suitable software is necessary. There are specialized programs for this purpose, such as STATISTICA Data Miner, SAS Enterprise Miner and SPSS Clementine.

Examples of non-commercial software are Weka and Orange. More information on data mining can be found in [6].

References to Chapter 6

1. Freund, J. E.: Modern elementary statistics. Prentice-Hall, Inc., Englewood Cliffs, New Jersey,1981(6th edition). 561 p.
2. Freund, J. E., and Perles, B. E.: Modern elementary statistics. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 2006 (12th edition). 576 p.
3. Suhir, E.: Applied Probability for Engineers and Scientists. McGraw-Hill, New York, 1997. 593 p.
4. Montgomery, D. C., and Runger, G. C.: Applied Statistics and Probability for Engineers. John Wiley, New York, 2006 (4th edition). 784 p.
5. Kupka, K.: Statistical quality control (in Czech: Statistické řízení jakosti). Trilobyte, Pardubice, 1997. 191 p.
6. Piatetsky-Shapiro, G., and Parker, G. "Lesson: Data Mining, and Knowledge Discovery: An Introduction". Introduction to Data Mining. KD Nuggets. 2011, Retrieved 30 August 2012. http://www.kdnuggets.com/data_mining_course/x1-intro-to-data-mining-notes.html

7. Fitting of Empirical Data by Regression Functions

This chapter shows how relationships among various quantities can be described by suitable regression functions. Typical functions are shown and the determination of regression constants is explained. As the input data exhibit variability for random reasons, the regression function does not give accurate values. The reliability of predictions can be increased if confidence band is created for the regression function or its values.

It is useful to describe the empirical data by an analytical expression, for example

$$y = f(x)\,, \text{ or } w = f(x, y, z, \ldots) \qquad (7.1)$$

depending on whether the investigated variable depends on one or several quantities. Such expression, **regression function**, provides concentrated information and facilitates further processing of the data. The steps in this curve fitting are: 1) proposal of a suitable form of the regression function *f*, and 2) determination of the best values of its constants. In some cases it is necessary: 3) to evaluate the quality of the fit, especially if the fit is not perfect or if it is necessary to decide, which of the several possible approximations is the best.

Proposal of regression function

The first idea can be obtained from a chart with all measured values $y(x)$. **A picture says more than thousand words**! The possible shape of the regression function of several independent variables, $w = f(x, y, z, \ldots)$, can first be assessed from the plots $w = f(x; y = const, z = const)$, $w = f(y; x = const, z = const)\ldots$, corresponding to the cuts through it. Figure 7.1 shows shapes of various functions and can thus help with the choice of the regression function. Before proposing this function, it is useful to think for a while about the general nature of the investigated phenomenon. Generally, the regression function can increase or decrease monotonically, it can have a power-law character or can grow in an exponential or logarithmic manner, it can decrease as a hyperbola or an exponential function with negative exponent, or it can have a maximum or minimum. It can approach asymptotically to a certain value for very high values of the independent variable. The solution of some problems leads to periodical functions (sin, cos…). Sometimes, the approximation is sought in the form of a series (e.g. polynomial, Fourier with trigonometric terms or Prony series with exponential terms). The knowledge of analytical solution of related problems can often help in the proposal of the regression function. Shapes of several simple functions are shown in Fig. 7.1; see also [1, 2]. In some cases, functions typical for probability distributions (either probability density or distribution function) can be useful, for example normal distribution or Weibull distribution (see Figures 4.7 and 4.8 in Chapter 4). Two examples, a function with several exponential terms and a cosine function are given later in this chapter.

Determination of regression constants

Some universal programs (including Excel) offer several regression functions for fitting empirical data and can determine the parameters using their own algorithms. In such case it is sufficient to create the chart for the measured values *x* and *y*. Then, after a right-click on the data series, a pop-up menu appears and a suitable function can be selected, e.g. linear, polynomial, power-law, exponential or logarithmic. The application with Excel was described in Chapter 5. It is important to demand (from the menu) that also the expression for the regression function is shown in the chart, as well as the coefficient of determination r^2, characterising the quality of the fit (the detailed explanation of r^2 was given in Chapter 6). This is useful especially if various regression functions should be compared.

If a regression function is to be proposed, various criteria are considered. One such criterion is that the calculations done with this function should be relatively simple. Functions easy to work with are polynomials, such as $y = a + bx + cx^2 + dx^3 + \dots$ Also universal programs for curve fitting offer this and several other functions. Fortunately, polynomials can be used for more complicated functions also, if the original data are transformed in a suitable way. The Table 1 on the page following Figure 7.1 with typical forms of analytical curves, shows several functions that can easily be transformed to polynomial form.

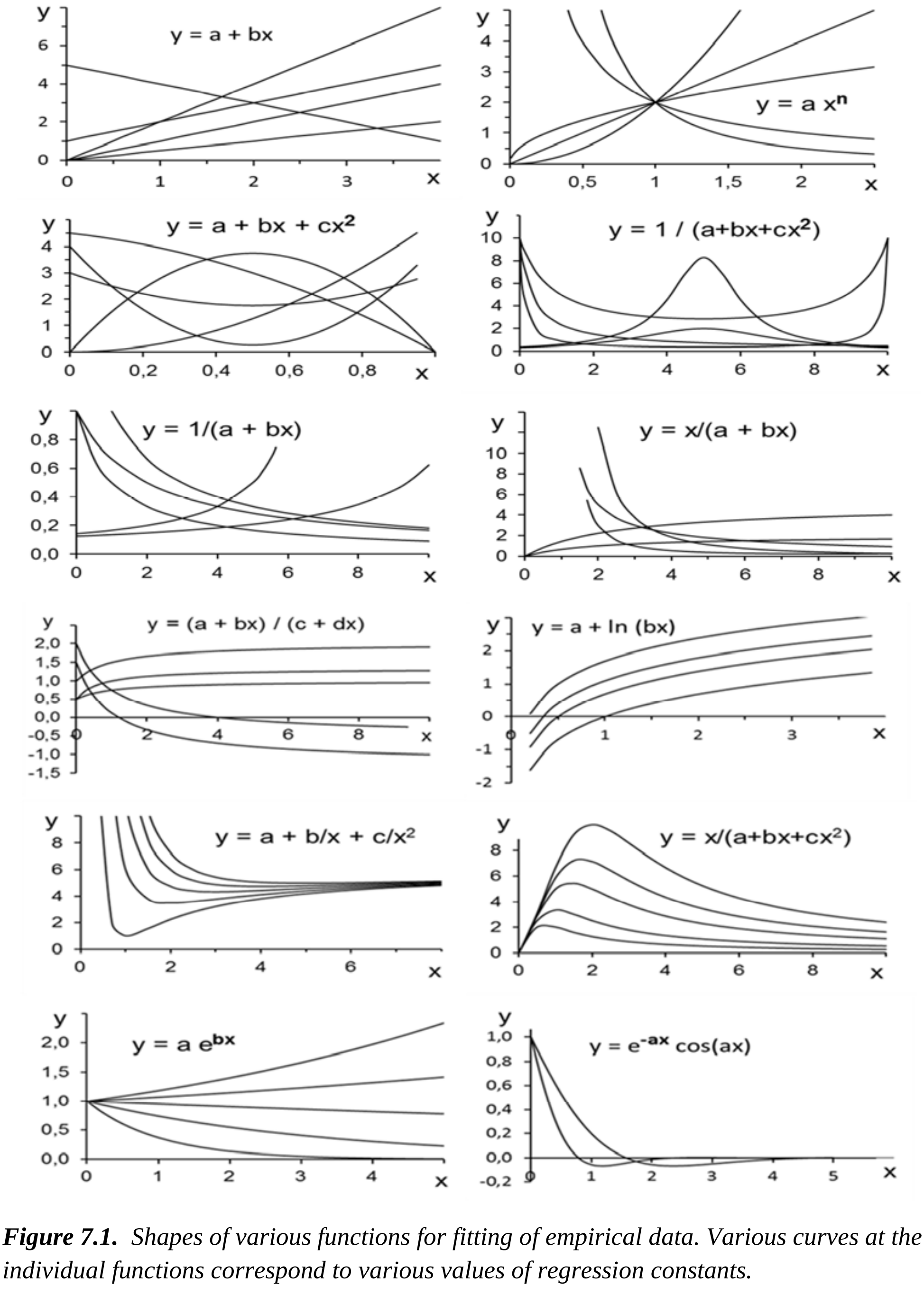


Figure 7.1. *Shapes of various functions for fitting of empirical data. Various curves at the individual functions correspond to various values of regression constants.*

Table 1. *Transformation of various functions to polynomial form [1].*

Original function	Transformation	Transformed function
$w = a + \frac{b}{x} + \frac{c}{x^2} + ... + \frac{k}{x^q}$	$t = \frac{1}{x}$	$y = a + bt + ct^2 + ... + kt^q$
$w = ae^{bx}$	$y = \ln w$	$y = \ln a + bx$
$w = \exp(a + bx + cx^2)$	$y = \ln w$	$y = a + bx + cx^2$
$w = ax^b$	$y = \ln w$ $t = \ln x$	$y = \ln a + bt$
$w = \frac{1}{a + bx + cx^2 + ... + kx^q}$	$y = \frac{1}{w}$	$y = a + bx + cx^2 + ... + kx^q$
$w^n = \frac{x^m}{a + bx + cx^2 + ... + kx^q}$	$y = \frac{x^m}{w^n}$	$y = a + bx + cx^2 + ... + kx^q$
$w = \ln(a + bx + cx^2)$	$y = \exp(w)$	$y = a + bx + cx^2$

Table 1 shows transformation formulae and transformed functions. However, such transformations change the character of the dispersion of the individual data points around the regression function. The classical determination of regression constants is based on the least squares method (see further), which gives the most accurate results if this dispersion is constant, independent of the values of the independent variable (= homoscedasticity). If the transformation has changed the dispersion significantly, transformed data *y* should be multiplied by appropriate weights. For more, the reader is referred to literature, e.g. [3].

The distribution of the measured values sometimes does not correspond to any of the predefined functions and it is better to propose one´s own expression (7.1). Figure 1 can help in search for a suitable expression; some of them are available in universal programs for curve fitting, such as Excel. The regression constants should be such that the distances between the individual measured values and the corresponding calculated values are minimal. For this purpose, usually the **least squares method** is used [4 – 6], which minimizes the sum of squared differences between the measured and calculated values of *y*:

$$SS_{\mathrm{res}} = \Sigma(y_{\mathrm{j,meas}} - y_{\mathrm{j,calc}})^2 \ = \mathrm{min}\ ! \qquad (7.2)$$

The subscript *res* means residual, *meas* – measured, and *calc* – calculated. The summation is done over all *n* values of $y(x_j)$. Solvers in universal programs enable this minimisation. An example was shown in Chapter 5 (Example 1); two other applications will be presented later in this chapter. These solvers have a big advantage: no transformation of regression function is necessary (and thus also no change of the character of dispersion).

Evaluation of the fit quality

Sometimes the visual evaluation is sufficient to give an unambiguous answer as to the fit quality. Also, a very simple quantity for such evaluation exists, the **coefficient of determination r^2**, explained in Chapter 6. The closer r^2 to 1, the better the fit. In a case of doubt, it is possible to make statistical test of significance of *r*, as explained in Chapter 8.

For more detailed characterisation, so-called **residuals** are suitable. They are defined as the differences between the measured and calculated values,

$$\Delta_j = y_{j,meas} - y_{j,calc} \qquad (7.3)$$

plotted as a function of the independent variable *x*. The differences between two curves are then more visible. This is useful especially if both curves, plotted in the original scales, overlap (see Figure 7.2a, b). Additional information follows from their distribution. Fit 2, with randomly dispersed positive and negative values is more suitable than Fit 1 with systematic gradual change of residuals from positive values to negative with increasing *x*.

Sometimes, **relative** (or standardised) **residuals** are used,

$$\Delta_{j,rel} = (y_{j,meas} - y_{j,\,calc})/y_{j,calc} \qquad (7.4)$$

which do not depend on the scale of *y*.

The differences between the measured values and those on the regression curve (7.1) can serve to three purposes: 1) their squares are used in the criterion in equation (7.2) for optimisation, 2) they can be used for verification whether the distribution of the individual points around the regression curve is normal (this is the condition for the use of the least squares method), and 3) they serve for the determination of confidence band around the regression curve. This band can be used for more reliable predictions than those based only on the regression function.

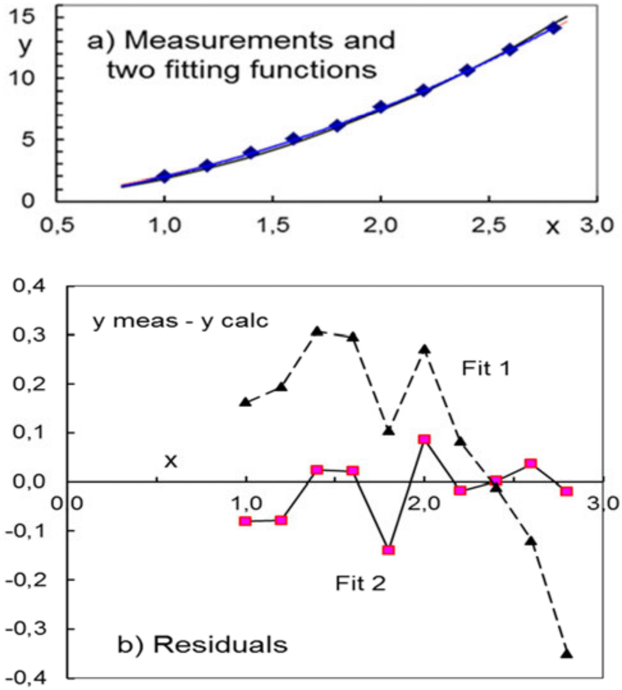


Figure 7.2. *a) Measured data fitted by two functions, b) Residuals; fit 2 is better than fit 1.*

Now, two methods for the determination of regression constants will be discussed in more detail: analytical and numerical.

Analytical determination of regression constants

A minimum of a function is usually found by the least squares method. This method is based on making partial derivatives of the sum (7.2) with respect to the individual regression constants, and putting each derivative equal zero [4 – 6]. For example, for linear regression

$$y = a + bx \tag{7.5}$$

Equation (7.2) changes to (with SS_{res} replaced by the symbol S):

$$SS_{\mathrm{res}} = S = \Sigma(a + bx_{\mathrm{j}} - y_{\mathrm{j,meas}})^2 \tag{7.6}$$

The partial derivatives $\partial S/\partial a$ and $\partial S/\partial b$ are

$$\partial S/\partial a = 2\Sigma(a + bx_\mathrm{j} - y_\mathrm{j,meas}) \ , \ \partial S/\partial b = 2\Sigma(a + bx_\mathrm{j} - y_\mathrm{j,meas})\, x_\mathrm{j} \tag{7.7}$$

Putting both expressions equal zero generates a system of two linear equations for two unknowns a and b. The solution gives

$$b = \frac{n\sum x_j y_j - \sum x_j \sum y_j}{n\sum x_j{}^2 - \left(\sum x_j\right)^2} \ , \ \ a = \frac{\sum y_j - b\sum x_j}{n} \tag{7.8}$$

Similar procedure can be used for other regression functions and leads, generally, to a system of n linear equations for n unknown regression constants. Simpler, however, is the use of a suitable computer solver, explained in Chapter 5 and in the next paragraph.

Computer-supported determination of regression constants

Today, universal computer programs (including Excel, Matlab or Mathcad) contain solvers which can find the minimum of an expression. This makes the determination of regression constants very easy. It is only necessary to prepare one series of measured data $y_\mathrm{j,meas}$ and a series of the $y_\mathrm{j,calc}$ values, calculated via Equation (7.1) for the same values x_j using the pertinent parameters, for example a, b in Equation (7.5). **Solver**, after the command to minimize the expression (7.2) or (7.5) by changing a and b, will find their best values by using its own algorithms. For these calculations, the cells for the regression constants (a, b) must be prepared in the worksheet in advance, as well as the cell containing the expression (7.1). The search for the best values of regression constants starts with assigning the initial values to the regression constants. Then, Solver is asked to find such values of the constants, for which the content of the cell with formula (7.6) is minimum. This value (SS_res) also characterises the quality of the fit (Chapter 6). This information is useful if the minimisation process is repeated. Examples are shown later.

The determination of regression constants with Solver needs some practice. The quality of the calculated “best” values of regression constants sometimes depends on their initial values. In the worst case, the optimisation process does not converge and different initial values must be chosen. Moreover, Solver looks for the constants ensuring a minimum from the mathematical point of view, and can propose values that have no real sense. In some cases it is necessary to define the interval of acceptable values of the constants. (Caution: the optimisation algorithm seeks only the nearest extreme, and does not know that several extremes can exist.)

Sometimes, the optimisation process must be done in two or more steps, as the first optimisation could not give the accurate values. The constants obtained in the first process are then used as input values for the repeated process. It can always be recommended to repeat the optimisation; the comparison of the sum of squared differences (7.2) from the subsequent optimisation cycles informs whether the search for the minimum has ended.

The determination of regression constants will be illustrated on two practical examples.

Example 1. Load response of a viscoelastic material

The time course of deformation of components from viscoelastic materials (for example plastics) under constant load resembles an exponential function. Often, however, simple exponential function is not sufficient. A better approximation may be a Prony series, which is a sum of several exponential functions,

$$y = a_0 + a_1 \exp(-t/\tau_1) + a_2 \exp(-t/\tau_2) + \ldots \qquad (7.9)$$

a_1, a_2… are constants and τ_1, τ_2,… are so-called relaxation times, which are also unknown. Figure 7.3 shows the time course of penetration of an indenter into polymethylmethacrylate (PMMA). The following regression function was used [7]:

$$y(t) = F\,K\,[A_0 + c_v t - \sum B_j \exp(-\,t\,/\tau_j)] \qquad (7.10)$$

F is the load, K is a constant for the indenter geometry, and A_0, c_v, B_j and τ_j (j = 1, 2, 3) are the regression constants, found by the least squares method. Figure 7.3 shows two approximations, with three and six regression constants. For comparison, also relative residuals Δ_{rel} in both cases are shown in the figures. They are defined by Eq. (7.4) and do not depend on the scale of y. Residuals can help in distinguishing various approximations, especially if they look nearly identical in the common $y(x)$ coordinates (see also Fig. 7.2). In the investigated case, the approximation with six constants is obviously better.

Remark: Commercial computer programs for the finite element analysis of structures enable work with Prony series.

Confidence band for predicted values

Regression function usually serves for future prediction. However, if this function was created from data exhibiting large scatter, the predictions will not be very

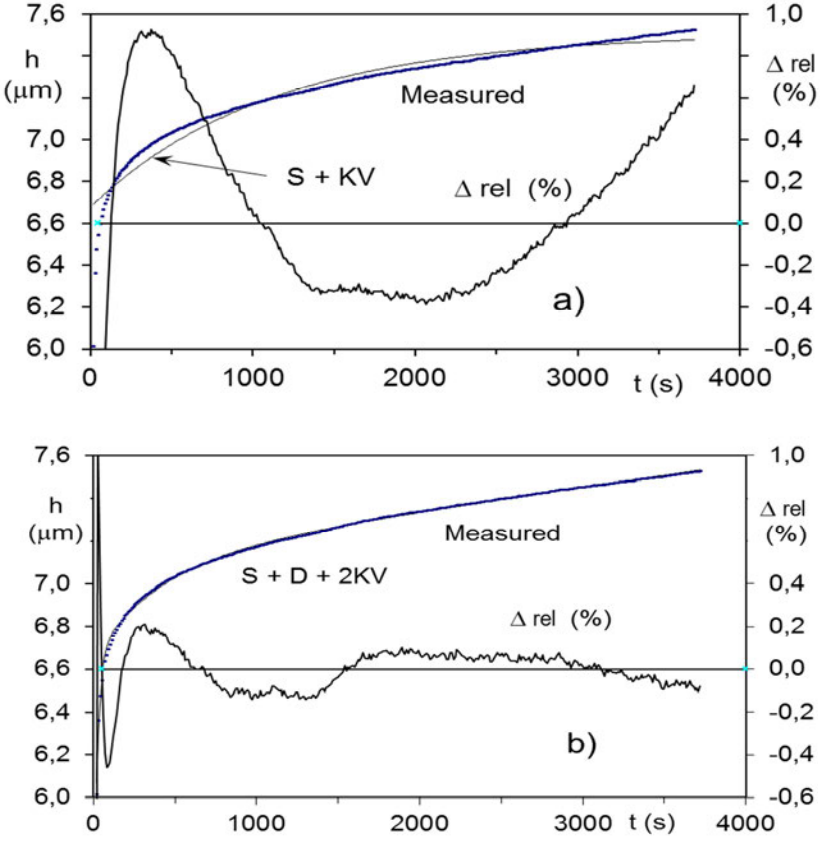


Figure 7.3. *Indenter penetration into PMMA under constant load [7]. Measurements (dotted curves), fits for two models (thin solid curves), and relative residuals Δ_{rel} (thin zig-zag lines). a) model S+KV (3 constants), b) Model S+D+2KV (6 constants). The S+D+2KV model fits the measured data very well; the differences are visible only via the residuals. h – depth, t – time. S - spring, D - dashpot, KV - Kelvin-Voigt body (spring and dashpot in parallel).*

reliable. Its reliability can be increased by creating confidence band for the individual points around the regression function. If one can assume that the distribution of the measured points y_j around the regression function $y_{reg}(x)$ is normal and the corresponding residual variance (determined from many points) is constant, independent of x, it is possible to construct the boundaries of the confidence interval (see Fig. 7.4 in Example 2) as

$$y_{\mathrm{U,L}}(x) \;=\; y_{\mathrm{reg}}(x) \pm |u_\alpha|\, s_{\mathrm{res}} \qquad (7.11)$$

U and L denote the upper and lower boundary limit, respectively, u_α is α-quantile of standard normal distribution, and s_{res} is the residual standard deviation defined as

$$s_{\mathrm{res}} = \sqrt{\,[\, SS_{\mathrm{res}}/\nu \,]} \qquad (7.12)$$

SS_{res} is the sum of squared differences between the experimental data points and the corresponding points on the regression curve, and ν is the number of degrees of freedom, equal to the number of measured values minus the number of regression constants (for example 2 in linear relationship). Probability that the predicted value will lie outside the limits ($y_{\mathrm{U,L}}$), is 2α.

Example 2. Fitting of the average daily temperatures during a year

The outside temperatures vary during a day and also during a year. Nevertheless, these variations exhibit some regularity (day, night, summer, winter…). If this regularity is taken into account, the predictions of temperatures at certain time can be more accurate. For example, temperatures in the town Ústí were monitored during the year 2008. It appeared that the average daily temperatures can be described by the following cosine function:

$$T \;=\; T_0 + A \cos\,[\pi(x_j - x_0)/186] \qquad (7.13)$$

x_j is the rank-order number of the day, and T_0, A and x_0 are constants. (NOTE: 186 days is the half-length of the analysed leap-year 2008). Figure 7.4 shows the individual temperatures and the regression function (7.13), found by the Solver in Excel. The regression constants were T_0 = 10.86°C, A = –9.41°C, x_0 = 14.17. The residual standard deviation was 3.236°C.

Now, let us predict the average temperature on 23rd October. Determine also the confidence interval, which will contain the true average temperature with probability 90%.

In Equation (7.13), 23rd October has the rank number x = 297, and the predicted temperature is

$$T = 10.86 - 9.41 \cos[\pi(297 - 14.17)/186] = 9.52°\mathrm{C}$$

Figure 7.4 also shows the confidence intervals for the temperatures (see further).

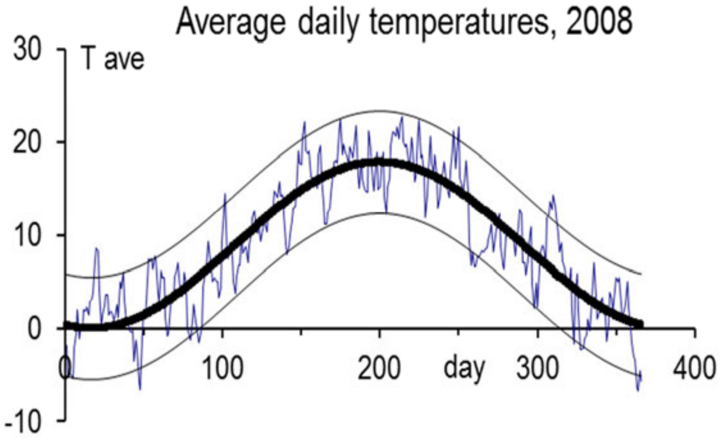


Figure 7.4. *Average daily temperatures in town Ústí during the year 2008. Measured data, the fit and confidence band. T ave – average daily temperatures.*

Construction of the confidence band. The confidence intervals for various characteristics will be explained in more detail in the next chapter. The confidence band for the temperatures can be constructed via Equation (7.10). With residual standard deviation s_{res} = 3.236°C and 5% quantile of standard normal distribution (1.645), the half width of confidence interval is $\Delta = s_{res} \times u_{0.05} = 3.236 \times 1.645 = 5.32$, and the lower and upper limits of the 90% confidence interval are: $T_L = 9.52 - 5.32 = 4.20$°C, $T_U = 9.52 + 5.32 = 14.84$°C. The temperature 7.43°C, measured for this day (297), lies within the confidence limits; cf. Fig. 7.4. One also can see from this chart that a few temperatures during the year are out of the limits. This is understandable because the limits were constructed for confidence 90%; ten percent of all measured values may lie outside this confidence band.

Multiple regression

Often, one must express how the variable y depends on several input quantities x_1, x_2, …:

$$y = f(x_1, x_2, \dots) \tag{7.13}$$

The simplest case is linear relationship

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + \dots \tag{7.14}$$

The constants a_0, a_0,… can be obtained easily by multiple linear regression, available by universal computer programs. Here, the use of Excel´s function *LINREGRESSION* will be shown on an example.

Example 3. Multiple linear regression.

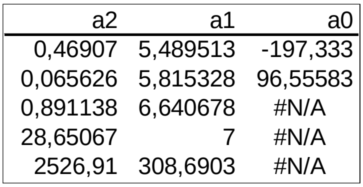
Let us have n = 10 values of strength y of an alloy measured for various content of Mg (x_1) and the temperature of thermal treatment (x_2); see the table below on the left. We shall assume that the strength depends on them as $y = a_0 + a_1x_1 + a_2x_2$.

Input data

y	x_1(%)	x_2(°C)
234	10.3	800
256	9.8	840
290	9.6	920
248	10.0	820
255	9.7	870
260	10.5	850
285	9.4	900
250	9.1	860
244	9.9	830
270	10.2	890

First, we create (by “click and drag”) an empty array (matrix) with five horizontal rows and the number of columns equal the number of regression constants; in this case 3. In the next step we open the Excel menu “Insert function”, find *LINREGRESSION*, insert the array of y values from the table into the upper window in this menu, insert the array containing all input values x_1, x_2 into the window below, and then we write the word *TRUE* into the two lowest windows and press simultaneously the keys *CTRL*, *SHIFT*, *ENTER*. That´s all. The table of results is given on the next page. The reader is encouraged to repeat the procedure.

The individual horizontal rows in the table of results contain: row R1: regression constants arranged from left to right as a_2, a_1, a_0; row R2: standard deviations of the individual regression constants, row R3: coefficient of determination r^2 and the standard deviation of y, row R4: F-statistics and the number of degrees of freedom, *Table of results*



a2	a1	a0
0,46907	5,489513	-197,333
0,065626	5,815328	96,55583
0,891138	6,640678	#N/A
28,65067	7	#N/A
2526,91	308,6903	#N/A

needed for testing whether the relationship among the dependent and independent variables is not only random, and row R5: the regression sum of squares SS_{reg} and the residual sum of squares SS_{res}, defined in Chapter 5. The three cells with #N/A do not contain any values. Detailed explanations can be found at the command *LINREGRESSION* in Excel menu. The coefficient of determination is $r^2 = 0.891$ (see R3), which is acceptable. The reader can calculate y for chosen values x_1, x_2 and compare it with the table of input data.

The regression function (see the constants in the first row of the above table) is

$$y = -197.333 + 5.48951\, x_1 + 0.46907\, x_2$$

The above facility for multiple linear regression can sometimes be used for finding constants in nonlinear regression also. For example, the product

$$y = a\, x_1 x_2 x_3 \tag{7.15}$$

can be changed by logarithmic transformation to the sum

$$z = b + u_1 + u_2 + u_2 \tag{7.16}$$

$z = \log y$, $b = \log a$, $u_1 = \log x_1$, $u_2 = \log x_2$ and $u_3 = \log x_3$. Similarly, it is possible to transform individual variables. For example, in equation $y = a_1 x_1 + a_2 x^3 + a_3 \sin(cx)$ new variables can be defined as $w = x^3$ and $v = \sin(cx)$. Note: The regression constant b must then be transformed back to the original system as $a = 10^b$. If the shape of the assumed expression does not allow multiple linear regression, the regression constants can be found by using a suitable solver, as described above.

REMARK. The direct determination of parameters in Weibull distribution from measured values, described in Chapter 5, is nothing else than the determination of constants in a regression function.

Moving averages

Sometimes we have no idea what function could be used for the approximation of the time course of some quantity expressed as time series. We just see plenty of data points. A trend can sometimes be revealed better if the original data are smoothened by replacing them by moving averages. Such average is calculated from *p* neighbouring data points; the number *p* is a matter of our choice. For example, the original series x_1, x_2, x_3, x_4, …x_n is replaced (for a chosen number p = 3) by the series $y_1 = (x_1+x_2+x_3)/3$, $y_2 = (x_2+x_3+x_4)/3$, $y_3 = (x_3+x_4+x_5)/3$,… $y_{n-2} = (x_{n-2}+x_{n-1}+x_n)/3$. This new series has only $n - 2$ terms, generally $n - (p - 1)$, and is smoother than the original one. It can be recommended to use several approximations, for various *p*, and choose the best looking one. Universal computer programs (including Excel) enable easy application of moving averages on empirical data.

References to Chapter 7

1. Kropáč, O.: Methods of experimental research. (In Czech: Metody experimentálního výzkumu.) ČVUT, Praha, 1979. 139 p.
2. Pechoč, V.: Evaluation of measurement and computing methods in chemical engineering. (In Czech: Vyhodnocování měření a početní metody v chemickém inženýrství.) SNTL, Praha, 1981. 226 p.
3. Meloun, M., and Militký, J.: Statistical Data Analysis. A practical Guide. Woodland Publishing India Pvt. Ltd. New Delhi, 2011. 773 p.
4. Felix, M., and Bláha, K.: Statistical methods in chemical industry. (In Czech: Matematickostatistické metody v chemickém průmyslu.) SNTL, Praha, 1962. 336 p.
5. https://en.wikipedia.org/wiki/Regression_analysis. January, 14, 2017.
6. Freund, J. E., and Perles, B. E.: Modern elementary statistics. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 2006 (12th edition). 576 p.
7. Menčík, J., He, L.H., and Němeček, J.: Characterization of viscoelastic-plastic properties of solid polymers by instrumented indentation. Polymer Testing, 30, 2010, pp. 101 - 109.

8. Confidence Intervals, Testing of Hypotheses and the Amount of Data

An important task at the initial stage of any research is finding the necessary extent of experiments. Experiments cost money (specimens and the related material, the devices and other equipment that must be purchased or hired), and they also need time and work capacity (plus corresponding personal expenses). Therefore, their extent should not be excessively large, but in proportion to the: 1) task of the research, 2) importance of the expected results, and 3) the demanded accuracy. All this should be clarified in advance.

Information on the accuracy of a measured parameter is provided by the confidence interval for this parameter. The width of such interval depends on the number of measurements. Vice versa, the formula for the necessary number of values can be derived from the expression for the confidence interval. Similarly, the decision about a tested hypothesis is based on the value of the pertinent test criterion, which also depends on the number of measurements. This number, important for ascertaining or increasing the test power, can again be derived from the test criterion.

In this chapter the formulae for confidence intervals and statistical tests will be shown, and also the numbers of measurements needed for ensuring the demanded accuracy of some characteristics (mean, standard deviation or other parameters), the number of values for a confidence interval for points on a regression line, and the amount of data for some statistical tests. Each section will start with the pertinent formulae, and the applications will be shown on examples.

Confidence intervals and the necessary numbers of values

Mean value

The confidence interval for the mean μ is [1 – 5]:

$$\bar{x} - t_{\alpha,\nu}\, s/\sqrt{n} \;\leq\; \mu \;\leq\; \bar{x} + t_{\alpha,\nu}\, s/\sqrt{n} \qquad (8.1)$$

$\bar{x}$ is the average value (= $\Sigma x_j/n$), n is the number of measured values, s is the sample standard deviation, and $t_{\alpha,\nu}$ is α-critical value of t-distribution for $\nu = n - 1$ degrees of freedom. Half-width of the confidence interval,

$$\Delta = t_{\alpha,\nu}\, s/\sqrt{n} \qquad (8.2)$$

expresses the uncertainty in the determination of the mean value and corresponds to the possible inaccuracy. This equation can be rewritten to express the number of values (or tests) necessary for ensuring that the true mean μ will differ from the average $\bar{x}$ not more than Δ:

$$n = (t_{\alpha,\nu}\, s/\Delta)^2 \qquad (8.3)$$

The probability that a larger difference can occur, is α. One can see that the necessary number of tests increases significantly with increasing dispersion of the individual values and with increasing demands for accuracy (i.e. with smaller allowable error Δ). It can be said roughly that the reduction of the inaccuracy to 50% needs four-times more tests. Certain mitigating role is played by the fact that $t_{\alpha,\nu}$ decreases with the increasing n, especially for small n.

According to Equation (8.3) it would be possible, in principle, to achieve any accuracy, but for a high price. Therefore a compromise must often be made. It is reasonable to make only a few tests first in order to obtain an estimate of standard deviation s and to calculate a preliminary number n of the necessary tests via Eq. (8.3). If n is high, it is reasonable to make about half of the tests at the beginning, to calculate the improved estimate of s and n (using the corrected value of t), and then to make the remaining tests.

Example 1.

Diameters of machined shafts, measured on 10 pieces, were: D = 16.02 – 15.99 – 16.03 – 16.00 – 15.98 – 16.04 – 16.00 – 16.01 – 16.01 – 15.99 mm. Calculate: a) the average value and standard deviation. Assume that the diameters have normal distribution, and calculate b) the 95% confidence interval for the mean value and also c) the interval, which will contain 95% of all diameters.

Solution.

a) The average value is $\bar{D} = (\Sigma D_i)/n$ = 16.007 mm and standard deviation is s = 0.01889 mm.

b) 5%–confidence interval for the mean, calculated by Eq. (1), is (for two-sided critical value $t_{0,05;\,10-1}$ = 2.2622):

$$16.007 - 2.2622\,\frac{0.01889}{\sqrt{10}} < \mu_D < 16.007 - 2.2622\,\frac{0.01889}{\sqrt{10}}$$

$$15.993 < \mu_\mathrm{D} < 16.020 \text{ mm, or } \mu_\mathrm{D} \in (\bar{D} \pm \Delta) = 16.007 \pm 0.0135 \text{ mm}$$

If we want to increase the accuracy in the determination of the mean value of the diameter so that the actual mean differs from the calculated average $\bar{D}$ not more than Δ = 0,005 mm, Equation (8.3) gives

$$n = (t_{\alpha,\nu}\, s/\Delta)^2 = (2.002465 \times 0.01889/0.005)^2 = 57.2$$

2.0025 is the critical value of *t*-distribution $t_{\alpha,\nu}$ for significance level α = 5% and the number of degrees of freedom $\nu = n - 1 = 57$. The standard deviation s = 0.01889 mm as before was used, as no better estimate was available. Therefore, the improved mean value will be obtained as the average of not less than 58 values.

c) The individual values can be expected (under assumption of normal distribution) to lie within the interval $\bar{D} - u_{\alpha/2}\times s < D < \bar{D} + u_{\alpha/2}\times s$, where $u_{\alpha/2}$ is $\alpha/2$ – critical value of standard normal distribution (corresponding to probability $\alpha/2$ that the diameter will be larger than the upper limit of the confidence interval, and $\alpha/2$ that it will be smaller than the lower limit). In our case, $u_{0.025} \approx 1.96$, so that $16.007 - 1.96\times 0.01889 < D < 16.007 + 1.96\times 0.01889$; that is $D \in (15.970;\ 16.044)$. The reliability of the prediction could be increased if ***tolerance interval*** is used instead of confidence interval; see later in this chapter.

Variance

The confidence interval for the variance σ^2 of normal distribution is [1 – 5]

$$(n-1)s^2/\chi^2_{\alpha/2,\nu} \le \sigma^2 \le (n-1)s^2/\chi^2_{1-\alpha/2,\nu} \qquad (8.4)$$

$\chi^2_{\alpha/2,\nu}$ is $\alpha/2$–critical value of chi-square distribution for $\nu = n-1$ degrees of freedom; $\chi^2_{1-\alpha/2,\nu}$ is $(1-\alpha/2)$-critical value. (NOTE: α–critical value is identical with $(1-\alpha)$-quantile.) The width of confidence interval depends on the number of values *n*. The number of measurements needed for obtaining the demanded width can be obtained using the relationship between the number of values and the corresponding critical values of chi-square distribution. Universal programs with statistical functions (including Excel) are suitable for this task.

Example 2.

Calculate the 90% confidence interval for the standard deviation *s* from Example 1.

Solution. Equation (4) gives the confidence limits for variance. The variance for standard deviation s = 0.01889 mm is s^2 = 0.0003658 mm^2. Further necessary values are: $\nu = n - 1 = 10 - 1 = 9$, $\chi^2_{0,05}(\nu = 9) = 19.9190$, $\chi^2_{0,95}(\nu = 9) = 3.3251$. Lower (L) and upper (U) confidence limits for the variance are

$$s^2(\mathrm{L}) = 9\times0.0003658/19.9190 = 0{,}0001612 \text{ mm}^2,\quad s_\mathrm{L} = 0.0126975 \text{ mm}$$
$$s^2(\mathrm{U}) = 9\times0.0003658/3.3251 = 0{,}0009901 \text{ mm}^2,\quad s_\mathrm{U} = 0.0314659 \text{ mm}$$

and the confidence limits for standard deviation, calculated as square roots of the variances, are s_L = 0.0127 mm, s_U = 0.0315 mm. (Note the large width of confidence interval for *s*, compared with the estimated value s = 0.0189 mm !)

Parameter of exponential distribution

Exponential distribution plays a very important role, for example, in reliability. It is usual for the times between failures occurring from many reasons in complex electrical, mechanical and other objects or systems consisting of many elements. Probability of failure during interval (0; *t*) is

$$R(t) = \mathrm{e}^{-t/T}{}_{\mathrm{mean}} \qquad (8.5)$$

t is the time and T_mean is the mean time to failure or between failures. In practice, the mean time is determined as the average of the measured times to failure,

$$T_\mathrm{mean} \approx T_\mathrm{ave} = \Sigma t_\mathrm{fj} / n \qquad (8.6)$$

the summation is done for all n values t_fj. However, the times to failure of individual elements vary, and T_ave, calculated from Eq. (8.6), is only an estimate of the mean time. The knowledge of **confidence limits** for T_0 is therefore needed. The lower (L) and upper (U) limit are given by the following formula [1 – 5]:

$$t_L = \frac{2r}{\chi^2_{\alpha/2}(\nu)}\bar{t} \le T_{mean} \le \frac{2r}{\chi^2_{1-\alpha/2}(\nu)}\bar{t} = t_U \qquad (8.7)$$

$\chi^2_{\alpha/2}(\nu)$ is the $\alpha/2$-critical value and $\chi^2_{1-\alpha/2}(\nu)$ is the $(1-\alpha/2)$–critical value of the chi-square distribution for ν degrees of freedom. The probability that the actual time to failure can be shorter than t_L or longer than t_U is α. The number of degrees of freedom depends on the arrangement of the tests. If they are terminated after failure of r pieces, it holds $\nu = 2r$ [6]. With increasing number r, the difference between the lower and upper critical values of chi-square distribution becomes smaller. Also the confidence interval for the mean time becomes narrower and the prediction of T_{mean} more accurate. In this way it is possible to determine in advance the number of failed specimens, at which the test may be terminated to achieve the demanded width of confidence interval for the mean time to failure.

Example 3.

Ten electrical components were tested to determine their times to failure and the failure rate. These tests can last very long for some components. Therefore, they are sometimes terminated after certain defined time. In this example, the duration of the tests was fixed as $t_T = 500$ hours. During this time, only 6 components failed ($r = 6$), in times: 65 – 75 – 90 – 120 – 250 – 410 hours. Four components survived the test. It is necessary to estimate the mean time to failure and construct two-sided confidence intervals (for the confidence 90%).

Solution. The mean value and standard deviation of times to failure of the 6 failed components were, respectively: 168.33 and 136.33 hours. It is thus possible to assume exponential distribution.

The cumulated duration of the tests was [6]:

$$t_{tot} = \sum_{i=1}^{6} t_i + 4 \times t_t = 60 + 75 + 90 + 120 + 250 + 410 + 4\times 500 = 3010 \text{ hours}$$

The average time to failure is $t_{ave} = t_{tot}/r = 3010 / 6 = 501.67$ h.

The lower and upper confidence limit for t_{mean}, with respect that the tests were terminated before the failure of all samples, are [6]:

$$t_L = \frac{2r}{\chi^2_{\alpha/2}(2r+2)} t_{ave} \le t_{mean} \le \frac{2r}{\chi^2_{1-\alpha/2}(2r)} t_{ave} = t_U \qquad (8.8)$$

where $\chi^2_{\alpha/2}(2r+2)$ is $\alpha/2$–critical value of chi-square distribution for $2r+2$ degrees of freedom, and $\chi^2_{\alpha/2}(2r)$ is $\alpha/2$ – critical value of chi-square distribution for $2r$ degrees of freedom. In the investigated case, with r = 6 and α = 10%, the critical values are $\chi^2_{0.05;\,14}$ = 23.685 and $\chi^2_{0.95;\,12}$ = 5.226. Inserting them, together with t_{ave} = 501.67 h into (6.8) gives t_L = 254.4 h and t_U = 1152.1 h. The mean time to failure thus can be expected to lie within the interval $t_{mean} \in$ (254 h; 1152 h).

This confidence interval, obtained from only six failures, is very wide. If it should be narrower (in order to get more accurate estimate), it is necessary to make a longer test so that more parts of the tested group fail, or to increase the number of tested parts; see [6] or Chapter 20 in [7].

Values predicted by a regression line

Regression line

$$y(x) = a + bx \qquad (8.9)$$

is often used for prediction of y-values corresponding to certain values of x. However, the constants a, b are determined from measured values that exhibit some dispersion. If another series of measurements would be used, less or more different regression line would be obtained. The confidence interval for a point of the regression line (Figure 1) is [1–4, 8]:

$$y = a + bx \pm t_{\alpha,\nu}\, s_{res} \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}} \qquad (8.10)$$

where $s_{res}(x)$ is the residual standard deviation, defined as

$$s_{res} = \sqrt{\frac{\sum(y_j - a - bx_j)^2}{n - 2}} \qquad (8.11)$$

All points of confidence limits form the confidence band, which is narrowest for $x = \bar{x}$. The formula for $x = \bar{x}$ is identical with the formula (8.1) for the confidence interval for the mean.

Similarly, the width of confidence interval at $\bar{x}$ decreases with the square root of $1/n$. The confidence band for the regression line can therefore by made narrower by using higher number of values for the determination of regression constants. The

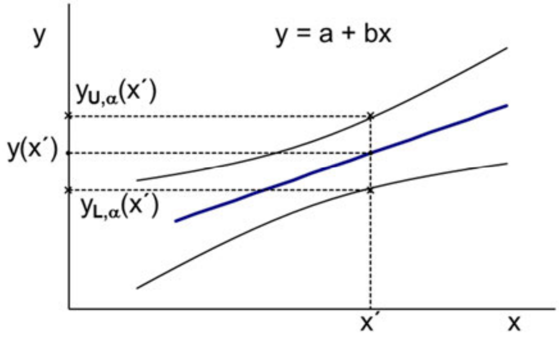


Figure 8.1. *Confidence interval for a regression line.*

necessary number of values for ensuring the demanded accuracy can be found similarly to the case of the mean value. A modification of Equation (8.3) gives

$$n = (t_{\alpha,\nu}\, s_{\mathrm{res}}/\Delta)^2 \tag{8.12}$$

Tolerance limits

Sometimes we need to know the interval that will include *P* percent of the population. A very important case is if the population has (approximately) normal distribution. If the parameters μ, σ of the population are known, then *P* % of the population lies within the limits $\mu \pm u_{1\text{–P/2}}\sigma$, where $u_{1\text{–P/2}}$ is the (1–*P*/2)–critical value of standard normal distribution.

Often, however, only sample characteristics $\bar{x}$ and *s* are known instead of μ and σ. In such case, it is impossible to determine the corresponding limits with certainty. We can only determine so-called tolerance limits, which will contain the fraction *P* of the population with a chosen probability γ. Two-sided tolerance limits (lower and upper) can be calculated via the formula [5, 9, 10]

$$x_{\mathrm{L}}, x_{\mathrm{U}} = \bar{x} \pm ks \tag{8.13}$$

$\bar{x}$ and *s* are the average and standard deviation of the sample of size *n*, and *k* is a constant, depending on *P*, *n* and γ. The coefficients *k* for selected values *P*, *n* and γ can be found in statistical tables, for example [9, 10].

Example 4.

In Example 1 the interval was calculated, which should contain 95% of all machined shafts: (15.970 mm; 16.044 mm). This interval was calculated under the assumption that $\overline{D}$ = 16.007 mm and s = 0.01889 mm are parameters of the population. These values, however, were calculated from a sample of only n = 10 pieces. More reliable interval will thus be obtained via the formula for the *tolerance interval*. For the fraction P = 0.95, reliability of the prediction γ = 0.90 and n = 10 is k = 3.18; see [9, 10]. The interval containing 95% of all pieces is

$$\overline{D} \pm k \times s = 16.007 \pm 3.18 \times 0.01889 = 16.007 \pm 0.060 = (15.947\text{ mm}; 16.067\text{ mm})$$

This is wider than the original interval. The difference between both intervals gets larger with larger standard deviation and smaller amount of empirical data, especially if n < 10.

Testing of hypotheses

Difference of two averages

The test criterion depends on whether the variance of both samples is (approximately) the same or not. Here, only the case with different variances will be considered, which is more universal. In this case, the test characteristic is

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{s_1^{\,2}}{n_1} + \dfrac{s_2^{\,2}}{n_2}}} \tag{8.14}$$

and it will be compared with the critical value of t-distribution for the significance level α and ν degrees of freedom, defined as [8, 11]

$$\nu = \frac{\left[(s_1^{\,2}/n_1) + (s_2^{\,2}/n_2)\right]^2}{\dfrac{(s_1^{\,2}/n_1)^2}{n_1 - 1} + \dfrac{(s_2^{\,2}/n_2)^2}{n_2 - 1}} \tag{8.15}$$

Example 5.

A modified procedure for preparation of a certain kind of plastic was proposed. The costs were lower, but also the measured strength was lower. It is necessary to verify whether the strength decrease is only random, or if it is statistically significant. The characteristics of both samples are as follows:

Sample 1: n_1 = 31, $\bar{x}_1$ = 31.03, s_1^2 = 1.41, s_1 = 1.19
Sample 2: n_2 = 32, $\bar{x}_2$ = 29.87, s_2^2 = 1.84, s_2 = 1.36

Insertion of these values into Equations (8.14) and (8.15) gives ν = 60.4 and t = 3.615. This test characteristic t is larger than the critical value of t-distribution for significance level α = 0.05 and the number of degrees of freedom ν = 60, which is $t_{0.05;\,60}$ = 2.0003. The difference between both average values is significant (on level 5%), which means that the way of preparation has influence of the strength.
NOTE: The difference is statistically significant even on confidence level 0.001.

Comparison of the accuracy of two measuring methods

This test is based on the comparison of variances of both methods. The ratio of two variances has F-distribution. The test criterion,

$$F = s_1^2 / s_2^2 \qquad (8.16)$$

will be compared with the critical value of F-distribution for $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ degrees of freedom, $F_\alpha(n_1 - 1,\ n_2 - 1)$. If $F > F_\alpha$, the null hypothesis (no difference between both variances) is rejected. Otherwise we conclude that the difference is not significant.

Test of significance of the coefficient of correlation of two quantities

The correlation coefficient r (= $\sqrt{r^2}$) is sometimes very high, for example 0.9 or more, and we can assume that the proposed functional relationship between both quantities is justified. Sometimes, however, the correlation coefficient is lower and we do not know whether the values of one quantity really depend on the values of the other quantity, or if they are correlated with it only loosely. Statistical test is then useful. Correlation coefficient r is statistically significant on significance level α, if

$$\frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} \geq t_{\alpha,\nu} \qquad (8.17)$$

$t_{\alpha,\nu}$ is one-sided α-critical value of t-distribution for $\nu = n - 2$ degrees of freedom; n is the number of pairs of values. If we want to be sure that the correlation coefficient r is statistically significant (on level α), it follows from (17) that the number of the pairs must be

$$n \geq 2 + (t_{\alpha,\nu})^2 (1 - r^2)/r^2 \qquad (8.18)$$

As $t_{\alpha,\nu}$ depends on n, several iterative steps are sometimes needed to obtain the right number of values n.

Example 6.

Correlation coefficient between strength and hardness of an alloy, determined from 30 pairs of values, was r = 0.7. Is this value statistically significant?

The test criterion (8.17) is

$$\frac{|0.7|\sqrt{30-2}}{\sqrt{1-0.7^2}} = 5.1867$$

One-sided critical value of t-distribution for $\nu = n - 2 = 30 - 2 = 28$ degrees of freedom is $t_{0.05,28} = 1.701$ for confidence level 5%, and 2.763 for confidence level 0.5%. The calculated value 5.1867 is much higher than the critical values. We can thus conclude that strong correlation exists between the strength and hardness of this material.

Tests of goodness-of-fit. These tests are used to check whether the experimental data have certain probability distribution. Two kinds of tests are used most often: Kolmogorov-Smirnov and chi-square. With **Kolmogorov-Smirnov test**, the differences between the empirical distribution function and the reference one are calculated for all values of the empirical distribution, and the maximum difference is compared with the critical value, which can be found in special tables [9, 10]. If it is larger, we reject the null hypothesis and say that the empirical population does not correspond to the assumed distribution. Otherwise we accept the hypothesis. The following example illustrates the application of Kolmogorov-Smirnov test.

Example 7.

It is necessary to verify whether the batch of NaOH comes from the supply with the mean concentration $\mu = 42.3$ and standard deviation $\sigma = 1.5$. The results of $N = 50$ analyses were:

44.0	45.0	42.5	41.9	41.9	41.2	41.0	43.3	42.3	40.0
42.2	43.6	44.4	40.2	41.5	42.7	45.0	43.0	43.2	41.1
42.4	42.4	40.8	39.2	43.2	44.2	42.3	43.3	44.7	41.1
41.9	42.7	41.4	44.4	43.0	40.1	42.0	39.7	42.9	42.7
41.5	38.8	43.4	43.9	40.8	40.5	42.8	41.6	43.0	42.8

The basic statistics are: average concentration $\bar{x}$ = 42.27, standard deviation s = 1.48, the minimum and maximum values: $x_{\min}$ = 38.8, $x_{\max}$ = 45.0.

Solution. For Kolmogorov-Smirnov test the values of the empirical and theoretical distribution functions are needed. First, the measured concentrations (x) were rank-ordered from minimum (x_1) to maximum (x_N) and the corresponding values of empirical distribution function were calculated as $F_{\mathrm{j,emp}} = j/(N+1)$. Then, the values of theoretical distribution function were calculated for the same quantiles x_{j}, but under the assumption that they pertain to the normal distribution with parameters μ = 42.3 and σ = 1.5. (In Excel, the command NORMDIST(x_{j},μ,σ,TRUE) can be used.) Finally, the differences of both distribution functions, $\Delta_{\mathrm{j}} = |F_{\mathrm{j,emp}} - F_{\mathrm{j,theor}}|$, were calculated. The table below shows a part of the complete table, and Figure 8.2 shows both distribution functions.

The maximum difference between the empirical and theoretical distribution function was $\Delta_{\max}$ = 0.05612 (for j = 28; see the table). This is much less than the critical value of the Kolmogorov-Smirnov criterion [10], which is $D_{0.05}(50)$ = 0.188 for confidence level α = 0.05 and the number of values N = 50. Therefore we can accept the hypothesis that the parameters of the batch are μ = 42.3 and σ = 1.5.

j	x_{j}	$F_{\mathrm{j,emp}}$	$F_{\mathrm{j,theor}}$	Δ_{j}
1	38.8	0.01961	0.00982	0.00979
2	39.2	0.03922	0.01938	0.01983
……………………..				
28	42.7	0.54902	0.60514	0.05612
……………………..				
50	45.0	0.98039	0.96407	0.01632

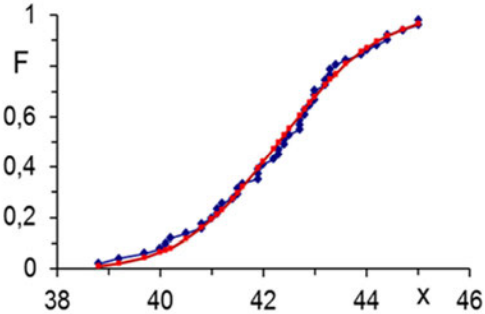


Figure 8.2. *Kolmogorov-Smirnov test – comparison of empirical and theoretical distribution for 50 values.*

Another goodness-of-fit test is **chi-square test** (χ^2–test). This test is based on the idea that if the sample has the assumed distribution, the differences of empirical and assumed (theoretical) values have standard normal distribution and the sum of their squares has therefore chi-square distribution. The application is as follows. The data are divided into *m* intervals and the frequency of their occurrence in the individual bins is determined. Then, the theoretical frequencies are computed for the assumed distribution. The differences of both frequencies in the corresponding classes are calculated. The criterion is [11]:

$$\chi^2 = \sum_{j=1}^{m} \frac{(n_j - np_j)^2}{np_j} \tag{8.19}$$

If the value χ^2, calculated via Equation (8.19), is higher than the α-critical value of the χ^2–distribution for (*m–k–1*) degrees of freedom, the hypothesis "the sample has the assumed distribution" is rejected on the level of significance 1–α. Otherwise, the hypothesis is accepted. REMARK: *k* is the number of parameters of the distribution function, calculated from the random sample. The condition for the use of chi-square test and the number *m* of intervals is that np_j must be equal or higher than 5 (i.e. $np_j \geq 5$) for every *j*.

Example 8.

The hypothesis from Example 7, "the 50 specimens come from the population with the parameters μ = 42.3 and σ = 1.50" will now be tested by the chi-square test. The input values are the same as in the previous example.

Solution. The range of possible concentrations (38.0 – 45.0) was divided into 13 subintervals of width 0.5 each, and the frequencies of occurrence in each were calculated, similarly to the above example. However, chi-square test may be used only if the number of values in every subinterval is equal or higher than 5. Therefore, new subintervals (7 altogether) were created by merging of some of them. The following table contains all data. Column 1 shows the concentrations, column 2 shows the corresponding "measured" numbers n_j of specimens, column 3 shows the theoretical probabilities p_j of concentrations, calculated via the values of distribution function, column 4 shows the theoretical numbers of values np_j for n = 50, column 5 gives the differences $n_j - np_j$ and column 6 shows the partial values $Z_j = (n_j - np_j)^2/np_j$ for the chi-square criterion.

x_j	n_j	p_j	np_j	$n_j - np_j$	Z_j
38.51-40.50	7	0.1151	5.755	1.245	0.2693
40.51-41.50	9	0.1818	9.090	–0.090	0.0009
41.51-42.00	5	0.1238	6.190	–1.190	0.2288
42.01-42.50	6	0.1323	6.615	–0.615	0.0572
42.51-43.00	9	0.1266	6.330	2.670	1.1261
43.01-43.50	5	0.1085	5.425	–0.425	0.0333
43.51-45.00	9	0.1760	8.800	0.200	0.0045

The resultant value of the criterion, given by Equation (8.19), is $\chi^2 = 1.720$. This is much lower than the critical value $\chi^2_{0.05}(4) = 9.488$, corresponding to confidence level $\alpha = 0.05$ and the number of degrees of freedom $\nu = m - p - 1 = 4$; m is the number of subintervals (8.7) and p (=2) is the number of parameters of the investigated distribution [10, 11]. Therefore, we can consider the tested sample of 50 specimens as being from the population with parameters $\mu = 42.3$ and $\sigma = 1.50$; similarly to the conclusion from the Kolmogorov-Smirnov test.

Bayesian methods

This term denotes probabilistic methods, which enable combination of information on some

event or quantity with previous information from measurement or experience. The use of additional information can increase reliability of our information, or reduce the extent of measurements needed for making conclusions on certain event.

Bayesian methods are based on the so-called **Bayes theorem** [7, 12, 13]. Let us assume that an event (B) can occur if another event (A) has occured. The event A, however, could occur by several ways (A_1, A_2, … A_n), which are mutually exclusive. The probability of simultaneous occurence of both events A_j and B, is calculated as:

$$P(BA_j) = P(A_j) \times P(B|A_j) \qquad (8.20)$$

$P(A_j)$ is the probability of event A_j, and $P(B|A_j)$ is (conditional) probability that event B can occur provided that event A_j has happened. The total probability of event B is

$$P(B) = \Sigma P(BA_j) \qquad (8.21)$$

the summation is done for all possible cases $j = 1, 2, \dots n$. Bayes theorem looks at the issue in the opposite way: *„If event B has happened, what is the probability that it was as a consequence of (or after) event* A_j?“ With the use of Eqs. (8.20, 8.21) and the fact that $P(BA_j) = P(A_jB)$, this probability can be expressed as [7, 12, 13]:

$$P(A_j|B) = P(A_j) \times P(B|A_j) \,/\, P(B) \qquad (8.22)$$

the total probability $P(B)$ in the denominator is calculated from individual probabilities via Eqs. (8.21) and (8.20). Equation (8.22) is the simplest form of **Bayes theorem**. Its use will be shown on the following example.

Increasing the reliability of non-destructive testing.

Welded components are tested for the occurrence of defects (cracks). The device used for non-destructive testing is not perfect. It classifies a defect correctly (as defect) only with probability 98%, while in 2% of all cases it does not recognise the crack and classifies the component as good. On the other hand, the device marks 96% of good parts as good, but 4% classifies erroneously as with a crack. According to long term inspection records, 3% of all tested components contain cracks. The questions are: If the tested part was classified as „wrong“ (i.e. with a defect), what is the probability that it is actually: a) wrong, b) good? And what about if the component was classified as „good“?

Solution. Event A_1: component contains a defect, A_2: component is good. $P(A_1) = 0.03$; $P(A_2) = 0.97$. Event B: component is classified as wrong. $P(B|A_1) = 0.98$; $P(B|A_2) = 0.04$. The fraction of tested components marked as wrong: $P(B) = 0.03\times0.98 + 0.97\times0.04 = 0.0682$.

Case 1a. Probability that the component marked as wrong is actually wrong, is $P(A_1|B) = P(A_1)\times P(B|A_1)/P(B) = 0.03\times0.98/0.0682 = 0.431 = 43.1\%$. Case 1b. Probability that the part marked as wrong, is actually good, is $P(A_2|B) = 0.97\times0.04/0.0682 = 0.569 = 56.9\%$. [Due to high proportion of good parts (98%), also the proportion of good, but rejected parts, is high. It may be useful to test the rejected parts once more, in order to reduce the total losses.]

Event B': component is classified as good. $P(B'|A_1) = 0.02$; $P(B'|A_2) = 0.96$. The total fraction of components, denoted as good, is: $P(B') = 0.03\times0.02 + 0.97\times0.96 = 0.9318$.

Case 2a. Probability that the component marked as good is actually wrong, is $P(A_1|B') = 0.03\times0.02/0.9318 = 0.00064 \approx 0.06\%$. Case 2b. Probability that the component marked as good is actually good, is $P(A_2|B')$ = 0.99936 ≈ 99.94%. Similar approach can be used in medicine, for example in screening for cancer.

Another example of combination of various kinds of information is the

Improvement of the estimate of parameters of normal distribution

Mean value μ and standard deviation σ of a population with normal distribution are usually unknown, so that they are replaced by their estimates m and s from a sample of size n. The estimate of the mean value can be refined via confidence interval (8.1). The estimate can be made more accurate if additional information is available, for example estimates of m_0 and s_0 from previous measurements or records. If the number n_0 of these values is known, and if the assumption can be made that all samples (new and old) belong to the same population, the updated average m_u can be calculated as the weighted average of both sample averages,

$$m_\mathrm{u} = (nm + n_0 m_0) / n_\mathrm{u} \quad ; \quad n_\mathrm{u} = n + n_0 \tag{8.23}$$

$n_{\mathbf{u}}$ is the updated number of values. The updated standard deviation is

$$s_u = \sqrt{\frac{(n-1)s^2 + (n_0-1)s_0^{\;2} + nm^2 + n_0 m_0^{\;2} - n_u m_u^{\;2}}{n_u - 1}} \tag{8.24}$$

Then, the updated confidence interval for μ can be calculated with m, s and n in (8.1) replaced by the updated values m_u, s_u a n_u. If $n_{\mathbf{0}}$ is unknown, the literature [14, 15] recommends the formula:

$$n_0 = s^2 / s_0^{\;2} \tag{8.25}$$

based on the idea that m_0 a s_0 carry information corresponding to a fictitious sample of certain size n_0. The smaller the variance $s_0^{\;2}$ compared to s^2, the more important are the original results, and the larger is the size of the fictitious sample.

More on Bayesian methods can be found in [7, 12 – 15] and in references quoted therein.

Acknowledgment. Parts of this chapter were previously published in Chapter 22 of Ref. [7].

References to Chapter 8

1. Freund, J. E.: Modern elementary statistics. Prentice-Hall, Inc., Englewood Cliffs, New Jersey,1981(6th edition). 561 p.
2. Freund, J. E., Perles, B. E.: Modern elementary statistics. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 2006 (12th edition). 576 p.
3. Suhir, E.: Applied Probability for Engineers and Scientists. McGraw-Hill, New York, 1997. 593 p.
4. Montgomery, D. C., and Runger, G. C.: Applied Statistics and Probability for Engineers. John Wiley, New York, 2006 (4th edition). 784 p.
5. ČSN 010250. Statistical methods in industrial practice. General principles. (In Czech: Statistické metody v průmyslové praxi.) ÚNM, Praha, 1972.
6. Bednařík, J. et al.: Reliability techniques in electronic practice. (In Czech: Technika spolehlivosti v elektronické praxi.) Praha, SNTL, 1990. 336 p.
7. Menčík, J.: Concise reliability for engineers. InTech, Rijeka, 2016, 204 p.
   ISBN 978-953-51-2278-4, *An Open Access publication*, *available* at:
   http://www.intechopen.com/books/concise-reliability-for-engineers.
8. Felix, M., Bláha, K.: Statistical methods in chemical industry. (In Czech: Matematickostatistické metody v chem. průmyslu.) SNTL, Praha, 1962, 336 p.
9. Jílek, M.: Statistical tolerance limits. (In Czech: Statistické toleranční meze.) SNTL, Praha, 1988. 280 p.
10. Likeš, J., Laga, J.: Statistical tables. (In Czech: Základní statistické tabulky.) SNTL, Praha, 1978. 564 p.
11. ČSN 010253. Statistical methods in industrial practice III. Basic distribution-free methods. (In Czech: Statistické metody v průmyslové praxi. Základní neparametrické metody.) Vydavatelství Úřadu pro normalizaci a měření, Praha, 1974. 114 p.
12. Martz, H. F., Waller, R. A.: Bayesian Reliability Analysis. John Wiley, New York, 1982. 745 p.
13. Press, S. J.: Bayesian statistics. Principles, models and applications. John Wiley & Sons, New York, 1989. 256 p.
14. Holický, M., Marková, J.: Principles of reliability theory and risk evaluation. (In Czech: Základy teorie spolehlivosti a hodnocení rizik.) ČVUT, Praha, 2005. 115 p.
15. Ang, A. H. S., Tang, W. H.: Probability Concepts in Engineering Planning and Design. Vol. 1, Basic Principles. John Wiley, New York, 1975. 574 p.

9. Dimensional Analysis and Theory of Similarity

Dimensional analysis and theory of similarity are powerful tools that significantly increase the efficiency of experimental research. Theory of similarity is very useful for effective creation of models and work with them. This is especially important if large structures should be studied or if experimentation with real objects or systems is impossible or very difficult. Dimensional analysis and the use of dimensionless variables simplify the experiments, can spare a lot of experimental work and make the results more general. This chapter defines various kinds of similarity, gives examples of dimensionless quantities and shows how they can be created. This is illustrated on practical problems. Also limitations of the principle of similarity are shown.

Dimensional analysis

Every physical quantity is described by a numerical value accompanied by a unit. The numerical value says how many times the considered quantity is larger than its unit. An example of length is 5.3 m, example of force is 25 N, of time is 15.6 ms. In addition to the fundamental units (meter, kilogram, second…), defined in the Système International (SI), also various derived units are used, as well as prefixes (μ, m, k, M…) denoting the order.

Every equation, describing a physical phenomenon, must be dimensionally homogeneous: its left side must have the same dimension as the right side. The check of this homogeneity should always be done before the first use of a newly derived formula. Such check also helps in formulating a correct relationship among the variables. Consider, for example, a formula for the deflection y of an elastic beam loaded by a force F. It is known from mechanics of materials that y will be directly proportional to F and indirectly proportional to the bending stiffness of the beam, defined as $E\times J$, where E is the Young modulus of the material and J is the moment of inertia of the cross section. The deflection will also be proportional to some power S of the beam length L. Now, imagine that we do not know the exponent S. In such case we could write the basic form of the formula:

$$y = C \times F \times L^S/(E \times J) \tag{9.1}$$

C is a non-dimensional constant. Replacement of the individual quantities in Eq. (9.1) by their units gives

$$\mathrm{m} = 1 \times \mathrm{N} \times \mathrm{m}^S/(\mathrm{Nm}^{-2} \times \mathrm{m}^4)$$

The dimension of the right side must be the same as that of the left side, i.e. meter, or, generally, m^1. The product of all terms containing m is $\mathrm{m}^S \times \mathrm{m}^2 \times \mathrm{m}^{-4} = \mathrm{m}^{S+2-4} = \mathrm{m}^{S-2}$. Comparison of the exponents on the left and right side of the equation gives $1 = S - 2$. From this it follows $S = 3$, so that $y = C \times F \times L^3/(EJ)$, a formula well known from mechanics.

If one side of an equation is created by a sum of several terms, then they all must have the same dimension. For example, vertical movement y of a body falling in gravitational field is described as

$$y = y_0 + v_0 t + ½\, g t^2 \tag{9.2}$$

t is time, y_0 and v_0 are the position and velocity of the body at $t = 0$, and g is the acceleration of gravity. The dimensional homogeneity demands that the individual quantities cannot exist in the physical equation independently, but only in groups of the same dimension. If Equation (9.2) is divided by one of the terms, for example y_0, it changes to non-dimensional form

$$y/y_0 = 1 + v_0 t/y_0 + ½\, g t^2/y_0 \tag{9.3}$$

with normalised quantities y/y_0, $v_0 t/y_0$ and gt^2/y_0.

Nearly every physical equation can be transformed to non-dimensional form. The use of normalised quantities has many advantages. Physical equations, expressed by means of non-dimensional variables, are more general than if they are expressed by dimensional quantities. The relative displacement, y/y_0, does not depend simply on v_0, t and y_0, but only on their certain combinations, shown in Eq. (9.3). Dimensionless quantities thus enable one to combine the results of experiments made with specimens of various initial velocity and position, the only condition being their proper combination. (In the above case of a beam, combination of its size and material play a role.) Therefore, more data and a wider range of parameters can be used for the formulation of a certain law (see Figure 9.2 later). The results expressed in non-dimensional form are also more universal, valid for the whole class of similar objects, with similar geometry or physical properties. Moreover – and this is very important – the use of non-dimensional quantities can spare experimental work, because

The relationship of *N* quantities, whose dimensions can be expressed by means of *D* basic dimensions, may usually **be replaced by a relationship of only**

$$\boldsymbol{P = N - D} \tag{9.4}$$

dimensionless parameters *Π*.

According to this, so-called ***Buckingham theorem*** [1 – 4], the determination of fewer regression constants needs fewer experiments. The reduction of experimental work is significant especially if the investigated relationship contains many quantities and if the number of variables, N, is closer to the number of basic dimensions, D. This can be illustrated on the previous example of falling body. Equation (9.2) represents relationship of 5 quantities: y, y_0, v_0, g and t; that is $N = 5$. These quantities can be expressed by means of two basic dimensions: meter and second; thus $D = 2$. According to Eq. (9.4), the number of non-dimensional parameters should be $P = N - D = 5 - 2 = 3$. And really, Equation (9.3) is the relationship of 3 dimensionless parameters: y/y_0, v_0t/y_0 and gt^2/y_0. The determination of the necessary number of experiments will be discussed in Chapter 11. Nevertheless, an idea can be obtained from a simple example. If the influence of six factors should be investigated, with each on two levels (low and high), the number of necessary experiments would be $2^6 = 64$. If the number of dimensionless factors would be only 4, the number of necessary experiments drops to $2^4 = 16$, i.e. to 25%!

Similarity

The use of non-dimensional quantities is also of prime importance in the study of behaviour of real objects by means of models. For example, building of a new large ship, a bridge, or a chemical reactor is accompanied with many uncertainties, and the potential losses due to wrong design would be very high. Therefore, usually a smaller model is built first and tested. However, if the model should adequately reflect the behaviour of the actual structure, similarity between them must exist. There are various kinds of similarity, for example:

Geometric similarity, which means identity of shape, equality of corresponding angles, and a constant proportionality between the corresponding dimensions (so-called scale factor). The following relation holds:

Model dimension = Scale factor × Dimension of the real object

For example, a model of a building, made in the scale 1:20, has all dimensions 20x smaller than the real building.

Static similarity means that the relative deformations of a model under constant stress is in the same proportion as the corresponding deformations of the object.

Kinematic similarity is based on the ratio of the time proportionality between corresponding events in the model and the object.

Dynamic similarity exists if the forces acting at corresponding times and locations in the model and object are in a fixed ratio.

Thermal similarity means that the temperature profiles in the model and the prototype must be geometrically similar at corresponding times.

Chemical similarity means that the rate of a chemical reaction in the model is proportional to the rate of the same reaction at the corresponding time and location in the object.

The **theory of similarity** works with so-called **similarity numbers**. Those, who have attended a college course of physics, know, for example, the Reynolds number (Re), which helps in assessing whether a flow of a liquid will be laminar or turbulent. More examples will be given at the end of this chapter. The similarity numbers are dimensionless; in fact, every non-dimensional quantity can serve as a similarity number.

Dimensionless variables can be created in various ways. The simplest case is the ratio of some quantity to its characteristic value, for example x/x_0 or $\Delta x/x_0$ for distance or displacement. Well known in mechanics are: strain, defined as relative elongation ($\varepsilon = \Delta L/L$), Poisson number μ (the ratio of relative shortening in transverse direction to the relative elongation in the direction of stress action), or coefficient of friction f, defined as the ratio of the force, needed to slide a body along another body, and the normal force pressing both bodies together. Another example is relative position of a point in a body, for example

$$\xi = (x - x_{\min})/(x_{\max} - x_{\min}) \qquad (9.5)$$

x, $x_{\max}$ and $x_{\min}$ represent the coordinates. Similarly it is possible to express time. Non-dimensional temperature, $\theta = (T - T_\infty)/(T_0 - T_\infty)$, is used for universal description of processes of heat transfer (T_0 is the initial temperature and T_∞ is the final temperature). In this case also the position of the investigated place and the time can be in non-dimensional form. Illustration of this approach follows.

Example. Determination of elastic modulus of thin coatings by instrumented indentation.

Modulus of elasticity E of various materials can be determined (among other methods) by instrumented indentation. An indenter is pressed into the specimen, and its displacement is measured during loading and unloading as the function of load. The elastic modulus is then determined by special processing of the measured data [5, 6]. The determination of elastic modulus of a coating, deposited on a substrate, is more complex. The response of the coated sample to indenter penetration, and thus the E value, obtained in a test, depends on the modulus of the coating (E_c) and the substrate (E_s), on the coating thickness (t) and on the depth h of indenter penetration into the specimen. The apparent E value gradually changes from the value of the coating (for "zero" indenter penetration) to the substrate modulus for very large depths of penetration (Fig. 9.1). Note that silicon (Si) has

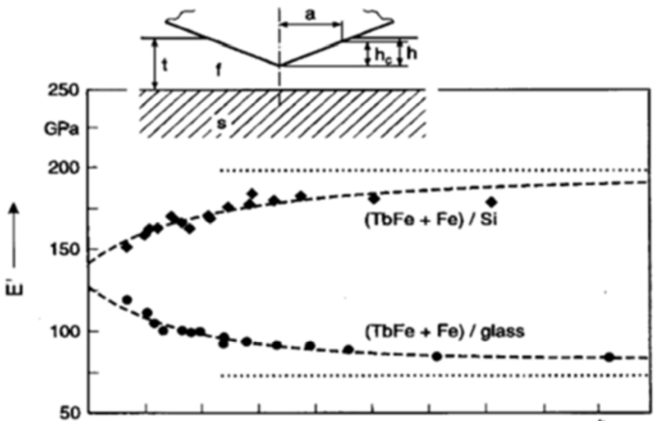


Figure 9.1. *Measurement of elastic modulus of coated samples [8]. The coating of TbTe/Fe was deposited on (a) silicone substrate and (b) glass substrate. Dotted horizontal lines correspond to the substrate moduli (Si, glass).*

higher modulus of elasticity than the coating, while glass has lower modulus.

The genuine value of the coating modulus E_c can be obtained by fitting several apparent E-values, measured for various depths, by certain function Φ and extrapolating them to zero depth of penetration. This non-dimensional function Φ can be defined as [7]

$$\Phi(h/h_c) = [E(h/t) - E_s] / [E_c - E_s] \qquad (9.6)$$

The importance of non-dimensional notation is demonstrated in Figure 9.2. This diagram shows the values measured on 25 specimens with various coating materials and thicknesses and various substrates and depths of indenter penetration [7]. One can see that all values plotted in standardised coordinates lie approximately on the curve, based on the theoretical solution of the contact [9].

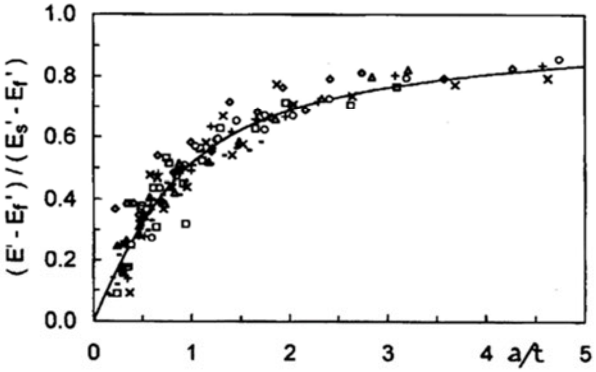


Figure 9.2. *Measurement of elastic modulus of coated samples; after [7]. The apparent values, measured for 25 various coatings, substrates and depths, are plotted in standardised coordinates. a – contact radius, t – film thickness, s – substrate, f – film (coating), E´– reduced modulus = E/(1 –* μ^2*).*

Dimensionless must also be the arguments in mathematical functions of type sin, cos, ln or exp. Otherwise any change of the units would change the numerical value of the result. Non-dimensional are also the arguments in continuous probability distributions. For example, normal distribution uses the argument $\{½[(x - \mu)/\sigma]^2\}$, where μ and σ are the mean value and standard deviation, respectively. However, the term in square brackets is nothing else than standardised variable, which expresses the distance of x from the mean value μ as the multiple of standard deviation σ. Similarly, the arguments in Weibull or exponential distribution are dimensionless. (NOTE: Non–dimensional quantities are used more often than we realise!)

A table at the end of this chapter shows examples of dimensionless quantities from various branches of physics – as inspiration for creation of parameters in other cases.

The following paragraphs, based on the works [1, 2], explain a formal procedure for creation of non-dimensional parameters and give further advice.

Creation of non-dimensional parameters

The steps are shown on the above case of movement of a body in gravitational field.

1. All quantities and their dimensions are written down:

$$y(\mathrm{m}),\ y_0(\mathrm{m}),\ v_0\ (\mathrm{m}\times\mathrm{s}^{-1}),\ t(\mathrm{s}),\ g(\mathrm{m}\times\mathrm{s}^{-2});\quad N = 5,\ D = 2$$

basic dimensions in this case are m and s.

2. Non-dimensional parameters (Π) will be assumed in general form:

$$\Pi = y^{x1} \times y_0^{x2} \times v_0^{x3} \times t^{x4} \times g^{x5}$$

3. The left and right side will be expressed by means of dimensions of the participating quantities:

$$1 = [\mathrm{m}]^0 \times [\mathrm{s}]^0 = [\mathrm{m}]^{x1} \times [\mathrm{m}]^{x2} \times [\mathrm{m}\times\mathrm{s}^{-1}]^{x3} \times [\mathrm{s}]^{x4} \times [\mathrm{m}\times\mathrm{s}^{-2}]^{x5}$$

4. The equality of both sides demands the equality of the exponents at the same bases. We shall here use the arrangement usual for systems of equations:

$$\textit{meter}: \quad x_1 + x_2 + x_3 \quad x_5 = 0 \qquad \text{(a)}$$

$$\textit{second}: \quad -x_3 + x_4 - 2x_5 = 0 \qquad \text{(b)}$$

These are two linear equations with 5 unknowns. If 5 equations were available instead of two (see above), the unknown values $x_1,\ldots x_5$ would be obtained directly by solving the system of five equations. In our case there are 3 more unknowns than the equations for their determination; $N - D = 5 - 2 = 3$. We thus can propose

3 exponents and calculate the remaining two. Such choice can be done three-times. In this example, we shall propose the values of x_1, x_3 and x_5, and want that they are as simple as possible. Therefore, one of these chosen constants will always be equal 1, and the remaining will be 0. It is reasonable if one of these exponents pertains to the variable of our interest.

Choice 1. $x_1 = 1$; $x_3 = 0$; $x_5 = 0$.
Inserting these constants into (b) gives $x_4 = 0$. Inserting x_1, x_3 and x_5 into (a) gives $x_2 = -1$. The first non-dimensional parameter thus is $\Pi_1 = y^1 \times y_0^{-1} \times v_0^0 \times t^0 \times g^0 = y/y_0$.

Choice 2. $x_1 = 0$; $x_3 = 1$; $x_5 = 0$.
Inserting them into (a, b) and solving this system in similar way as above gives $x_4 = 1$ and $x_2 = -1$. The second parameter is $\Pi_2 = y^0 \times y_0^{-1} \times v_0^1 \times t^1 \times g^0 = v_0 t/y_0$.

Choice 3. $x_1 = 0$; $x_3 = 0$; $x_5 = 1$.
Inserting them into (a, b) and solving this system gives $x_4 = 2$ and $x_2 = -1$. The third parameter is $\Pi_3 = y^0 \times y_0^{-1} \times v_0^0 \times t^2 \times g^1 = gt^2/y_0$. The reader is encouraged to repeat the solutions.

Thus, the movement of the falling body can be expressed as

$$\Phi(\Pi_1, \Pi_2, \Pi_3) = \Phi(y/y_0, v_0t/y_0, gt^2/y_0) = 0, \text{ or } y/y_0 = f(v_0t/y_0, gt^2/y_0)$$

These choices would be suitable if t could be changed easily and y measured. It is also possible to choose other parameters. For example, if we could easily change y and measure the duration t of the fall, we could first define x_4, x_1 and x_5 (similarly as above), find x_2, x_3 and obtain $\Pi_1 = v_0\,t/y_0$, $\Pi_2 = y/y_0$, $\Pi_3 = y_0\,g/v_0^2$.

Further advice

1) Sometimes the form of the non-dimensional parameters does not correspond to our intentions or experimental possibilities. Generally, it is possible to create new parameters (or similarity numbers) by making a product or ratio of the original ones, or to change them by making their reciprocal or some power. As they are dimensionless, the new parameters obtained by such transformations will be dimensionless, too.

2) If several quantities of the same dimension appear in one problem, it is also possible to create non-dimensional parameters directly as their ratios. This can reduce the number of arguments, which must be determined by solution of the system of equations such as those given under point 4 above. This will be illustrated on an example of the deflection y of a beam with rectangular cross section ($w \times h$) and length L loaded by a point force P. The modulus of elasticity is E. The variables and their dimensions are: y(m), w(m), h(m), L(m), P(N), E(Nm^{-2}); that is 6 variables with 2 dimensions. The number of non-dimensional parameters needed for the description of the problem is $P = N - D = 6 - 2 = 4$. We can immediately create three parameters $\Pi_1 = y/h$, $\Pi_2 = b/h$ and $\Pi_3 = L/h$. Two quantities remain (P and E), which must be contained in the fourth parameter. With respect to their dimensions and the condition of non-dimensionality also one geometric quantity must be included in Π_4, for example h or its power. We obtain this parameter as $\Pi_4 = P/(Eh^2)$. The studied relationship can thus be written in the following non-dimensional form:

$$y/h = f\,[P/(Eh^2), L/h, w/h] \tag{9.7}$$

One should remember that for the study of relative deflection y/h are important not the individual quantities L or P, etc., but their ratios.

3) In some problems always non-dimensional quantities appear. Examples are coefficient of friction, Poisson´s number μ for lateral contraction, or angle φ (rad). These quantities automatically become arguments in the dimensionless relationships.

4) When creating dimensionless parameters, one can use the existing knowledge on the investigated or similar problem. For example, we may know that deflection of an elastic beam is directly proportional to the load and indirectly to the modulus of elasticity. Sometimes, analytical solution is known for very small or very large values of certain variable. This can help in searching for proper form of the arguments. Sometimes it is known that some quantities must appear in certain combination. This combination can be considered as a new variable, which can enable reduction of the total number of variables. Consider, for example, force acting in the contact area of two bodies. If friction should be investigated, the force F (N) and contact area A (m^2) can be replaced by contact pressure $p = F/A$ (N/m^2).

5) When preparing an experiment, it is necessary to include all quantities, which can play a role. Otherwise wrong and misleading results can be obtained. It is less dangerous to include a quantity, whose importance is uncertain (and, perhaps, it appears later that it may be omitted), than to omit a quantity, which could be found later as important. The use of dimensional analysis sometimes reveals serious shortcomings. For example, if some dimension appears only at one quantity, this quantity falls out and will not be included in any non-dimensional parameter. However, if this quantity is obviously necessary for the description of the investigated phenomenon (e.g. as dependent variable), it is necessary to add another quantity having the same dimension. This can be illustrated on a **study of wear rate of a cutting tool**. The quantities playing a role are: wear rate *w* (m/s), velocity of mutual sliding *v* (m/s) and the pressure in the contact area *p* (N/m^2). The non-dimensional parameter could be searched in the form $\Pi = w^{x1}\ x^{x2}\ p^{x3}$. We can rewrite this expression by means of the dimensions of the individual quantities (m, s, N):

$$[\mathrm{m}]^0\ [\mathrm{s}]^0\ [\mathrm{N}]^0 = [\mathrm{m}\times\mathrm{s}^{-1}]^{x1} \times [\mathrm{m}\times\mathrm{s}^{-1}]^{x2} \times [\mathrm{N}\times\mathrm{m}^{-2}]^{x3} \qquad (9.8)$$

The left side corresponds to nondimensional notation. It follows from the condition of equality of exponents at the same base, $\mathrm{N}^0 = \mathrm{N}^{x3}$, that $x_3 = 0$. But it is well known from experiments that the wear rate does depend on the contact pressure *p*, so that x_3 cannot equal 0. It is thus necessary to include one further quantity, which would also have the dimension Nm^{-2}. This could be, for example, hardness *H* (Nm^{-2}), which characterises the resistance of the material. Now, the general form of the non-dimensional parameter is

$$\Pi = w^{x1}\ v^{x2}\ p^{x3}\ H^{x4} \qquad (9.9)$$

From this expression, we can easily formulate the appropriate relationship of dimensionless parameters as $w/v = f(p/H)$, and perform a series of experiments in order to find the appropriate form of the function *f*. Nevertheless, as an exercise, we shall also find here the non-dimensional arguments by the formal procedure described above.

Expressing the left and right side of Equation (9.9) by means of dimensions of the participating quantities gives:

$$[\mathrm{m}]^0\ [\mathrm{s}]^0\ [\mathrm{N}]^0 = [\mathrm{m}\times\mathrm{s}^{-1}]^{x1} \times [\mathrm{m}\times\mathrm{s}^{-1}]^{x2} \times [\mathrm{N}\times\mathrm{m}^{-2}]^{x3} \times [\mathrm{N}\times\mathrm{m}^{-2}]^{x4}$$

The condition of equality of exponents on the left and right side leads to the following system of equations:

$$\begin{array}{lrcl} m: & x_1 + x_2 - 2x_3 - 2x_4 & = & 0 \\ s: & -x_1 - x_2 \qquad\qquad\quad & = & 0 \\ N: & x_3 + \;\; x_4 & = & 0 \end{array}$$

These three equations are not sufficient for the determination of four exponents. Generally, one could choose one exponent and then obtain the remaining three by solving the system of 3 equations. Unfortunately, in our case this is not possible, as linear relationship exists among the equations. (The first equation can be obtained as the sum of the second one and twice of the third equation.) Thus, there must be 2 dimensionless parameters, so that 2 exponents must be chosen, for example x_1 and x_3.

Choice 1. $x_1 = 1$; $x_3 = 0$.
This choice gives $x_2 = -1$ and $x_4 = 0$, so that $\Pi_1 = w/v$

Choice 2. $x_1 = 0$; $x_3 = 1$.
This choice gives $x_2 = 0$ and $x_4 = -1$, so that $\Pi_2 = p/H$

We can thus investigate the relationship $\Pi_1 = f(\Pi_2)$, that is $w/v = f(p/H)$, as above.

Limitations of similarity principle

The principle of similarity holds only under some conditions, and outside them it loses its validity [10]. A good example is the transition from elastic to elastic-plastic deformations in components from ductile materials. If the stresses are lower than the yield strength, the deformations are elastic; linear relationship exists between stresses and strains, and the similarity principle may be used. However, the relationships in the elastic-plastic region are nonlinear and the situation must be solved for various loads individually. Another case is elastic contact of two bodies. If the stresses are low and the loaded area is large, the formulae for homogeneous isotropic elastic bodies are suitable. However, if the size of the loaded volume becomes smaller, comparable with the size of the crystalline grains and other components of the microstructure, the heterogeneity cannot be neglected. Examples are concrete and other composite materials tested by nanoindentation, but also a crystal of pearlite if the indent size is comparable with the thickness of the ferrite and cementite lamellas. The nonhomogeneity is manifested by higher scatter of individual measured values. Fortunately, statistical analysis of hundreds of tests can reveal the properties of individual phases [11]. Here, the distribution of microstructural units becomes an additional parameter for the material characterization. The increasing hardness of metals with decreasing imprint size, known as indentation size effect, is partly caused by decreasing amount of dislocations that facilitate plastic flow. For very small depths of indentation, the surface roughness becomes important, as well. The measurements under very low loads can also be significantly influenced by adhesive forces, especially when testing very compliant materials, such as gels.

Also other cases exist, where the principle of similarity does not hold. Well known is the strength dependence of brittle components on the size of loaded area or volume. Brittle fracture usually starts at a pre-existing weak point, e.g. a broken crystalline grain in ceramics or a tiny scratch on the glass surface. Smaller size of the loaded area or volume means a lower probability of occurrence of a larger defect. A smaller defect can act as a starting point only at higher stress level. Therefore, very small objects are stronger. For similar reasons, also the fatigue limit of metal components increases with their decreasing size.

Generally, one must have in mind that sometimes the investigated quantity changes with the changes of a certain parameter relatively slowly, but at its certain level it can change very quickly. The relationship, describing some behaviour or process, is often valid only within certain range of parameters. If the pertinent process is described by means of non-dimensional quantities, the conditions for a transition from one mode to another are characterised by a *critical value* of some of these quantities. A well-known example is the change from laminar to turbulent flow at the critical value of Reynolds number. One must therefore always consider all possible influences, and reduce their number only after a thorough analysis.

Examples of dimensionless quantities

Material properties

E_1/E_2, H_1/H_2	ratio of elastic moduli or hardnesses; subscripts denote the components,

$E(x)/E_0$, $H(x)/H_0$ ratios as above, subscript 0 denotes the characteristic value,

H/Y, E/Y, E/H ratio of hardness and yield strength or elastic modulus,

σ/Y, σ/σ_u, Y/σ_u ratio of stress to yield strength Y, ultimate strength (σ_u), surface stress…

Geometry

x/d x – distance, depth of indenter penetration, d – characteristic length of the specimen or material (contact radius, diameter of a crystal grain, pore or fibre, specimen length, width, height or diameter, coating thickness, size of plastic zone, distance of dislocations or other material defects, distance from the specimen edge…),

$\Delta l/L$ relative displacement, relative elongation (strain ε), Δl – elongation, L – basic length,

h/R, h/t_c ratio of indenter penetration h to the tip radius R or coating thickness t_c,

h_c/h ratio of the contact depth h_c to indenter penetration h.

Forces and stresses

F/F_0 ratio of load F and force of adhesion ($F_0 = F_{ad}$) or another characteristic force,

σ/σ_m ratio of the stress σ to the nominal or mean stress or pressure σ_m.

Time

t/t_0 t_0 – characteristic time (time of load increase, relaxation time…).

The reader can find more examples.

Similarity numbers appearing often in physics and technology

Important similarity numbers were given names of prominent scientists, and are denoted by the first two letters of the pertinent name. Some examples follow.

Archimedes Ar = $gd^3\rho'(\rho - \rho')/\eta^2$; ρ, ρ' – density of liquid and the body, g – acceleration of gravity, d – characteristic dimension, η – dynamic viscosity

Biot Bi = $\alpha d/\lambda$; α – coefficient of heat transfer, d – characteristic dimension, λ – thermal conductivity of the body

Deborah De = t_r/t ; t_r – relaxation time, t – time

Euler Eu = Δp / ru^2 ; Δp – pressure difference, ρ – density, u – characteristic velocity

Fourier Fo = $a\tau/d^2$; a – thermal diffusivity, τ – time, d – characteristic dimension

Froude Fr = u^2/gd ; u – characteristic velocity, g – acceleration of gravity, d – characteristic dimension

Galilei Ga = gd^3/ν^2 ; g – acceleration of gravity, d – characteristic dimension, ν – kinematic viscosity

Grashoff Gr = $\beta\Delta Tgl^3/\nu$; β – thermal expansion of the liquid, ΔT – temperature difference, g – acceleration of gravity, d – characteristic dimension, $\nu = \eta/\rho$ = kinematic viscosity

Nusselt Nu = $\alpha d/\lambda$; α – coefficient of heat transfer, d – characteristic dimension, λ – coefficient of thermal conductivity of the surrounding medium

Péclet Pe = ud/a ; u – velocity, d – characteristic dimension, a – thermal conductivity

Prandtl Pr = ν/a ; ν – kinematic viscosity, a – thermal diffusivity

Reynolds Re = $ud\rho/\eta = ud/\nu$; u – characteristic velocity, d – characteristic dimension, ρ – density of the liquid, η – dynamic viscosity, $\nu = \eta/\rho$ = kinematic viscosity

Stanton St = $\alpha/(\lambda u)$ = Nu/(Re.Pr) ; α – coefficient of heat transfer, λ – thermal conductivity of the fluid, u – velocity of the fluid

Stokes Stk = ut/d ; u – velocity, t – relaxation time, d – characteristic dimension

Weber	We = $\rho u^2 d/\sigma$; ρ – density, u – velocity, d – characteristic dimension, σ – surface stress

References to Chapter 9

1. Kožešník, J.: The Theory of Similarity and Modeling. (In Czech: Teorie podobnosti a modelování.) Academia, Praha, 1983. 216 p.
2. Zlokarnik, M.: Scale-up in Chemical Engineering. 2nd Edition, Wiley, 2006, 296 p.
3. Szirtes, T.: Applied Dimensional Analysis and Modeling, McGraw-Hill, New York, 1997, 2nd Ed. 2007. 856 p.
4. Y.T. Cheng, Y.T., Cheng, C.M.: Scaling, dimensional analysis, and indentation measurements, Mat. Sci. Eng. R44 (2004) 91 – 149.
5. Oliver, W.C., and Pharr, G.M.: An improved technique for determining hardness and elastic modulus using load and displacement sensing indentation experiments. J. Mater. Res. 7 (1992), No. 6, 1564 – 1583.
6. Fischer-Cripps, A.C.: Nanoindentation. Springer, 2004, 282 p.
7. Menčík et al: Determination of elastic modulus of thin layers using nanoindentation. J. Mater. Res., Vol. 12 (1997), No. 9, 2475 – 2484.
8. Menčík, J., Munz, D., Quandt, E., and Ludwig, A.: Determination of elastic modulus of thin layers. Z. Metallkd. 90 (1999), No. 10, 766 – 773.
9. Gao, H., Chiu, C.H., and Lee, J.: Elastic contact versus indentation modelling of multilayered materials. Int. J. Solids Structures, 29 (1992), 2471 – 2492.
10. Menčík, J.: Limitations of Similarity Principle in Indentation Testing of Small Samples. Key Engineering Materials (Trans Tech Publications), Vol. 586 (Local Mechanical Properties IX), 2014. 47 – 50.
11. Němeček, J.: Nanoindentation based analysis of heterogeneous structural materials. Chapter 4 (p. 89 – 108) in Nanoindentation in materials science (J. Němeček, editor). Intech, Rijeka, 2012. Open access, available at http://www.intechopen.com/books

With the help of physical theories, we try to find the way through the maze of observed facts and to understand the world.

Albert Einstein

Physical laws should have mathematical beauty.

Paul Dirac

10. Analysis of Variance (ANOVA)

A frequent problem in research is to evaluate the influence of various factors, to find, which factor has the strongest influence, which one has negligible influence, etc. If the influence of only two factors should be compared, *t*-test for the difference between means (Chapter 8) is suitable. Sometimes, however, it is necessary to evaluate the effect of three or more **factors**. For example, various raw materials, various technological procedures and various apparatuses can be used in the production of a chemical compound, and we want to know, which of these factors have stronger influence on the product. It would be possible to test separately the differences between the individual pairs of factors. However, more efficient is so-called **analysis of variance** (ANOVA), which well be briefly explained.

If the results of a certain group of tests can be sorted according to one or more criteria, then also the total variability can be sorted with respect to these criteria. The basic idea of the analysis of variance is to decompose the total variance σ_{tot}^2 of the investigated quantity into the parts caused by the individual factors (σ_{fj}^2) and a residual part σ_{res}^2 caused by unidentified (random) influences:

$$\sigma_{\mathrm{tot}}^2 = \sigma_{\mathrm{f1}}^2 + \sigma_{\mathrm{f2}}^2 + \ldots + \sigma_{\mathrm{res}}^2 \qquad (10.1)$$

Comparison of the variances corresponding to the individual factors with the residual variance caused by random influences can reveal whether the former are really due to the effect of the pertinent factors, or if they are only random.

The total variability of the results of experiments can be represented by the sum of squared differences between the individual observations and the total average of all values. The influence of the individual factors can be represented by the squared differences between the average effect of the pertinent factor and the total average. Then, residual variance remains, which is based on the squared differences between the individual observations and the averages for the individual factors.

(NOTE: the variance is obtained as the sum of squared differences divided by the number of degrees of freedom.)

Analysis of variance is based on testing of hypotheses. It tests the null hypothesis: "there is no significant difference among the influences of individual factors", which also means "all measured values come from the same population", and tests it by comparing various variances. Sample variances have chi-square distribution, and the ratio of two quantities with chi-square distribution has *F*-distribution. Therefore, *F*-tests are used to check the influence of the individual factors.

The procedure can be explained on a **one-way analysis of variance** [1]. The formulae for the necessary calculations are summarised in Table 1.

TABLE 1.

Source of variation	*Sum of squares*	*Degrees of freedom*	*Average variance*	*F*
Factor	$S_{\mathrm{fj}} = n\Sigma(y_{\mathrm{j}}. - y..)^2$	$(p - 1)$	$s_{\mathrm{fj}}^2 = S_{\mathrm{fj}}/(p - 1)$	$s_{\mathrm{fj}}^2/s_{\mathrm{res}}^2$
Residual	$S_{\mathrm{res}} = \Sigma(y_{\mathrm{jk}} - y_{\mathrm{i}}.)^2$	$(N - p)$	$s_{\mathrm{res}}^2 = S_{\mathrm{res}}/(p - 1)$	
Total	$S_{\mathrm{tot}} = \Sigma(y_{\mathrm{jk}} - y..)^2$	$(N - 1)$		

The individual symbols have the following meaning:

S_{tot} – total sum of squared differences between the individual values and the total average.
S_{fj} – sum of squared differences between the individual values and the average corresponding to the *j*-th investigated factor.
S_{res} – residual sum of squared differences between the individual measured values and the average values of the groups corresponding to the individual factors.
N – number of all tests (or observations)
n – number of observations (or tests) for the individual factors
p – number of factors
s_{fj}^2 – average variance of the factor j
s_{res}^2 – average residual variance
y_{jk} – k-th value of the j-th factor
$y_{\mathrm{j}}.$ – average of the values of j-th factor
$y..$ – total average of all values

ν – number of degrees of freedom (equal the number of the values for the determination of the characteristic minus the number of the regression constants used in this determination).

If no significant difference exists between the variances, the test criterion has *F*-distribution. If the *F*-value, calculated from the measured data, is higher than the α–critical value of *F*-distribution, an event has happened that was expected only with very low probability α, and we conclude that the difference is not random – the null hypothesis is rejected on the significance level α. Otherwise we conclude that the difference among the individual factors is not substantial. The application will be illustrated on the following example.

Example. It is necessary to find whether the kind of motor oil has influence on the fuel consumption. Three oil brands (A, B, C) were compared, each tested in five vehicles. The individual fuel consumptions and the average values (all given in *l*/100 km) were as follows:

Oil brand	Average consumption					Average
A:	7.7	8.1	7.1	7.6	8.0	$y_{jA} = 7.7$
B:	7.0	5.8	7.4	6.6	7.0	$y_{jB} = 6.8$
C:	7.6	8.5	8.2	8.0	7.7	$y_{jC} = 8.0$

The total average was $y_{..} = 7.5$ *l*/100 km. These values were inserted into the formulae of Table 1, together with $p = 3$, $n = 5$, $N = pn = 3\times5 = 15$. The results are in Table 2 below.

TABLE 2.

Source of variation	Sum of squares	Deg. of freedom	Average variance	*F*
Factor	$S_{f1} = 4.1853$	$3 - 1 = 2$	$s^2{}_{f1} = 2.0927$	9.5410
Residual	$S_{res} = 2.6320$	$15 - 3 = 12$	$s_{res}^2 = 0.2193$	
Total	$S_{tot} = 6.8173$	$15 - 1 = 14$		

The value of the test criterion is $F = s_{f1}^2/s_{res}^2 = 9.54$. This is much more than the critical value of *F*-distribution for the reliability level $\alpha = 5\%$ and degrees of freedom $\nu_1 = 3 - 1 = 2$, $\nu_2 = 15 - 3 = 12$, which is $F(0.05; 2; 12) = 3.885$. The calculated value $F = 9.54$ is even higher than the critical value for level $\alpha = 1\%$: $F(0.01; 2; 12) = 6.927$. Therefore, we can be nearly sure that the kind of the oil has influence on the fuel consumption. (The reader can repeat the procedure.)

The comparison of the calculated value with the critical value, corresponding to some level of confidence, is a classical approach, developed at the time when only tables of critical values of some distributions (e.g. F or t), corresponding to certain probabilities (e.g. 5%), were available. Today, universal programs (including Excel) can calculate the values of distribution functions of many distributions. And the distribution function gives the probabilities of non-exceeding. One can thus directly determine the probability that the differences among the individual factors are significant. For $F = 9.54$ this probability is $0.9967 \approx 99.7\%$; therefore, the probability that the measured differences were only random, is $\alpha = 1 - 0.9967 = 0.0033 \approx 0.3\%$ (see the Excel function F.DIST(F;ν_1;ν_2;TRUE) for $F = 9.54$, $\nu_1 = 2$, $\nu_2 = 12$). REMARK: α–quantile equals $(1 - \alpha)$-critical value.

Moreover, universal statistical programs enable direct application of the analysis of variance. Only the input data (e.g. the measured fuel consumptions for the oils A, B, C) must be known. The pertinent programs perform all necessary calculations (including the determination of the degrees of freedom) and give the resultant value of F and critical F-value for the chosen confidence level α, together with the probability P that the influence of the factor is insignificant. For example, Excel offers several kinds of the analysis of variance; they are available in the menu Data, submenu Data analysis. The above problem of three oils belongs to the category "Anova: One factor". It is sufficient to write the measured oil consumptions into an array of 3 rows × 5 columns (3 oil brands and 5 tests for each), put this array into the pertinent input cell in the menu, mark the command "merge the rows", and to demand the confidence level of the test. After pressing ENTER, the results are shown in two tables. The first table gives the number of values in the individual compared cases (i.e. in the rows), their sums, averages and variances. The second table gives all important values mentioned above; the sum of squares is denoted SS and average variances are denoted MS. The reader is encouraged to solve this example with the same input data and to compare the own results with those in this chapter – for better understanding and practice.

This was a simple problem, with only one factor, just for illustration. Analysis of variance can be used for various problems, with sorting according to two, three or even more factors. More can be found in textbooks on statistics, e.g. [1 – 5]. Brief explanations are also available via Help command in computer programs for statistical analysis.

References to Chapter 10

1. Felix, M., Bláha, K.: Statistical methods in chemical industry. (In Czech: Matematickostatistické metody v chemickém průmyslu.) SNTL, Praha, 1962. 336 p.
2. Freund, J. E.: Modern elementary statistics. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1981(6th edition). 561 p.
3. Freund, J. E., Perles, B. E.: Modern elementary statistics. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 2006 (12th edition). 576 p.
4. Suhir, E.: Applied Probability for Engineers and Scientists. McGraw-Hill, New York, 1997. 593 p.
5. Montgomery, D. C., and Runger, G. C.: Applied Statistics and Probability for Engineers. John Wiley, New York, 2006 (4th edition). 784 p.

The supreme judge of every physical theory is an experiment.

Lev Landau

Chance serves those, who are prepared.

Louis Pasteur

11. Design of Experiments (DOE)

Any experimental investigation starts with preliminary experiments to obtain better insight into the problem in question. Much of the work at this initial stage is mainly intuitive and aims at better defining the problem. As soon as the goals of the investigation have been defined, the next step is to reduce the large number of possible variables to the several most important ones. Statistical analysis is useful, because it can help in choosing from all possible models. As the experimentation moves into the optimisation stage, statistical design of experiments is again effective in finding the optimum parameters. This chapter shows how experiments can be organised efficiently so that the demanded information is obtained with minimum effort. Important terms, such as blocks, randomisation and Latin squares are explained, as well as the principal rules and tables for design of experiments. Their use is illustrated on examples.

The **variables**, which play a role in the experiments, can be classified as quantitative, qualitative, or binary. ***Quantitative response****,* which is measured by a continuous scale, is the most common and easiest to work with in statistical analysis. ***Qualitative response****,* like glitter or odour, can be ranked on an ordinal scale, for example from 0 for the worst alternative to 10 for the best one. ***Binary response*** produces one of two values, e.g. pass or fail, go or no-go, men or women.

Factors are experimental variables controlled by the investigator. An important part of planning an experimental program is the identification of the important variables that affect the response, and deciding how to exploit them in the experiments. The scientific model of the problem is examined for important variables. Previous experience is very useful. We often take the advantage of dimensional analysis in establishing the factors (see Chapter 9).

Factors may be independent in the sense that the level of one factor is independent of the level of other factors. However, two or more factors may interact with one another. This means that the effect of one variable on the response depends on the levels of the other variables.

Random sampling

In every experimental program using a large number of tests or measurements made on individual specimens it is important that any of the specimens involved in the experiment has the same chance of being selected for the given test. This, so-called random sampling, can be achieved by **randomisation**. One way to randomize a batch of specimens is to assign a number to each specimen, put the numbered tags into a jar, mix them, and then withdraw them like in a lottery. (Today, computer-generated random numbers can be used instead.) In this way, bias caused by uncontrolled second-order variables is minimised (e.g. that due to subtle changes in the characteristics of the testing equipment or in the proficiency of the operator). If metal specimens are taken from large forgings, the possibility of the variation of properties with the position in the forging must be considered. If average properties of the entire forging should be determined, randomisation of the specimens positions will minimize the bias due to the position in the forging. In addition to the primary variables that are under control of the experimenter there are other variables which may not be under control. Examples are small differences in the way different operators run an experiment or carry out a test, or differences in humidity or other environmental factors. These effects can be reduced using blocking design, described further.

Block design, Latin squares

In order to increase the reliability of the conclusions, the experiment is often repeated several times. However, a frequent problem in these experiments is to maintain the identical conditions. Often, one batch of a homogeneous raw material is sufficient only for one series of tests and another batch must be used for another test series. The properties of individual batches often vary. If this fact is neglected and all measured values are evaluated together, the results will exhibit bigger variance due to the combination of the natural scatter and the differences of properties among the batches. This drawback can be reduced or eliminated if the experiments are designed so that the differences among the repetitions are separated from the differences due to various batches. (See the chapter Analysis of variance and [1 – 5].) The experiments are divided into groups with approximately the same conditions.

The word **block** denotes a set of conditions that creates a homogeneous entity (from the experimental point of view). A block can be raw material from one batch, compared with the materials from other batches. A block can also be experiments done in the same chemical reactor under the same thermal conditions, analyses made by the same technician, or samples taken from a continuous process during a short time.

In order to avoid systematic errors that could be caused by the same order of procedures in every block, it is recommended to carry out the individual experiments in random order, as described in the previous section. We speak about *randomized blocks*. For example, if we want to study the influence of temperature on the result of a chemical reaction, and if this investigation should be based on the raw materials from four batches (B_1, B_2, B_3, B_4) and four temperatures (T_1, T_2, T_3, T_4) for every batch, the following random arrangement can be used:

Batch	*Order of temperatures in each batch*			
B_1	T_1	T_3	T_4	T_2
B_2	T_3	T_1	T_4	T_2
B_3	T_4	T_2	T_3	T_1
B_4	T_1	T_4	T_2	T_3

This arrangement resembles so-called **Latin squares** [1 – 5]. In such experiments, the same number of variants for every factor is used. An example of a Latin square "4 × 4" for three factors is below. The rows are assigned to the levels of the first factor, the columns are assigned to the second factor, and the letters *A*, *B*, *C*, *D* are assigned to the third factor. The creation of such system (rotation of the letters for the third factor) is obvious from the table.

Rows	*Columns*			
	1	2	3	4
1	*A*	*B*	*C*	*D*
2	*B*	*C*	*D*	*A*
3	*C*	*D*	*A*	*B*
4	*D*	*A*	*B*	*C*

Factorial experiments

The necessary extent of experiments increases with the number of factors that play a role. If we want to determine whether the factor x has an influence on the quantity w, and how strong this influence is, we must make (at least) two experiments, i.e. for two values of x. If the influence of two factors should be revealed, we must make at least three or four experiments. This number will grow significantly with an increasing number of factors. The use of the rules for **design of experiments**, or DOE, can make the process more effective.

DOE means creation of schemes with such combinations of input quantities, which generate the required information with minimum experimental effort. This can be illustrated for the case of three independent variables x, y, z and a dependent variable w. The independent quantities are usually called **factors**, and their individual values (**levels**) are denoted by subscripts, e.g. 1 for the lower level and 2 for higher level if two levels are used. For example, the dependent variable corresponding to the experiment with the factor x on the lower level, with y on the upper level and with z on the upper level, is denoted $w_{1,2,2}$.

Also other notations are used to denote the arrangement of input quantities, for example –1 for lower level and +1 for upper level. If three levels are used, the subscript 0 is used for the intermediate level. Symbols + and – are also used and then arranged into tables for various numbers of factors (see at the end of this chapter).

REMARK: The situation with three factors is used here for illustration, as it can be imagined easily in our “3-D” space. The described procedures can be extended for more factors.

The simplest arrangement for three factors is depicted in Fig. 11.1. Four experiments are made, with the following combinations: w_{111}, w_{211}, w_{121}, w_{112}. The influence of x on w is obtained as the difference of values $w_{211} - w_{111}$, the influence of y is obtained as $w_{121} - w_{111}$, and the influence of z is $w_{112} - w_{111}$. In these cases, always the influence of only one quantity is investigated, whereas the remaining quantities keep their original levels. If we want to obtain information on the variance, we must repeat the experiments at least twice; that is we have to make at least eight experiments.

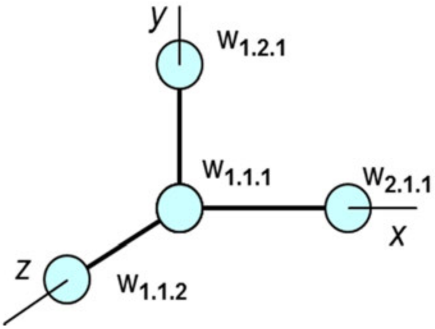


Figure 11.1. *Simple experiment with three independent factors (x, y, z), each on two levels.*

More information is obtained from a **full factorial experiment**. The quantity of our interest, *w*, is determined for all possible combinations of all factors and levels (Fig. 11.2). The necessary number of experiments is, generally,

$$n_{\mathrm{exp}} = (n_{\mathrm{levels}})^{N\mathrm{factors}} \qquad (11.1)$$

if the number of levels n_{levels} is the same for every factor. For three factors (N_{factors} = 3), each at two levels, the number of experiments is $n_{\mathrm{exp}} = 2^3 = 8$, with the combinations w_{111}, w_{211}, w_{121}, w_{112}, w_{221}, w_{212}, w_{122}, w_{222} (see Fig. 11.2 and also the table at the end of this chapter). If various numbers of levels are used for the individual factors, the number of experiments is

$$n_{\mathrm{exp}} = n_{\mathrm{level\ 1}} \times n_{\mathrm{level\ 2}} \times n_{\mathrm{level\ 3}} \times \ldots \qquad 11.(2)$$

where $n_{\mathrm{level,j}}$ denotes the number of levels for the *j*-th factor.

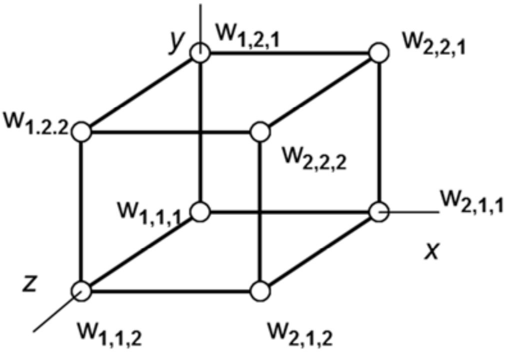


Figure 11.2. *Full factorial experiment with three independent factors.*

An advantage of full factorial experiments is that always the results of all experiments are used for ascertaining the influence of any factor. This increases the accuracy of results. For example, the average influence of factor *x* in our case with three factors is obtained by summing its effects for various values of factors *y* and *z* and dividing by four (Fig. 11.2):

$$U_x = [(w_{221}–w_{121}) + (w_{211}–w_{111}) + (w_{222}–w_{122}) + (w_{212}–w_{112})] / 4 \qquad (11.3)$$

The result can be rewritten:

$$U_x = [(w_{221} + w_{211} + w_{222} + w_{212}) – (w_{121} + w_{111} + w_{122} + w_{112})] /4 \qquad (11.4)$$

The term in the first brackets is the sum of all results obtained with *x* on the upper level, and the term in the second brackets is the sum of all results obtained with *x* on the lower level. Equation (11.4) can also be rewritten as

$$U_x = (w_{221} + w_{211} + w_{222} + w_{212} – w_{121} – w_{111} – w_{122} – w_{112}) / 4 \qquad (11.5)$$

Similarly the influences of *y* and *z* are obtained as:

$$U_y = (w_{121} + w_{221} + w_{122} + w_{222} – w_{111} – w_{211} – w_{112} – w_{212}) / 4 \qquad (11.6)$$

$$U_z = (w_{122} + w_{222} + w_{112} + w_{212} – w_{121} – w_{221} – w_{111} – w_{211}) / 4 \qquad (11.7)$$

Another advantage of factorial experiments is the possibility of revealing interactions from the same experiments. **Interaction** means that the influence of a certain factor, say *x*, depends also of the values of factor *y* or *z*, or both. The situation is schematically depicted in Fig. 11.3 with curves for various values of *z*; the left illustration is without interaction, and the right one corresponds to the *x*–*z* interaction. If, in our experiment with three factors, the interaction among the factors *x* and *z* should be revealed, the solution is as follows:

1) The effect of *x* at one level of *z* is subtracted from the effect of *x* at the second level of *z*: [(w_{212}–w_{112}) minus (w_{211}– w_{111})].

2) As the influence of *y* is not considered, similar effects must be added for the second level of *y*: [(w_{222}–w_{122}) minus (w_{221}– w_{121})]. The result must again be divided by four:

$$U_{xz} = [(w_{212}–w_{112}) – (w_{211}– w_{111}) + (w_{222}–w_{122}) – (w_{221}– w_{121})] / 4 \qquad (11.8)$$

Expression (11.8) can be rewritten as:

$$U_{xz} = (w_{212} + w_{222} + w_{111} + w_{121} - w_{211} - w_{221} - w_{112} - w_{122}) / 4 \qquad (11.9)$$

The effects of other interactions could be obtained in a similar way.

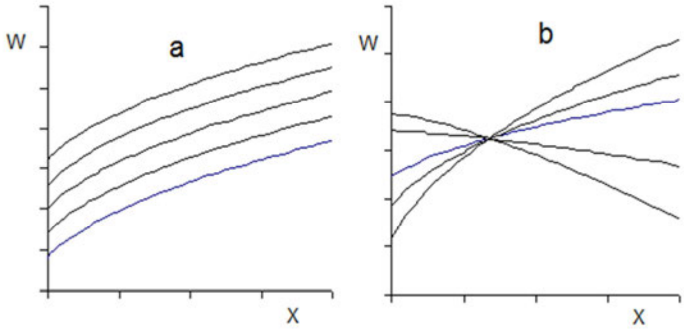


Figure 11.3. Experiments without interaction (a) and with interaction (b) of some factors. The individual curves correspond to various values of factor z.

A practical illustration of design of experiments (DOE) follows.

Example. It is necessary to reveal the cause of creation of surface cracks on steel springs during quenching. The three most influential factors were: temperature of steel before quenching (Ts), temperature of oil bath (To), and carbon content in the steel (C).

For quantitative characterisation of their influence, a full factorial experiment was proposed, with each factor on two levels:

Level	*Ts (°C)*	*To (°C)*	*C (%)*
Low (–)	830	70	0.5
High (+)	910	120	0.7

The number of experiments is $2^3 = 8$. The combinations of the levels and the corresponding numbers of cracks *N* found on the hardened springs are given in the table on the next page (cf. also Fig. 11. 2; *x* corresponds to Ts, *y* corresponds to To, and *z* corresponds to C; + corresponds to higher level and – corresponds to lower level):

Test No.	Ts	To	C	Ts(°C)	To(°C)	C(%)	N
1	–	–	–	830	70	0.5	67
2	+	–	–	910	70	0.5	79
3	–	+	–	830	120	0.5	59
4	+	+	–	830	120	0.5	90
5	–	–	+	830	70	0.7	61
6	+	–	+	910	70	0.7	75
7	–	+	+	830	120	0.7	52
8	+	+	+	910	120	0.7	87

The average influences of steel temperature (Us), oil temperature (Uo) and carbon content (Uc), calculated via Equations (11.5) – (11.7), are:

Us = (79+90+75+87–67–59–61–52)/4 = 23.0

Uo = (59+90+52+87–67–79–61–75)/4 = 1.5

Uc = (52+87+61+75–59–90–67–79)/4 = –5.0

The steel temperature has the strongest influence; the influence of carbon content (in the range 0.5 ÷ 0.7 %) is small and the influence of oil temperature (in the range 70°C ÷ 120°C) is negligible.

The interaction of steel temperature and carbon content, following from Equation (11.9), is

Usc = [(75+87+67+59)–(79+90+61+52)]/4 = 1.5

i.e. also negligible. The other interactions can be found in a similar way.

Very informative is a graphical representation (Figure 11.4). The horizontal axis represents the average of all values, i.e. (67+79+59+90+61+75+52+87)/8 = 71.25. The influence of steel temperature is depicted by two points at the left: one, giving the average number of cracks in the cases where the temperature was on the lower level, i.e. (59+67+52+61)/4 = 59.75, and the other, corresponding to the higher temperature, (90+79+87+75)/4 = 82.75. For better visibility they are connected by a straight line. NOTE: 82.75 – 59.75 = 23.0 = Us. Figure 11.4 also depicts the influence of oil temperature and carbon content.

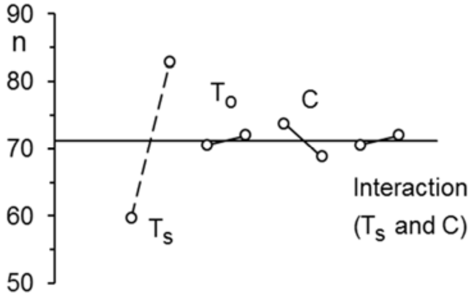


Figure 11.4. *Influence of individual factors on the quality of springs.*

Sometimes it is necessary to investigate the influence of a higher number of factors. The following table gives the levels of individual factors for full-factorial experiments with two, three and four factors, each on two levels. Four tests are necessary for two factors (columns A,B), eight tests for three factors (A,B,C), and 16 tests for four factors (A,B,C,D). The table can be easily extended for more factors if one looks how pluses and minuses vary at the individual factors, beginning from A.

Test	A	B	C	D .
1	–	–	–	–
2	+	–	–	–
3	–	+	–	–
4	+	+	–	–
5	–	–	+	–
6	+	–	+	–
7	–	+	+	–
8	+	+	+	–
9	–	–	–	+
10	+	–	–	+
11	–	+	–	+
12	+	+	–	+
13	–	–	+	+
14	+	–	+	+
15	–	+	+	+
16	+	+	+	+

In industrial research, for example in optimisation of manufacturing conditions, more factors often play a role, and the number of experiments for a full-factorial experiment would be very high (e.g. 256 experiments for 8 factors, each on two levels). Here, **reduced factorial experiments** are often used, where some combinations of levels are omitted. For this purpose, special schemes (so-called orthogonal arrays) have been developed. This topic goes beyond the scope of this book and the reader is referred to the books on design of experiments, and robust design and the relevant methods developed by Genichi Taguchi and other authors [6 – 9]. For design of experiments (DOE) in general, an engineering handbook [10] and a comprehensive monograph [11] can be recommended.

References to Chapter 11

1. Freund, J. E.: Modern elementary statistics. Prentice-Hall, Inc., Englewood Cliffs, New Jersey,1981(6th edition). 561 p.
2. Freund, J. E., Perles, B. E.: Modern elementary statistics. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 2006 (12th edition). 576 p.
3. Suhir, E.: Applied Probability for Engineers and Scientists. McGraw-Hill, New York, 1997. 593 p.
4. Montgomery, D. C., Runger, G. C.: Applied Statistics and Probability for Engineers. John Wiley, New York, 2006 (4th edition). 784 p.
5. Felix, M., Bláha, K.: Statistical methods in chemical industry. (In Czech: Matematickostatistické metody v chemickém průmyslu.) SNTL, Praha, 1962. 336 p.
6. Taguchi, G.: Introduction to Quality Engineering: Designing Quality into Products and Processes. Asian Productivity Organization, 1986. 191 p.
7. Ross, P. J.: Taguchi Techniques for Quality Engineering. McGraw-Hill, New York, 1996. 329 p.
8. Taguchi, G., Chowdhury, S., Wu, Y.: Taguchi´s Quality Engineering Handbook. John Wiley & Sons, Hoboken, New Jersey, 2005. 1804 p.
9. Fowlkes, W. Y., Creveling, C. M.: Engineering methods for robust product design. (Using Taguchi methods in technology and product development.) Addison-Wesley, New York, 1995. 432 p.
10. Dieter, G. E.: Engineering design. 2nd edition. McGraw-Hill, New York, 1991. 721 p.
11. Montgomery, D. C.: Design and analysis of experiments. 8th edition, Wiley, 2012. 730 p.

12. Experimental Finding of Maximum or Minimum

A usual problem in research is finding a maximum or minimum of some quantity, for example certain parameter of a machine or a chemical compound, efficiency, or costs. This chapter shows how an extreme of a function (maximum or minimum) can be found in cases when the analytical form of this function is not known and its values can be obtained only by experiments or by computer modelling for concrete values of the input quantities. An intuitive method of successive changes of the input values is explained first, then the method of the steepest gradient, and the simplex method, which is very efficient for cases with several independent variables. The procedures will be illustrated on an example of search for a maximum. Also the methods of simulated annealing or genetic algorithms are explained.

A. Gradual changes of the individual variables

The experimental search for a maximum of a function of one variable, $y = f(x)$, is very easy (Figure 12.1). We start with two experiments, for values x_0 and x_1. If the value $y(x_1)$ was higher than $y(x_0)$, we make the next experiment with x changed in

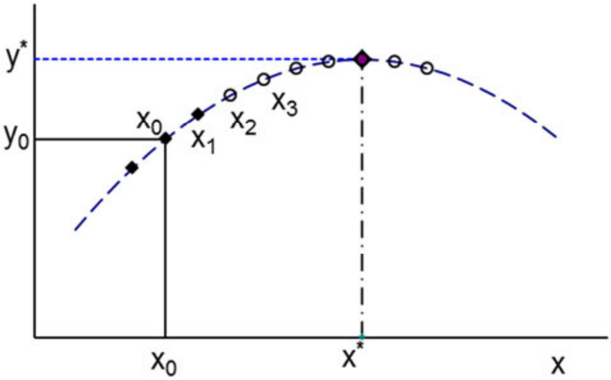


Figure 12.1. *Search for a maximum of a function of one independent variable.*

the same direction; $x_2 = x_1 + \Delta x$, where Δx is a suitably chosen increment. In this way we proceed until y starts decreasing; the maximum lies approximately at the value of the preceding step or near it (Fig. 12.1). A more accurate position of the maximum can be found by making more experiments here.

If we look for a minimum, we move in the opposite direction.

With **two independent variables** we assume that the function $z(x, y)$ can be approximated by a polynomial, at least in the vicinity of a chosen starting point x_0, y_0. Several experiments are made, in which only one variable, say, x, is changed, while y keeps its initial value. In this way we proceed in the direction of increasing z until z starts decreasing. The preceding step corresponded to the local maximum of z. Now, we move from this point in perpendicular direction and change only y and proceed until the local maximum of z is attained, and so on. The situation is depicted in Figure 12.2.

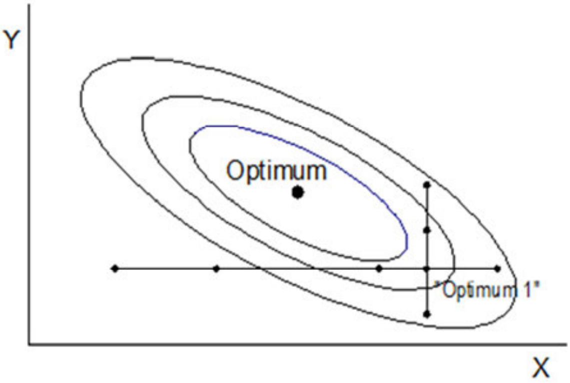


Figure 12.2. *Experimental search for a maximum of a function of two independent variables x and y.*

If the approximate position of the maximum (or minimum) is roughly known, it is also possible to make several experiments around this point, for example 4 to 8 for two independent variables, and fit the obtained values by a response surface (Fig. 13.1 in Chapter 13); a second order polynomial is often sufficient. The accurate position of its extreme can then be found using standard mathematical methods or a suitable solver. More on this topic can be found in Chapter 13 and in the literature, recommended there.

B. Gradient method

This method tries to approach to the maximum (or minimum) in the fastest way, which is in the direction of the gradient to the response surface [1–3]. This gradient must be found first, as it will be shown here for two independent factors. Several experiments are made around a suitably chosen point x_0, y_0. The z (x, y) values can be fitted by a polynomial, for example

$$z = a_0 + a_1x + a_2x^2 + a_3x^3 + \ldots + b_1y + b_2y^2 + \ldots \qquad (12.1)$$

The gradient vector is obtained generally by means of partial derivatives,

$$\boldsymbol{grad\ z} = \partial z/\partial x\ \mathbf{i} + \partial z/\partial y\ \mathbf{j} + \ldots \qquad (12.2)$$

i and **j** are the unit vectors in directions x and y, respectively. For the polynomial (12.1) the gradient is

$$\boldsymbol{grad\ z} = (a_1 + 2a_2x + 3a_3x^2 + \ldots)\ \mathbf{i} + (b_1 + 2b_2y + \ldots)\ \mathbf{j} \qquad (12.3)$$

In a small vicinity of the point x_0, y_0, often a first-degree polynomial (tangential plane) is sufficient,

$$z = a_0 + a_1x + b_1y \qquad (12.4)$$

with the gradient

$$\boldsymbol{grad\ z} = a_1\ \mathbf{i} + b_1\ \mathbf{j} \qquad (12.5)$$

Now, we proceed in this direction towards the local maximum, with steps proportional to a_1 in direction x and simultaneously to b_1 in direction y, until the values of z start decreasing. Again several experiments are made around this point, the direction of fastest growth is found, etc. The application of this approach for three and more variables is similar.

C. Simplex method

This is a simple method, in which the input variables approach to the optimum stepwise according to an algorithm proposed by Spendley et al. [4]. The knowledge of gradient is not necessary. In the first step, a **simplex** is created. This is a simple fictitious convex body with n + 1 vortices; this number is by one higher than the number n of input variables (for example, a triangle for two independent variables and a tetrahedron for three input variables). The coordinates of the vortices correspond to the values of input parameters. For all these points, the output quantity *z* is calculated. In the next step, a new simplex is created by replacing the vortex with the worst value of *z* by a new one, whose coordinates are mirror-symmetrical. (In a two-dimensional space, the new simplex is obtained by skipping the original one over the edge opposite to the worst vortex, Fig. 12.3a.) For this new point, the dependent variable is calculated. Now, the values of the dependent variable for all vortices of the new simplex are compared, and again the worst vortex is omitted and the new one is created in the same way. In this manner we proceed until the quantity of interest attains the extreme or acceptable value. The reaching of optimum is usually indicated by the oscillation of the simplex between two positions, or by the movement of the simplex bodies along the closed curve (Fig. 12.3b).

REMARK: A thoughtful reader will notice the similarity between this method and Figure 1 for one independent variable.

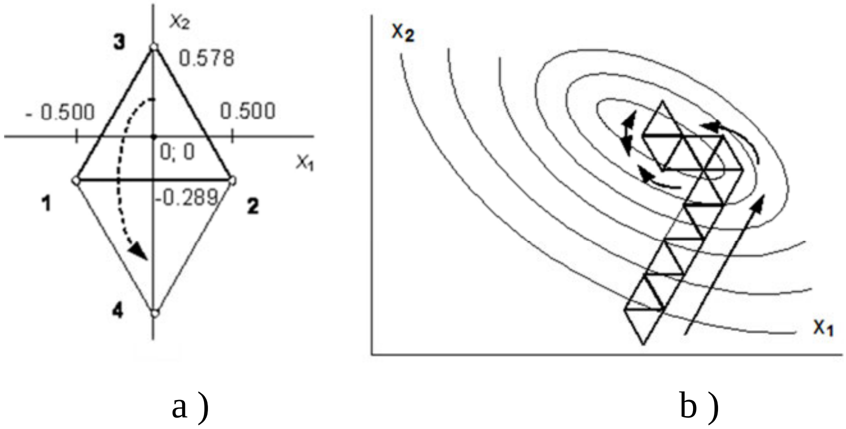


Figure 12.3. *Simplex method. a) Non-dimensional simplex 1-2-3 for two independent variables (x_1, x_2) and one optimisation step (creation of vortex 4); b) movement towards the optimum.*

The practical procedure is as follows [4, 5]. First, the coordinates $x_{i,0}$ of the starting point (= the centroid of the simplex) are chosen, as well as the increments of individual variables Δx_i. (The first subscript denotes the variable; *i* = 1, 2,… *n*, where *n* is the number of independent variables; the second subscript denotes the step.) Then, the coordinates of the individual n+1 vortices of the first simplex are calculated by means of the associated non-dimensional regular simplex as [4, 5]:

$$x_{i,1} = x_{i,0} + z_i \, \Delta x_i \qquad (12.6)$$

z_i denotes the radius of the sphere inscribed (r_i) or circumscribed (R_i) to the associated simplex. The centroid of this simplex is in the centre of the coordinate system, and the pertinent radii are calculated using the following matrix and formulae:

$$\begin{bmatrix} -r_1 & -r_2 & -r_3 & \ldots\ldots & -r_{n-1} & -r_n \\ R_1 & -r_2 & -r_3 & \ldots\ldots & -r_{n-1} & -r_n \\ 0 & R_2 & -r_3 & \ldots\ldots & -r_{n-1} & -r_n \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots\ldots & R_{n-1} & R_n \\ 0 & 0 & 0 & \ldots\ldots & 0 & R_n \end{bmatrix}, \quad r_i = \frac{1}{\sqrt{2i(i+1)}} \quad ; \quad R_i = \sqrt{\frac{i}{2(i+1)}} \qquad (12.7)$$

The column denotes the variable, and the row denotes the vortex number; the matrix has n columns and n + 1 rows. The response of the structure is then calculated for all n + 1 combinations of input values. The x_i coordinate of the new, generally (j + 1)-st vortex, is then determined as

$$x_{i,j+1} = \frac{2}{n} \sum_{i=1}^{n} x_{i,j} - x_{i,j}^{*} \qquad (12.8)$$

The first subscript denotes the variable (x_1, x_2, …), the second subscript denotes the vortex numbers. $x_{\mathrm{i,j}}^{*}$ is the coordinate of the point with the worst value of the optimisation criterion (y), and ($\Sigma x_{\mathrm{i,j}}$)/n is the average of the coordinates of all vortices (of the j-th simplex) except the worst one. In this way, all n coordinates of the new vortex are obtained.

The procedure is best shown on an example. Let us have n = 2 independent variables, x_1 and x_2. (For example, x_1 is the width of a cross section of a load carrying component, and x_2 is its height). Let the coordinates of the starting point be $x_{1,0}$ = 200 mm, $x_{2,0}$ = 500 mm, and their increments in the optimisation steps let be Δx_1 = 20 mm, Δx_2 = 30 mm. (Generally, the individual variables can have different dimensions.) The non-dimensional matrix (12.7) in this case has n = 2 columns and n + 1 = 3 rows:

$$\begin{bmatrix} -r_1 & -r_2 \\ R_1 & -r_2 \\ 0 & R_2 \end{bmatrix} = \begin{bmatrix} -0.500 & -0.289 \\ 0.500 & -0.289 \\ 0 & 0.577 \end{bmatrix}$$

The left (right) column expresses the non-dimensional coordinates for x_1 (x_2), the rows pertain to the first, second and third vortices. The r_j and R_j values were calculated using the formulae at the right side of Equation (12.7).

The coordinates of the three vortices of the first simplex are (with respect to Figure 12.3a) as follows: the x_1 value of the first vortex is $x_{1,1} = x_{1,0} - 0.500 \times \Delta x_1 = 200 - 0.500 \times 20 = 190$ mm, the x_1 value of the second vortex is $x_{1,2} = x_{1,0} + 0.500 \times \Delta x_1 = 200 + 0.500 \times 20 = 210$ mm, and the third vortex is $x_{1,3} = x_{1,0} + 0 \times \Delta x_1 = 200 + 0 \times 20 = 200$ mm. The corresponding values of x_2 are: $x_{2,1} = x_{2,0} - 0.289 \times \Delta x_2 = 500 - 0.289 \times 30 = 491.33$ mm, $x_{2,2} = x_{2,0} - 0.289 \times \Delta x_2 = 500 - 0.289 \times 30 = 491.33$ mm, and $x_{2,3} = x_{2,0} + 0.578 \times \Delta x_2 = 500 + 0.578 \times 30 = 517.34$ mm. All values are arranged in the table:

vortex	x_1	x_2
1	190	491.3
2	210	491.3
3	200	517.3

Now, if the worst value of the optimisation criterion belonged, e.g., to the vortex No. 3, then the coordinates of the new vortex (No. 4) would be $x_{1,4} = 2(190 + 210)/2 - 200 = 200$ mm, and $x_{2,4} = 2(491.3 + 491.3)/2 - 517.3 = 465.3$ mm. The new simplex is defined by the vortices 1, 3 and 4 [see also Fig. 33a; here the non-dimensional coordinates of vortex 4 are 0 and –0.289 – (0.289 + 0.578) = –1.156. Then, the value of optimisation criterion (y) at this vortex is computed, the y values for the points 1, 2 and 4 are compared and the coordinates of the vortex 5 are calculated, and so on, until the optimisation criterion stops growing.

The described simplex method has several advantages. The algorithm is very simple and the coordinates of the new vortex are calculated directly from those of the previous simplex. No gradients must be determined. One gets closer to the optimum in every step. The method is suitable also for higher number of variables.

As the construction of a new vortex is based not on the exact values of the dependent variable at the individual vortices, but on their comparison, there are no excessive demands on the accuracy of results in this stage of optimisation. If constraints exist for some of the variables, and the coordinates of a new vortex would move outside the allowable limits, it is possible to omit not the vortex with the worst value, but with the second (or third) worst one.

D. Further methods

Also other methods exist for finding the extreme of a function. They are more demanding and need a computer and a suitable program. Here, only two will be mentioned: simulated annealing and genetic algorithms.

Simulated annealing. The methods, described here until now, tried to approach the maximum stepwise so that in each step the values of the independent variables were changed in the direction of the increase of the investigated quantity. A drawback is that this procedure can find only local maximum. Sometimes, more local extremes can exist, and the task usually is to find the global maximum.

REMARK. The search for minimum is analogous. The term simulated annealing is used in analogy with heat treatment (annealing), in which slow controlled cooling of a hot body leads to the state with minimum internal energy and defects. Simulated annealing proceeds in steps. In contrast to the gradient methods, described before, this method enables random search in various directions, and accepts (with some probability) even worse solutions than current ones. Thanks to this approach one can (in the following steps) get out from the local minimum and consider also other possible solutions. Gradually the solution approaches to the global maximum (or minimum) in the investigated region.

The method of simulated annealing needs a suitable computer program, such as Simulated Annealing Solver [6], which is a part of Global Optimization Toolbox within the universal computing tool Matlab. The details can be found at http://www.mathworks.com,. The first information on the method can be found, e.g., in Wikipedia and sources quoted there.

Genetic algorithms solve optimisation problems by mimicking principles of biologic evolution. Such algorithm generates various solutions of the investigated problem. It works with so-called population, every member of which constitutes one solution of the problem. The solution is usually represented by binary numbers, i.e. an array of 0 and 1, but also other representations are used, for example with matrices. At the beginning of optimisation process (the first generation) the population consists of totally random members. Then, several individuals are chosen from it, which are then modified by mutations and crossover. (Mutation means random change of a part of the individual, and crossover means mutual exchange of parts of several individuals.) In this way, a new population arises. After this creation of new generation, so-called fitness function is calculated for each individual, which characterises its ability and thus the quality of the solution. Further selection and modification is done with respect to the obtained value of the fitness function. This procedure is repeated, so that the quality in the population gets gradually better („upgrading“). The process is usually stopped on attaining a sufficient quality of solution, or after the elapse of certain time.

Optimisation via genetic algorithms can be done, for example, using the Genetic Algorithm Solver program [7], a part of Global Optimization Toolbox of Matlab. Details can again be found at http://www.mathworks.com or, generally, in Wikipedia and sources quoted there.

References to Chapter 12

1. Felix, M., and Bláha, K.: Statistical methods in chemical industry. (In Czech: Matematickostatistické metody v chemickém průmyslu.) SNTL, Praha, 1962. 336 p.
2. Arora, J.: Introduction to optimum design (third edition), Elsevier, 2011, 896 p.
3. Myers, R. H., and Montgomery, D. C.: Response surface methodology. Wiley, 2016 (4th Edition). 856 p.
4. Spendley W., Hext, G. R. and Himsworth, F. R.: Sequential Application of Simplex Designs in Optimisation and Evolutionary Operation *Technometrics*, Vol. 4 (1962), No. 4, 441 – 461.
5. Tichomirov, V. B. Planning and analysis of experiments (in Russian: Planirovanije i analiz eksperimenta pri provedenii issledovanij v legkoj i tekstilnoj promyšlennosti. Legkaya industria, Moscow, 1974. 262 p.
6. Matlab – Simulated Annealing Solver; http://www.mathworks.com
7. Matlab – Genetic Algorithm Solver; http://www.mathworks.com

13. Sensitivity Analysis

Input quantities often vary or deviate from nominal values, which causes deviations of the investigated quantity from its nominal value. The task of sensitivity analysis is to show how these variations contribute to the deviations of the investigated quantity from its nominal or assumed value. This is important for the prediction of the response under real conditions, i.e. with some uncertainties or variations of input quantities. Sensitivity analysis can be made using analytical expressions or simulation probabilistic methods. In the former case the relationship between the output quantity z and input variables x_1, x_2, ... x_n, so-called **response function**, should be known. The exact analytical expression,

$$y = f(x_1, x_2, \dots x_n) \tag{13.1}$$

is available only for simple problems. Often, the response must be found by experiments or by time-consuming numerical solution. In such cases, an approximate expression is used, obtained by regression-fitting the response for several combinations of input variables.

The simplest form of a response function is a polynomial, for example

$$y_i \quad = \quad a_0 + a_{i1}x_i + a_{i2}x_i^{\,2} + a_{i3}x_i^{\,3} + \dots. \tag{13.2}$$

or

$$y_i \quad = \quad y_0 + a_i(x_i - x_{i,0}) + b_i(x_i - x_{i,0})^2 + \dots. \tag{13.3}$$

Equation (13.3) expresses the changes of y as a function of deviations of input variable x_i from the nominal value $x_{\mathrm{i},0}$. Subscript i denotes i-th variable, and y_i corresponds to this variable. These regression functions represent the sections through the response surface (Fig. 13.1). In the vicinity of the design point ($x_{\mathrm{i},0}$), polynomials up to the second order are often sufficient.

Polynomial, or even a linear function can also be used for the approximation of other relationships (e.g. $1/x$ or $\sqrt{x}$) if suitable transformation is made. Solvers in universal programs enable easy determination of regression coefficients in complex functions by direct use of the least squares method, without transformations.

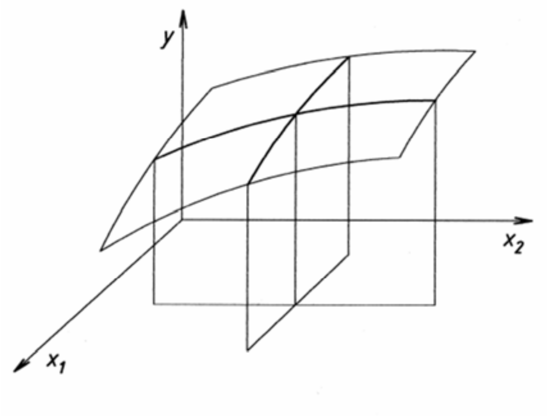


***Figure 13.1.** Response surface for 2 independent variables, with cuts x_1, x_2 = const.*

Further comments to response functions can be found in Chapter 14.

Sensitivity analysis depends on whether the deviations of individual quantities from their nominal values are considered as deterministic or random [2, 3].

Deterministic deviations of input quantities

The sensitivity of the response to the variations of individual variables is obtained from partial derivatives at the pertinent point,

$$c_i = (\partial y / \partial x_i) \tag{13.4}$$

The sensitivity coefficients c_i correspond to the constants $a_{i,1}$ in (13.2) and a_i in (13.3). Further information is obtained from relative sensitivities,

$$c_{ri} = \frac{\partial y}{\partial x_i} \frac{x_{i,0}}{y_0} \approx \frac{\Delta y}{y_0} \Big/ \frac{\Delta x_i}{x_{i,0}} \tag{13.5}$$

y_0 and $x_{i,0}$ are the values corresponding to the design point. Coefficient c_{ri} expresses the relative change of y (in %, for example) caused by 1% deviation of x_i from the nominal value $x_{i,0}$. For linear approximation, $c_{ri} = a_i \times (x_{i,0} / y_0)$.

Sensitivity analysis also reveals the input variables that have negligible or small influence on the variability of the output quantity y, and may be considered as constants in the more complex analysis with more input quantities. (Note that the variance of the output depends on the variances of the input quantities x_i and also on the sensitivities c_i !)

If the increments Δx_i are small, the response surface may be approximated by a linear expression

$$y \quad = \quad a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_n x_n \qquad (13.6)$$

which represents a plane for two independent variables ($n = 2$) and a hyperplane for more input variables. The constants correspond to the sensitivity coefficients (except a_0), and are obtained by fitting the $n + 1$ values of response by a multiple linear regression function. (Also Excel can be used for this purpose). The increments of y are calculated via the first derivatives. For $y = f(x_1, x_2, \ldots, x_n)$, the infinitesimal increment of y is generally

$$\mathrm{d}y = (\partial y/\partial x_1)\mathrm{d}x_1 + (\partial y/\partial x_2)\mathrm{d}x_2 + \ldots + (\partial y/\partial x_\mathrm{n})\mathrm{d}x_\mathrm{n} \qquad (13.7)$$

$\partial y/\partial x_\mathrm{i}$ expresses partial derivatives. In practical analysis, the differentials are replaced by small finite increments Δ,

$$\Delta y = (\partial y/\partial x_1)\Delta x_1 + (\partial y/\partial x_2)\Delta x_2 + \ldots + (\partial y/\partial x_\mathrm{n})\Delta x_\mathrm{n} \qquad (13.8)$$

The application of sensitivity analysis can be illustrated on an example [1, 3] of a small flat spring for a measuring device (Fig. 13.2). We want to know the sensitivity of its compliance to the variations of its dimensions and elastic modulus of the material. This compliance is

$$C = y/F = 4L^3/(Ewt^3) \qquad (13.9)$$

y is deflection, F – load, L – length, E – elastic modulus, w – spring width, t – spring thickness. The partial derivative of Equation (13.9) with respect to the first variable ($x_1 = L$) is

$$\partial C/\partial L = 3L^2 \times 4/(Ewt^3) = [4L^3/(Ewt^3)] \times 3/L = (3/L) \times C \qquad (13.10)$$

and the increment of compliance due to an increment of the beam length ΔL is thus

$$\Delta C = 3C\,(\Delta L/L) \qquad (13.11)$$

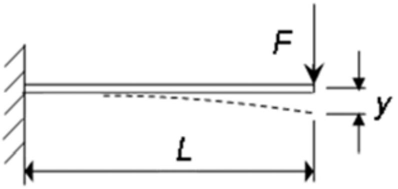


Figure 13.2. *Spring for a measuring device [1].*

The formulae for other variables are obtained in a similar way. The resultant expression, involving the changes of all variables, is

$$\Delta C = C\,(3\Delta L/L - \Delta E/E - \Delta w/w - 3\Delta t/t) \qquad (13.12)$$

The relative sensitivity

$$\Delta C/C = 3\Delta L/L - \Delta E/E - \Delta w/w - 3\Delta t/t \qquad (13.13)$$

shows illustratively the influence of the individual quantities. If the spring will be longer by 1% than the nominal value, the compliance will be higher by 3%; if the elastic modulus *E* will be higher by 1%, the compliance will be lower by 1%, etc. The constants in the individual terms correspond to their exponents in Equation (13.9), and the signs depend on whether the quantity was in the numerator or denominator.

Influence of random variations of input variables

The combined influence of random variations of input quantities can be evaluated via the expression for the variance of a function of several random variables. For small variance,

$$s_y{}^2 = \left(\frac{\partial y}{\partial x_1}\right)^2 s_{x1}{}^2 + \left(\frac{\partial y}{\partial x_2}\right)^2 s_{x2}{}^2 + \ldots + 2\left(\frac{\partial y}{\partial x_1}\right)\left(\frac{\partial y}{\partial x_2}\right) cov(x_1, x_2) + \ldots \qquad (13.14)$$

s_{xi} is the standard deviation of x_i. The far right-hand term is nonzero if the variables are correlated.

The response surface function for all factors can be written approximately as a polynomial:

$$y = y_0 + \sum a_i (x_i - x_{i,0}) + \sum b_i (x_i - x_{i,0})^2 + \sum c_i (x_i - x_{i,0})(x_j - x_{j,0}) + ... \qquad (13.15)$$

the summation is done for all independent variables. The constants are obtained by regression fitting the points around the design point. For non-correlated variables and linear approximation (6) of *y*, Equation (13.14) becomes

$$s_y{}^2 = a_1{}^2 s_{x1}{}^2 + a_2{}^2 s_{x2}{}^2 + ... + a_n{}^2 s_{xn}{}^2 + ... \qquad (13.16)$$

The individual components, $s_{yi}{}^2 = a_i{}^2 s_{xi}{}^2$, give the variances of *y* caused by random variations of *i*-th variable. The contribution of $s_{xi}{}^2$ to the total variance $s_y{}^2$ is larger for larger variance of the variable x_{i} and for larger sensitivity (a_i) of the output *y* to the changes of x_i. Division of Eq. (13.16) by $s_y{}^2$ gives the relative proportions of individual factors in the total variance,

$$1 = a_1{}^2 \frac{s_{x1}{}^2}{s_y{}^2} + a_2{}^2 \frac{s_{x2}{}^2}{s_y{}^2} + ... + a_n{}^2 \frac{s_{xn}{}^2}{s_y{}^2} + ... \qquad (13.17)$$

The influence of variance of the individual input factors can be assessed by means of the ratio of the variation coefficient of the *i*-th variable and the variation coefficient of the output, corresponding to the variance of this variable only,

$$\omega_i = \frac{v_y}{v_{xi}} = \frac{s_y}{y_0} \bigg/ \frac{s_{xi}}{x_{i,0}} \qquad (13.18)$$

Sensitivity analysis using simulation methods

The influence of random variability of input quantities can be assessed even without analytical expression for the response function – by means of the probabilistic simulation technique Monte Carlo, described in the next chapter. In this case, the sensitivity analysis consists of making *m* trials, the only random variable being x_i, and then calculating the partial variance $s_{yi}{}^2$ of the obtained values *y*. Then, using the characteristics s_{xi}, $x_{i,0}$ and y_0, and Equations (13.16) and (13.18), one can determine the ratios of variation coefficients v_i or the sensitivity coefficients a_i (= s_y/s_{xi}) and the coefficients of relative sensitivity [2].

The approximate value of the total variance is obtained by summing up the partial variances,

$$s_y{}^2 \;=\; s_{y1}{}^2 + s_{y2}{}^2 + \ldots + s_{yn}{}^2 + \ldots \qquad (13.19)$$

A more accurate value is obtained if all input variables, x_1, x_2, ... x_n, are simultaneously considered as random quantities in the Monte Carlo simulations. Dividing Equation (13.19) by the total variance $s_y{}^2$ gives the relative influence of individual factors, similarly to Eq. (13.17).

Examples of applications of uncertainty analysis for the prediction of lifetime can be found, for example, in [4, 5] and in Chapter 19 of [1].

Acknowledgment. Parts of this chapter were previously published in Chapter 19 of Ref. [1].

References to Chapter 13

1. Menčík, J.: Concise reliability for engineers. InTech, Rijeka, 2016, ISBN 978-953-51-2278-4. *An Open Access publication*, available via: http://www.intechopen.com/books/concise-reliability-for-engineers. 204 p.
2. Novák, D., Teplý, B. and Shiraishi, N. Sensitivity Analysis of Structures: A Review. *Int. Conf. CIVIL COMP"93, August 1993*: p. 201-207. Scotland, Edinburgh.
3. Menčík, J.: Reliability-based parameter optimisation and tolerancing in structural design. *8th Int. Conf. ICOSSAR 2001*. Newport Beach, USA. *Structural Safety and Reliability* (R.B. Corotis, G.I. Schuëller and M. Shinozuka, eds), A.A. Balkema Publishers, Lisse, p. 188.
4. Jacobs, D. F., Ritter, and J. E. jr.: Uncertainty in minimum lifetime predictions. Am. Ceram. Soc., 59 (1976) No. 11/12, p. 481 – 487.
5. Wiederhorn, S. M., Fuller, E. R. jr., Mandel, J., and Evans, A. G.: An error analysis of failure prediction techniques derived from fracture mechanics. J. Am. Ceram. Soc., 59 (1976) No. 9/10, p. 403 – 411.

14. Simulation Methods for Study of Random Quantities and Influences

Today, a great part of research is made by computer simulations. Probabilistic simulation methods can be used to study the influence of random variability of various quantities on the properties of machines, on chemical or biological processes, on the load carrying capacity or reliability of a structure and in many other cases. A very powerful tool for study of random phenomena or processes is the Monte Carlo technique. In some cases, useful results can be obtained with less effort using the Latin Hypercube Sampling.

Monte Carlo simulation method

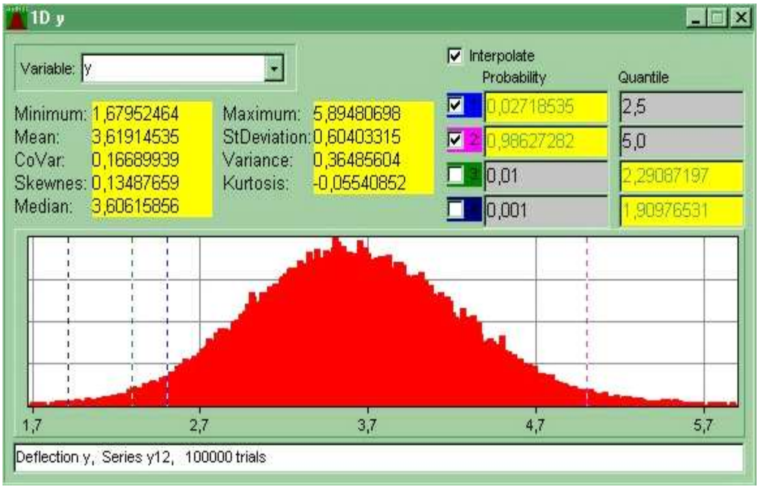
The Monte Carlo method is a simple computer technique based on performing numerous fictitious experiments with random numbers [1 - 3]. Its use is universal and does not need a special knowledge of probability theory. The only information one needs is the relationship between the output and input quantities,

$$y = f(x)\ ,\ \text{ or }\ \ y = f(x_1, x_2, x_3, \ldots) \qquad (14.1)$$

and the knowledge of probability distributions of the input variables. The method repeats trials with computer-generated random numbers processed by the relevant mathematical operations. In each ”trial“, the input variables x_1, x_2, …, x_n are assigned random values, but such that their distributions correspond to the probability distribution of each variable. With these values, the output quantity y is calculated via Equation (14.1). From the results, a histogram can be constructed (Fig. 14.1), which corresponds to the distribution of y.

The generated values can be used for the determination of the average value or of the probability that y will be lower or higher than a chosen value y^*, or for the determination of values, which will be exceeded (or not achieved) with some probability (e.g. the time to failure, maximum expectable load or deformation).

Various commercial computer programs exist for Monte Carlo simulations [5 – 7],



Variable: y

Minimum: 1,67952464
Mean: 3,61914535
CoVar: 0,16689939
Skewnes: 0,13487659
Median: 3,60615856

Maximum: 5,89480698
StDeviation: 0,60403315
Variance: 0,36485604
Kurtosis: -0,05540852

☑ Interpolate

	Probability	Quantile
☑ 1	0,02718535	2,5
☑ 2	0,98627282	5,0
□ 3	0,01	2,29087197
□ 4	0,001	1,90976531

Deflection y, Series y12, 100000 trials

Figure 14.1. *Histogram obtained by the Monte Carlo simulation program Ant-Hill [4, 5].*

but they can also be created. The base of such programs is a generator of ***random numbers***. Actually, these numbers are not truly random, but created via a suitable algorithm. The principle of these generators is simple. For example, the so-called congruential generator gives random numbers with uniform distribution in the interval (0; 1) in the following way. One number is chosen as the base for the series of random numbers u (e.g. $u_0 = 0.5284163$). Now, in the first step, this number is multiplied by some suitable number Q, for example 997. The product is $997 \times 0.5284163 = 526.8310511$. The first random number u_1 is then created as the part of the result, lying behind the decimal point; in our case, $u_1 = 0.8310511$. In the second step, u_1 is again multiplied by the same number Q, $997 \times 0.8310511 = 828.5579467$, and the second random number is created as the decimal part of the result (i.e. $u_2 = 0.5579467$). The reader is encouraged to make several steps in this way; for a check, $u_3 = 0.2728599$. A long series of these numbers has approximately uniform distribution. Many other algorithms exist; e.g. one for normal distribution is based on central limit theorem. Generators of random numbers are also a part of universal computer programs, such as Matlab. The use of commercial generators is strongly recommended, as they have undergone thorough statistical testing to prove that they behave nearly as really random. Even Excel has its own generator, though with limited possibilities.

Creation of random numbers with specified distributions

The commercial programs offer often-used distributions, for example uniform or normal. The random numbers, corresponding to other analytically defined distributions, can be generated via uniform distribution. The basic idea is that the distribution function *F* for any continuous random variable is also a random quantity, distributed uniformly in the interval (0; 1). Thus, if the distribution function of random quantity *x* is $z = F(x)$, then the random numbers *x* can be obtained from the random numbers *z* with uniform distribution in the interval (0; 1) using the inverse formula (Fig. 14.2):

$$x = F^{-1}(z) \tag{14.2}$$

For example, the distribution function for exponential distribution is $z = F(x) = 1 - \exp(-x/x_0)$, with the parameter x_0. The inverse transformation for this distribution is $x = -x_0 \ln(1 - z)$.

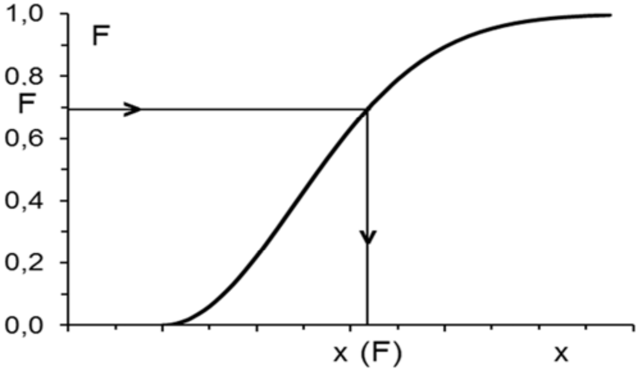


Figure 14.2. *Generation of random numbers x by inverse probabilistic transformation [4].*

In some cases, the distribution of a random quantity *x* has a complex shape and can only be described by a histogram (obtained from measurements). This histogram is then used for the construction of distribution function *F*(*x*). This function can be approximated either by constant values of *F* in the individual subintervals of *x* or by interpolation within each class,

$$F(x) = F_i + \frac{F_{i+1} - F_i}{x_{i+1} - x_i}(x - x_i)\ , \quad x = x_i + \frac{F(x) - F_i}{F_{i+1} - F_i}(x_{i+1} - x_i) \qquad (14.3)$$

$i = 1, 2, \dots, n$ denotes the interval. The formula at the right generates x corresponding to the probability F. The F values are generated as random numbers with a uniform distribution.

A typical feature of the Monte Carlo method is that the characteristic values (average, quantiles, probabilities corresponding to certain values of y, etc.), obtained as a result of n trials, are never the same in two sets of simulations. The results are thus only approximate, but they are closer to the actual values for more trials. The number of simulation trials n, needed for achieving some accuracy of results, is given approximately by the formula

$$n = u_{\alpha/2}^{\ 2}\,(1 - P)\,/\,(P\delta^2) \qquad (14.4)$$

P is the expected (estimated) probability of the investigated phenomenon, δ is the allowed relative error in the determination of P, and $u_{\alpha/2}$ is the $\alpha/2$–critical value of standard normal variable for the probability α that the actual value of P will lay outside the interval $P \pm \delta$. The necessary number of simulations significantly grows with decreasing probability. For example, if the assumed probability $P = 0.01$, the allowed relative error $\delta = 10\%$ and confidence level $\alpha = 5\%$ (with $u_{\alpha/2} = 1.96$), then ≈ 40,000 simulation trials are necessary. For $P = 0.0001$, it is as many as 4,000,000 trials, etc. [Note: Equation (14.4) is based on the fact that the number of outcomes of an event of probability P in n repetitions has binomial distribution, and this distribution can be approximated for high n by normal distribution.]

More complex cases, Response Surface Method

The direct use of the Monte Carlo method is suitable for simple relationships $y = f(x_1, x_2, \dots)$. Often, the response y must be obtained by numerical solution. If one such trial lasts minutes or more, then thousands of simulations would consume too much time. In these cases, more effective is the combination of the MC technique with the response surface method (RSM), mentioned in Chapter 13. The principle is that the "accurate" response is calculated only for selected values of input variables, the results are fitted by a simple regression function (response surface, Fig. 13.1 in Chapter 13), and the Monte Carlo trials are done with this function.

The relationship between the output quantity y (deformation, load-carrying capacity of a structure, amplitude of vibrations) and the input variables can often be fitted by a polynomial function:

$$y = a_0 + \sum a_i x_i + \sum b_i x_i^2 + \ldots + \sum c_{ij} x_i x_j + \ldots \qquad (14.5)$$

This approximation is possible if the relationship between input and output has a similar character (e.g. $y \sim x^3$) or if the output quantity changes in the considered interval only little. If it differs from a polynomial significantly (e.g. $y \sim 1/x^3$ or $y \sim x^{1/2}$), equation (14.5) cannot give a good approximation in a wider interval. Several ways for improvement exist. Linear or polynomial function may be used for the approximation of other relationships if suitable transformations are made. For example, the relationship $y = a/x^3$ can be expressed as $y = az$ by introducing a new variable $z = 1/x^3$; the relationship $y = ax_1/x_2^2$ can be converted to multiple linear regression $Y = A_0 + A_1X_1 + A_2X_2$ using logarithmic transformations, etc. The fitting of response function can sometimes be improved by dividing the definition interval of input quantities into subintervals and using different regression functions for each.

The quality of the fit can be evaluated by means of residual standard deviation s_{res}. Also, the differences between the ”accurate“ values and those on the response surface can serve as a criterion. With good response surface, these differences are randomly positive and negative. (See also the residuals and Fig. 7.2 in Chapter 7.)

Application of the Monte Carlo method for correlated quantities

The application of Monte Carlo technique to problems with several input variables is simple if the individual input quantities are mutually independent (e.g. material properties and the geometry of a component). Sometimes, however, correlation between them exists (for example between mass density and Young’s modulus of concrete). A special case is autocorrelation, when the value of a random quantity at some point is related partly to the values at neighbouring points or in preceding times. Examples are the properties of soil at foundations or the temperature of a building structure: it varies during a day or from a day to day, but depends partly also on the season in the year.

The omission of correlations can lead to errors. For example, a very low value of elastic modulus of concrete could be generated simultaneously with a very high value of strength, but this does not correspond to reality. If correlations are respected, the calculations reflect the reality better and the predictions are more accurate. Sometimes, also, a quantity needed for the analysis is unavailable, but can be replaced by a correlated quantity. For example, if the direct measurement of the tensile strength of an existing massive steel structure is impossible, the information from hardness tests can sometimes be used.

The strength of the relationship of two quantities is characterized by the correlation coefficient r, defined as

$$r = \mathrm{cov}(xy) / (s_x s_y) \qquad (14.6)$$

where cov(xy) is the covariance of x and y, and s_x and s_y are the standard deviations. The correlation coefficient r ranges from –1 to +1. For $r = 0$, no mutual relationship exists, whereas $r = +1$ or –1 corresponds to deterministic (functional) relationship. For $r > 0$, the x values grow with growing y, and decrease for $r < 0$. (NOTE: The correlation coefficient is equal to the square root of the coefficient of determination r^2, explained in Chapter 6.) Three examples with the same mean values and standard deviations and different values of r are shown in Figure 14.3.

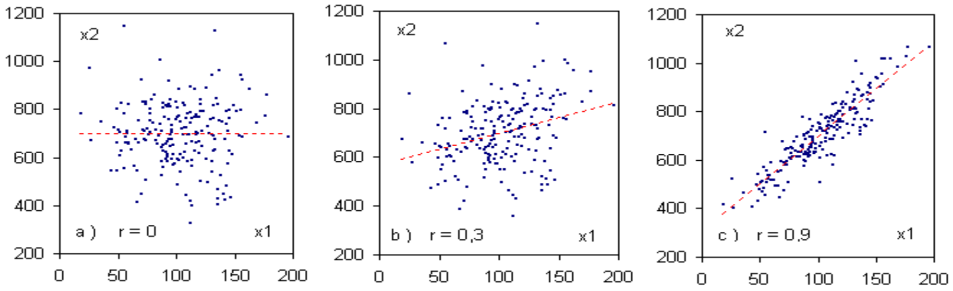


Figure 14.3. *Two correlated quantities x_1 and x_2 with the same means (μ_1 = 100 and μ_2 = 700) and standard deviations (σ_1 = 30 and σ_2 = 150) and various correlation coefficients r [8, 9].*

If two correlated random quantities x_1 and x_2 should be generated, and if the regression function $x_{2,\mathrm{reg}} = f(x_1)$ is known, as well as the coefficient of determination r^2 of this approximation, the following procedure may be used [7, 8]. First, the random value of x_1 is generated. Then, the corresponding value of x_2 is generated as

$$x_2 = f(x_1) + \Delta x_2 = f(x_1) + us_{2,\mathrm{res}} = f(x_1) + us_2(1 - r^2) \; ; \qquad (14.7)$$

$s_{2,\mathrm{res}}$ is the sample residual standard deviation of x_2 around the regression function f, and u is

the random quantile of standard normal distribution (provided that the distribution of individual values x_2 around f is normal). The right-hand part of Equation (14.7) uses the fact that the residual deviation $s_{2,\mathrm{res}}$ of x_2 can be expressed by means of the standard deviation s_2 and the coefficient of determination r^2 pertaining to the regression function $x_2 = f(x_1)$.

For more information on the Monte Carlo method, the books [1, 2] can be recommended. Many examples of practical applications can be found in [3]. Various commercial Monte Carlo programs for engineering applications exist, for example [5 – 7]; several others are mentioned in Chapter 26 of the book [4].

Latin Hypercube Sampling (LHS)

The Monte Carlo technique has two disadvantages. First, it usually needs a very high number of simulations. If the output quantity must be obtained by time-consuming numerical computations, the simulations can take a very long time. Also the response surface method is not always usable. Second, it can happen that the random numbers of distribution function F (which serve for the creation of random numbers with nonstandard distributions) are not distributed sufficiently uniformly in the definition interval (0; 1). Sometimes, more numbers are generated in one region than in others, and the generated quantity has thus a somewhat different distribution than demanded. This problem can appear especially if the output function depends on many input variables.

A method called Latin Hypercube Sampling (shortly LHS) removes this drawback [10, 11]. The basic idea of LHS is similar to the generation of random numbers with nonstandard distribution via the inverse probabilistic transformation (14.2), as shown in Fig. 14.2 above. The difference is that LHS creates the values of F not by generating random numbers dispersed in chaotic way in the interval (0; 1), but by assigning them certain fix values. The interval (0; 1) is divided into several **layers** of the same width, and the x values are calculated via the inverse transformation (F^{-1}) from the F values corresponding to the centre of each layer (Fig. 14.4). With reasonably high number of layers (tens or hundreds) the created quantity x will approximately have the proper probability distribution. This approach is called Stratified Sampling. If the output quantity y depends on several input quantities, x_1, x_2, … x_n, it is necessary that each quantity is assigned values of all layers, and that the quantities and layers of individual variables are randomly combined. This is done by random assigning the order numbers of layers to the individual input quantities.

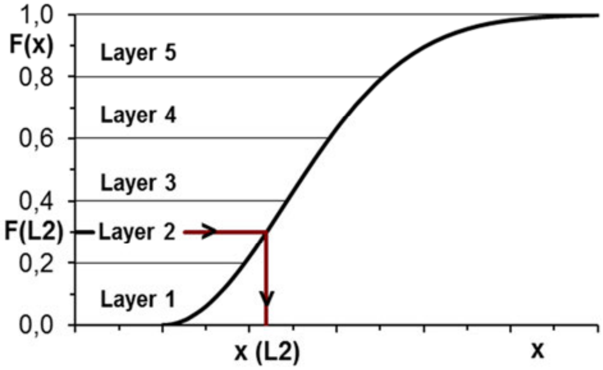


Figure 14.4. *Latin Hypercube Sampling method (LHS) – principle [4].*

The procedure is as follows. The definition interval of the distribution function F for each of m variables is divided into N layers. N, the same for all variables, also corresponds to the number of trials (= simulation experiments). In each trial, the order numbers of layers are assigned randomly to the individual variables (X_1, X_2,.., X_m). In this way, various layers of the individual variables are always randomly combined. In practice, this is achieved by means of random numbers and their rank-ordering. Then, each input variable is assigned the value corresponding to the centre of the pertinent layer of its distribution function.

The application is illustrated on a case with four random quantities (X_1, X_2, X_3, X_4) and the definition interval of F divided into 5 layers (Fig. 14.5). Five layers are used here for simplicity; usually several tens of layers are used. In our case, Y will be calculated for five combinations of the four input quantities. Thus, 5 × 4 = 20 random numbers with uniform distribution in interval (0; 1) are generated (see the table on the left part of Fig. 14.5). Then, the layer numbers for variable X_1 (for example) for individual trials are assigned with respect to the order of random values (for X_1) ranked by size from the maximum to minimum. Here, layer No. 2 (with the highest number 0.885) for the first trial, No. 3 for the second, No. 1 for the third trial, etc., corresponding to the numbers 0.382 – 0.885 – 0.863 – 0.032 – 0.285 in the column for X_1. Similar operations are done for each variable. Thus, in the first trial, variables X_1, X_2, X_3 and X_4 are assigned the values corresponding to the 3rd, 5th, 1st and 1st layer of their distribution functions, respectively. Inverse probabilistic transformation F^{-1} is then used for the determination X_1 from $F_{1,1}$, etc.; see the table on the right. Now, the investigated quantity $Y = Y(X_1, X_2, X_3, X_4)$ is calculated 5-times. The obtained values Y_1, Y_2, Y_3, Y_4, Y_5 can be used for the determination of statistical characteristics (mean, standard deviation…).

Random numbers (RN)

Variable -------------- Layer/trial	X_1 *RN*	X_2 *RN*	X_3 *RN*	X_4 *RN*
1	0.382	0.101	0.596	0.899
2	0.885	0.958	0.014	0.407
3	0.863	0.139	0.245	0.045
4	0.032	0.164	0.220	0.017
5	0.285	0.343	0.554	0.357

Layer numbers for individual layers (LN)

Variable ---------------- Layer (trial)	X_1 *LN*	X_2 *LN*	X_3 *LN*	X_4 *LN*
1	3	5	1	1
2	1	1	5	2
3	2	4	3	4
4	5	3	4	5
5	4	2	2	3

Figure 14.5. *LHS method – assignment of layers to individual variables and trials.*

Usually several tens or hundreds of trials are made, which enable construction of distribution function $F(Y)$ and determination of the mean value, standard deviation, various quantiles and other characteristics.

More on the LHS method can be found in [10, 11].

Acknowledgment. Parts of this chapter were previously published in Chapters 15 and 16 of the book [4].

References to Chapter 14

1. Hammersley, J. M., Handscomb, D.: Monte Carlo Methods. John Wiley, New York, 1964. 184 p.

2. Marek, P., Guštar, M., Anagnos, T. Simulation-based reliability assessment for structural engineers. CRC Press, Boca Raton, 1996. 384 p.
3. Marek, P., Brozetti, J., Guštar, M., Tikalsky, P., editors. Probabilistic Assessment of Structures using Monte Carlo Simulation. ITAM CAS CR, Prague, 2003. ISBN 80-86246-19-1. 471 p.
4. Menčík, J.: Concise reliability for engineers. InTech, Rijeka, 2016, *An Open Access publication*,: http://www.intechopen.com/books/concise-reliability-for-engineers, ISBN 978-953-51-2278-4. 204 p.
5. AntHill program for Monte Carlo simulations: http://www.sbra-anthill.com
6. Novák, D., Rusina, R., Vořechovský, M.: FReET. A multipurpose probabilistic software for analysis of Engineering problems. http://www.freet.cz. 2015.
7. Petschacher, M.: VaP (Variables Processor). http://www.petschacher.at. 2016.
8. Menčík, J.: Simulační posuzování spolehlivosti při korelovaných veličinách. In: Spolehlivost konstrukcí, 23. - 24. 4. 2003, Dům techniky Ostrava, s. 151 – 156.
9. Čačko, J., Bílý, M., Bukoveczky, J.: Measurement, evaluation and simulation of random processes. (In Slovak: Meranie, vyhodnocovanie a simulácia náhodných procesov.) Bratislava, VEDA, 1984. 210 p.
10. Florian, A.: An efficient sampling scheme: Updated Latin Hypercube Sampling. Probabilistic Engineering Mechanics, 7 (1992), issue 2, p. 123 – 130.
11. Olsson, A., Sandberg, G., Dahlblom, G.: On Latin hypercube sampling for structural reliability analysis. Probabilistic Engineering Mechanics. 2002; 25; issue 1, p. 47 – 68.

Index (the numbers in brackets denote the chapters)

Book title:	Introduction to Experimental Analysis
Author:	prof. Ing. Jaroslav Menčík, CSc.
Publisher:	University of Pardubice
Print:	May 2017
No. of pages:	142
No. of copies:	20
Edition:	first
Printed by:	Polygrafické středisko Univerzity Pardubice

ISBN 978-80-7560-068-4 (print)
ISBN 978-80-7560-069-1 (pdf)

The book is freely accessible on http://hdl.handle.net/10195/66961 .

ISBN 978-80-7560-069-1 (pdf)