KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

Politechnika Wrocławska

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

ROZWÓJ POTENCJAŁU I OFERTY DYDAKTYCZNEJ POLITECHNIKI WROCŁAWSKIEJ

Wrocław University of Technology

Environmental Quality Management

Monika Maciejewska

ENGINEERING APPLICATIONS OF MATHEMATICAL STATISTICS

Wrocław 2011

Projekt współfinansowany ze środków Unii Europejskiej w ramach
Europejskiego Funduszu Społecznego

Wrocław University of Technology

Environmental Quality Management

Monika Maciejewska

ENGINEERING APPLICATIONS OF MATHEMATICAL STATISTICS

Wrocław 2011

Reviewer: Andrzej Szczurek

ISBN 978-83-62098-67-5

Published by PRINTPAP Łódź, www.printpap.pl

TABLE of CONTENTS

PREFACE

The aim of this book is to build the capacity of applying statistical methods and tools in the professional practice of an engineer. Therefore, the focus is on understanding and the development of relevant skills.

This book covers a selection of statistical methods and tools. Their theoretical description is provided together with examples of application in solving engineering problems. When advantageous, hints for using statistical software are given.

From the scientific point of view, the presented methods and tools are elements of more advanced methodologies in engineering statistics which are subject to continuous development. It is intended that in the course of studying this book the Reader learns the appropriate language and lays the foundation for further development of knowledge and skills in the domain of engineering applications of statistics.

ORGANISATION OF THE BOOK

This book consists of several chapters with their order corresponding to the increasing complexity of the discussed statistical methods and tools as well as engineering problems which may be solved with their application. The following is a brief overview of the content found in the chapters.

- Random variable and its variability

A random variable is a principal entity in statistics. The concept of a random variable is presented and different types of random variables are described.

- Data collection

Data collection is necessary for obtaining values of random variables. Selected strategies of data collection are reported.

- Descriptive statistics

The statistical description of data may be used for characterizing real objects. Basic tools for the statistical description of data sets are presented.

- Theoretical distributions of discrete variables

Theoretical variables are available which may be used as models of real random discrete variables. A selection of distributions of theoretical discrete variables is presented.

- Theoretical distributions of continuous variables

Theoretical variables are available which may be used as models of real random continuous variables. A selection of distributions of theoretical continuous variables is presented.

- Confidence interval and confidence level

The confidence level represents the trust that a parameter of statistical distribution of a random variable remains within certain limits. The method of calculating confidence intervals on the mean and on the variance is explained.

- Statistical hypotheses and their testing

The testing of statistical hypotheses allows for comparing objects. Statistical tests are presented which allow for comparing the average states of objects and for comparing variabilities of the states of objects.

- Analysis of variance

The analysis of variance is used for detecting the change of objects due to the influence of nonrandom factors. The demonstrated methodology refers to cases when one or two nonrandom factors are considered simultaneously.

- Regression analysis

Regression analysis allows for the quantitative description of object change, which results from the influence of nonrandom factors. The principles of building regression models and their diagnostics are provided.

INTRODUCTION

The ENCYCLOPEDIA BRITANNICA defines engineering in the following way: "Engineering is the application of science to the optimum conversion of the resources of nature to the uses of humankind". The definition of statistics provided by ENCYCLOPEDIA BRITANNICA states "Statistics is a branch of mathematics dealing with gathering, analyzing, and making inferences from data". Statistics enters engineering by being a substantial fragment of mathematical knowledge applied in engineering. It is used for analyzing measurement/observation data concerning objects. Objects are fragments of the world, e.g. materials, structures, machines, devices, systems, phenomena and processes. They are studied by engineers in order to design, implement and control the 'use' of nature by humankind.

For an engineer, statistics provides aid in solving a number of problems, for instance

- characterizing objects,
- comparing objects,
- detecting change in objects,
- describing relationships within and between objects.

The engineering application of statistics consists of using statistical analysis for solving engineering problems. The following steps are required to implement this approach: (1) an engineering problem is expressed as a statistical problem, (2) a solution of the statistical problem is obtained, (3) the solution of the statistical problem is translated to the solution of the engineering problem. These principal elements of the approach are shown in Fig.1.

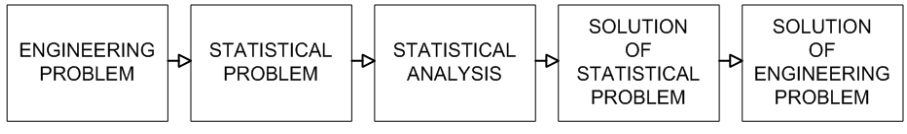


Figure 1 Pathway for solving engineering problems using statistical analysis.

For making use of engineering statistics, it is necessary that an engineer, a specialist in his/her own field, is additionally familiar with statistical methodology and is able to fuse these two domains in a proper way.

1 VARIABLE AND VARIABILITY

A **variable** may be used for representing a feature of an object or its surroundings. For example, let the object be *a chemical substance.* Such an object has many features, for instance *volatility.* This feature may be represented by the variable *saturated vapor pressure.*

A variable has a name, takes values or levels and is usually expressed in some units. For example, the levels of the variable *saturated vapor pressure* may be expressed in [Pa].

A variable taking a value or level is called realization. For example, the realization of the variable *saturated vapor pressure* may be 10150 Pa. There must be a possibility to observe/measure and record realizations of a variable.

The recorded realizations of variables are **data**. As already stated, statistical analysis operates on data.

Establishing the correspondence between features of an object and variables is the key point for transitioning between an engineering problem and a statistical problem.

1.1 SCALES AND TYPES OF VARIABLES

There are different types of variables. One of the most useful classifications divides variables according to scale providing levels/values of a variable. The following scales are available:

1. Nominal scale,
2. Ordinal scale,
3. Interval scale,
4. Ratio scale.

The **nominal scale** has levels that are different, but incomparable. There is no way to judge the size or direction of the difference. An example of a variable which takes levels from the interval scale is *sex*. Another example is *race*.

The **ordinal scale** also has levels. Levels of the ordinal scale are different and comparable. It is possible to rank the levels of an ordinal variable and to order them; however, it is not possible to measure the difference between the levels. An example of an ordinal variable is the freshness of air. Provided the air in room A is very fresh, the air in room B is medium fresh and the air in room C is not fresh, the rooms may be put in order according to the increasing freshness of air. However, the difference between the freshness of air in the rooms is unknown.

An **interval scale** has values that are different. It is possible to order the values and calculate the difference between levels. However, it is not possible to use the ratio of levels from the interval scale. In other words, the starting point of the interval scale is not absolute zero. The classical example of an interval variable is temperature measured in degrees Celsius. For example, assume liquid A has a

temperature of 40 °C and liquid B has a temperature of 70 °C. Clearly, the temperatures of liquid A and B are different. The temperature of liquid A is lower than the temperature of liquid B. The difference between the temperatures of liquids A and B is 30 °C. However, the ratio of the temperatures is not 40/70. It is 313/343. The ratio may be calculated if the absolute, Kelvin temperature scale is used.

The **ratio scale** has values and is an absolute scale with an absolute origin. Values from the ratio scale are different, can be ordered and subtracted and additionally their ratios can be calculated. An example of a ratio variable is the distance from a fixed point. Assume the distance between points A and O is 10 m and the distance between points B and O is 2 m. The following is concluded: the distances of points A and B from point O are different. Point A is located farther from point O than point B. There is an 8 m difference in the distance of points A and B from point O. Point A is located five times farther from point O than point B. The ratio scale is the most informative scale. The interval scale may be transformed into the ratio scale if the absolute reference point is defined.

Another method of classification uses **qualitative** and **quantitative** variables. Qualitative variables have levels and nominal or ordinate variables are qualitative. Quantitative variables have values and include interval or ratio variables. In general, statistics operates on quantitative variables. Qualitative variables may be used for representing features which have qualitative character. Oftentimes, they are applied for labeling classes, groups or sets of elements.

It is important to distinguish **discrete** and **continuous** variables. Discrete variables take values/levels from finite or countably infinite sets. Continuous variables take values from infinite sets. There are substantial differences in the logic of statistical analysis for discrete variables and continuous variables (see Chapter 4 and Chapter 5).

Using still another classification system, one may describe **independent** and **dependent/response** variables. Independent variables represent factors which influence the investigated objects. Dependent variables represent features of objects which are influenced by the factors. If jointly considered, symbol X is used for indicating the independent variables and letter Y refers to the dependent variables.

The type of variable determines the selection of methods which may be used in its statistical analysis. Therefore, it is very important to correctly identify the type of variable before attempting the analysis.

1.2 VARIABILITY OF VARIABLES

Variables exhibit variability in their values. There are two sources of variability considered in statistics: **random factors** and **nonrandom factors**.

Random factors are always present and there is no way to eliminate or control them. The magnitude and direction of their influence on objects changes in a nondeterministic manner. Contrarily, nonrandom factors may be controlled. It is possible to change the magnitude and direction of their influence on objects in a deterministic manner.

The **random variable** is represented by the following formal model:

$$X = \mu + \varepsilon$$

The first element of the sum, μ represents the influence of nonrandom factors on the variable. The second element of the sum, ε represents the influence of random factors. There are the two following possibilities:

1. Nonrandom factors remain at a constant level. In such circumstances variable X shows the variability exclusively caused by random factors which is equal to ε. Variable X does not show the variability caused by nonrandom factors. The value of μ is constant. Observed values of the variable randomly change around the constant level μ.
2. The level of a nonrandom factor is changed but the object is insensitive to this factor. See *case 1*.
3. The level of a nonrandom factor is changed and the object is sensitive to this factor. In such circumstances X shows the variability caused by nonrandom and random factors together. The variability caused by random factors is equal to ε. The variability caused by nonrandom factors is observed as the change of μ. Observed values of the variable randomly change around various levels of μ.

Statistics provides a means of detecting and analyzing the variability of variables. In this way, it is possible to make inferences about objects with analysis performed on a number of values/levels of variables. The set of actions aimed at their acquisition is usually referred to as data collection.

2 DATA COLLECTION

An elementary step of data collection is a single observation or measurement from which a single value of a variable is acquired.

There are different strategies for collecting data that depend on many factors, for example: the purpose of data collection, the constraints associated with the object, available methods and techniques of observation/measurement.

From an engineering point of view, it is particularly important to distinguish between a passive and an active strategy for data collection.

The data collected in a passive way provide extensive information about the 'natural' behavior of an object and may be used for characterizing the object. However, it is not possible to study the cause-response relationship between the object and its surroundings using data collected in a passive manner. Only a relationship which has a correlation character may be analyzed. The exception is the availability of the theory which describes the relationship.

With **passive data collection** the object is just observed. Its surroundings change without any deliberate action aimed at influencing the object. The recorded changes of the object (variability of the observed variable) usually result from a wide range of random and nonrandom factors. Nevertheless, the observed variability may not be undoubtedly attributed to changes of particular factors.

The data collected in an active manner provide information about the object being influenced by known nonrandom factors. Active data collection allows for studying cause-response relationships between the object and its surroundings.

Active data collection consists of observing the object while it is deliberately influenced by known nonrandom factors. The observer is in control of selected factors which may influence the object and manipulates these factors to see whether and how the object responds to their change.

The discipline of science that develops the methodology of planning active data collection is called Experimental Design. The reader will be presented with selected elements of experimental design in the chapter dedicated to the Analysis of Variance (see Chapter 8)

Another important distinction is made between collecting data for the entire population and sampling, i.e. collecting data for a part of a population.

In statistics, **population** is understood as the total set. The population can be fully characterized if each element of the set is known. However, populations usually consist of a large or even an infinite number of elements. This makes the investigation of every element impractical or even impossible. In such cases only a representation of the population is considered. A **sample** is a set of elements drawn from the population. The set shall be small enough to investigate each of its elements. Furthermore, it is expected that the sample is representative of the population.

The representative character of a sample is assured by the appropriate strategy of drawing with various strategies available. The most frequently used is called **random drawing**. In order to secure random drawing, the likelihood of pulling out an element from the population has to be the same for all elements. It is not known in advance which element will turn out from the draw, although the respective likelihood may be known.

Tables of random numbers and random number generators implemented in computer software are helpful in selecting random samples.

The majority of statistical methods and tools were developed for analyzing data provided by random sampling.

2.1.1 EXAMPLE.

Problem. A factory employs 700 workers. They all work in similar conditions. An employer was asked to select 50 workers who will be subject to a very detailed medical examination. The sample shall be representative for the entire group of employees.

Solution. In the considered problem the best representativeness is secured by random drawing. In order to solve the problem, we are going to use the generator of pseudorandom numbers, which is available in the DATA ANALYSIS TOOL in Excel. The path for obtaining the solution is the following:

- There is one variable – the id of the worker.
- The variable takes values of ordinal numbers between 1 and 700.
- There has to be a random sample drawn consisting of 50 elements, i.e. there are 50 requested values of the variable.
- The probability of drawing any single worker shall be constant and identical for all workers; therefore, the distribution of the variable is uniform.

Random numbers provided by the generator shall be rounded to integers. The results obtained by the author are shown in Table 2.1. The reader is encouraged to generate his/her own solution.

Table 2.1 Sample of 50 randomly selected numbers. The population consisted of 700 numbers from 1 to 700.

87	503	45	364	389	62	362	577	410	243
104	239	631	358	120	94	483	276	386	433
191	566	693	504	189	152	457	587	225	477
621	551	625	404	526	253	146	652	421	479
570	571	375	699	599	488	687	36	374	105

3 DESCRIPTIVE STATISTICS

An important category of engineering problems which may be addressed by the statistical methods and tools is related to characterizing objects. The realization of this task is possible by applying descriptive statistics to data sets. The data shall be realizations of variable, which represents a selected feature of the characterized object.

A number of numerical, graphical and combined tools allows for describing the principal features of the data set. Their use is recommended if nothing is known in advance about the variable represented by the recorded data. Otherwise, theoretical variables may be applied for representing the empirical variable (see Chapter 4 and Chapter 5) and the statistical analysis is performed in a different way.

The following tools are presented in this chapter: measures of centre in the data set, measures of spread in the data set, histogram, box and whisker plot.

3.1 CENTER

The center is a value representing the middle of a data set. There are a number of possibilities concerning the location of this feature. Three of the most frequently applied measures are the following:

- **Median** – The value of a variable such that 50 % of all recorded values are smaller than the median and 50 % of them are larger than the median. If the values of a variable are ordered decreasingly or increasingly, the median is the value from the middle. For an even number of measured values, the median is located half way between the two adjacent middle values. The median is a very good measure of center location and it is robust regarding extreme values of the variable.

- **Mode** or **modal value** – The value of a variable which occurs most frequently. It may happen that there are two or more modes. The mode is an adequate measure only in the case of discrete variables.

- **Mean** – The mean is calculated in the following way:

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

where: x_i is a single observation of variable X, n is the number of observations in the sample.

The symbol μ denotes the mean of the variable in the entire population. The symbol $\bar{x}$ denotes the arithmetic mean of the variable in the sample.

Oftentimes, the mean is automatically used as the indication of center in a set of data. However, this measure is sensitive to extreme values of the variable which may result in a false evaluation of center when the extreme values are actually faulty measurements.

3.2 SPREAD

The spread indicates the range of variability in the data set. There are a number of possibilities concerning the evaluation of spread. Three of the most frequently applied measures are the following:

- **Minimum** and **maximum** – The minimum is the smallest and the maximum is the largest value of the variable. These two limits indicate the range of recorded values of the variable. Minimum and maximum are very sensitive to extreme values of the variable. If the largest and the smallest values originate from faulty measurements, the actual variability of the variable may be much smaller than delimited by the $< min, max >$ range in the data set. Minimum and maximum values may be used together with any measure of center.
- k^{th} order **percentile** – A value of a variable such that k % of all recorded values are smaller than the percentile. This definition strictly refers to the so called lower percentile. For the case of the k^{th} upper percentile, k % of variable values exceeds the percentile. The spread is indicated by the pair of symmetric k^{th} percentiles: lower and upper.
  Most popular is the 25^{th} percentile, called the quartile. The minimum and maximum are actually the 0^{th} and 100^{th} percentiles, respectively.
  Percentiles are usually used together with the median. The distance from the center to the k^{th} order lower and upper percentiles indicates whether values of the variable are symmetrically distributed around the center of the data set.
- **Standard deviation** – Standard deviation is calculated in the following way:

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

where: x_i is a single observation of variable X, n is the number of observations in the sample.

The symbol σ denotes the standard deviation of a variable in the entire population. The symbol s denotes the standard deviation of a variable in the sample.

Oftentimes, standard deviation is automatically used as the indicator of spread in a data set and it is considered together with the mean. Standard

deviation does not indicate the symmetry or asymmetry of the distribution of variable values around the center.

- **Outliers** – These are observations which lie an abnormal distance from other values in a data set. There are mild and extreme outliers. Using the following notation: Q_L is the lower quartile, Q_u is the upper quartile and $IQ = Q_U - Q_L$ is the inter-quartile range, the following holds:
  - mild outliers belong to the interval $\langle Q_L - 1.5IQ) \cup (Q_u + 1.5IQ\rangle$
  - extreme outliers belong to the interval $\langle Q_L - 3IQ) \cup (Q_u + 3IQ\rangle$.

  An outlier is a 'strange' observation. The engineer has to decide whether it resulted from a faulty measurement or is a trace of abnormal object behavior. In the first case, the outlier shall be removed from the data set prior to any statistical analysis. Otherwise, the outlier shall be considered with special care.

3.3 HISTOGRAM

By quoting the measures of center and spread in a data set, the essential information is provided about the variable thus also about the investigated object. Namely, the value is known around which the variable varies and the magnitude of variation is given. In other words, the usual state of the object is indicated and it is also known how far from this state the object wanders.

Still a more detailed picture may be obtained by means of a histogram. In order to build a histogram, the range of values of the variable $< min, max >$ is divided into intervals of the same size. The number of intervals depends on the size of the data set. It is recommended to use odd numbers for the number of intervals. The **histogram of frequency** shows the frequency of occurrence, i.e. the number of times the values of the variable fall into different intervals. The frequency histogram is convertible into a **histogram of relative frequency**. The relative frequency histogram shows the relative frequency of occurrence, i.e. the percentage of values of the variable which fall into different intervals. In addition, the **histogram of cumulative frequency** is sometimes used. This shows the cumulative frequency of occurrence, i.e. the number of values of the variable which are smaller or equal to the right limit of the particular interval. The **histogram of cumulative relative frequency** is built similarly by using the cumulative relative frequency of occurrence. The principles of construction for the frequency histogram, relative frequency histogram, cumulative frequency histogram and cumulative relative frequency histogram are summarized in Table 3.1.

Table 3.1 The principles of constructing the frequency histogram, relative frequency histogram, cumulative frequency histogram and cumulative relative frequency histogram.

Indicator of interval	1	...	k	...	m
Limits of interval	$\langle x_{min}, x_{min} + 1\Delta x)$		$\langle x_{min} + (k-1)\Delta x, x_{min} + i\Delta x)$		$\langle x_{min} + (m-1)\Delta x, x_{max})$
Frequency of occurrence	n_1		n_k		n_m
Relative frequency of occurrence	$\frac{n_1}{n}$		$\frac{n_k}{n}$		$\frac{n_m}{n}$
Cumulative frequency of occurrence	n_1		$n_1 + n_2 + \cdots + n_k$		$\sum_{k=1}^{m} n_k$
Cumulative relative frequency of occurrence	$\frac{n_1}{n}$		$\frac{n_1}{n} + \frac{n_2}{n} + \cdots + \frac{n_k}{n}$		$\sum_{k=1}^{m} \frac{n_k}{n}$
Probability	$\frac{n_1}{n\Delta x}$		$\frac{n_k}{n\Delta x}$		$\frac{n_k}{n\Delta x}$

The following notation was used in Table 3.1: m is the number of intervals; x_{min} and x_{max} are minimum and maximum values of variable X, $\Delta x = \frac{x_{max} - x_{min}}{k}$ is the size of a single interval, n_k is the number of values of the variable which fall into the k^{th} interval, n is the number of all observations of variable X.

Histograms are plotted using a bar plot. The x axis represents variable X and the limits of the intervals are marked on this axis. A bar is plotted for each interval. The height of the bar represents the frequency of occurrence, relative frequency of occurrence, cumulative frequency of occurrence or cumulative relative frequency of occurrence, depending on the type of histogram. Graphical representations of frequency histograms and cumulative frequency histograms are shown in Fig. 3.1 and Fig. 3.2, respectively.

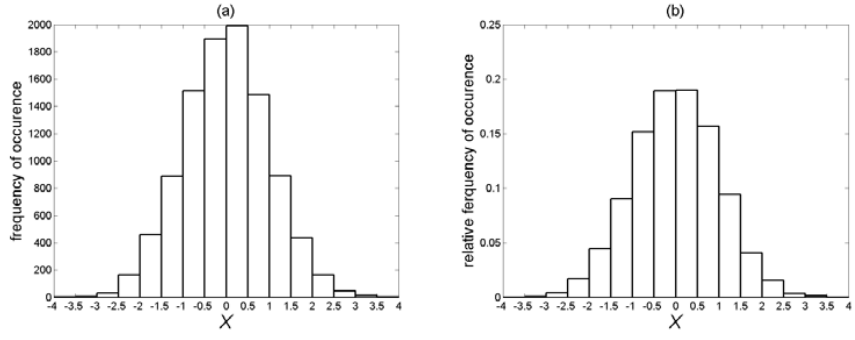


Figure 3.1 Graphical representation of (a) frequency histogram, (b) relative frequency histogram.

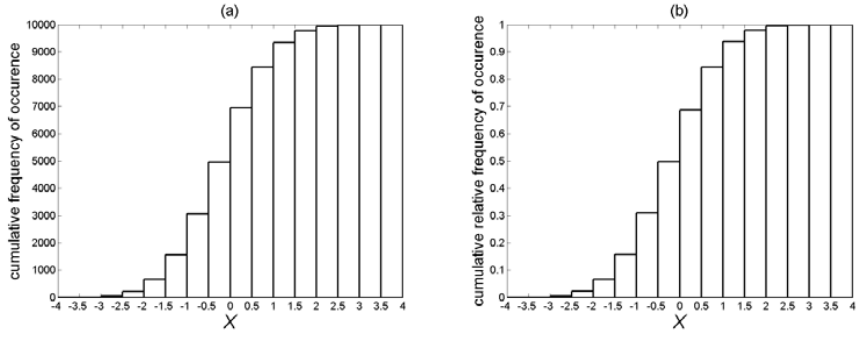


Figure 3.2 Graphical representation of (a) cumulative frequency histogram, (b) cumulative relative frequency histogram.

The relative frequency histogram provides the basis for calculating the **probability distribution** of a variable. The probability associated with an interval is calculated as the ratio between the relative frequency of occurrence in the interval and the interval length.

The cumulative relative frequency histogram is synonymous with the **cumulative probability distribution** of the variable. The height of the bar over the interval on the histogram plot is the probability that the value of the variable is smaller or equal to the right limit of that interval. The height of rightmost bar is always 1. It represents the fact that all the values in the sample are lower or equal to the maximum value of the variable. The associated probability is equal to one.

The principle of calculating probability distribution is shown in the last row in Table 3.1.

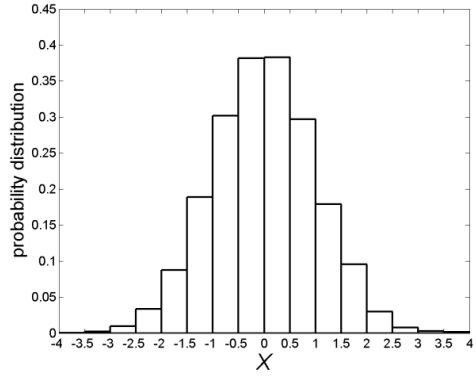


Figure 3.3 Graphical representation of empirical probability distribution.

A graphical representation of empirical probability distribution is shown in Fig. 3.3. Please note that the probability variable X takes a value from a selected interval is equal to the surface of the bar over this interval in the probability distribution plot. The total surface under the probability distribution plot is 1. It is the probability that all values of the variable in the sample fall between the minimum and the maximum value.

3.4 BOX AND WHISKER PLOT

The **box and whisker plot** is a convenient synthetic graphical presentation for the empirical distribution of a variable including measures of center and spread for the data set. The main components of the box and whisker plot are shown in Fig. 3.4. The bottom axis displays values of the considered variable. The plot itself consists of a rectangle (box) and two horizontal lines (whiskers) which stretch left and right from the box. The vertical line inside the box represents the median. Two sides of the box represent quartiles. The left side refers to the lower quartile and the right side refers to the upper quartile. The left part of the box contains 25 % of the values of the variable while the other 25 % of values belong to the right part of the box. The left horizontal line extends between the minimum value of the variable and the lower quartile while the right horizontal line extends between the upper quartile and the maximum value of the variable. 25 % of the values of the variable are contained in the left whisker while another 25 % belong to the right whisker. The minimum and maximum are calculated for the data set after excluding outliers which are marked with crosses on the box and whisker plot.

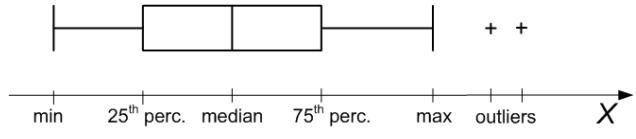


Figure 3.4 Principle of constructing a box and whisker plot.

The box and whisker plot is much more comprehensive compared to numerical representations of population center and spread. It is also more synthetic than a histogram. With this plot the empirical distributions of different variables may be easily compared. An example of such a comparison is shown in Fig. 3.5 using three imaginary variables X_A, X_B and X_C.

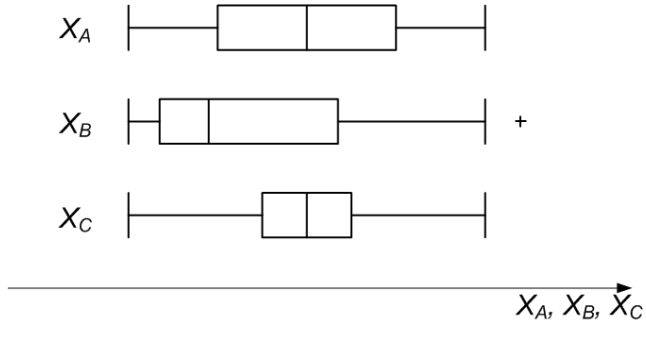


Figure 3.5 Comparison of the empirical distributions of three variables $\boldsymbol{X_A}$, $\boldsymbol{X_B}$ and $\boldsymbol{X_C}$ using box and whisker plots.

The distribution of variable X_A, shown in Fig. 3.5, is rather symmetric. The distance of both quartiles from the median is the same. So is the distance of minimum and maximum from the median. Contrarily, the distribution of variable X_B, also shown in Fig. 3.5, is asymmetric. The median is not located in the middle between the minimum and maximum value or half way between the lower and upper quartile. The distance between the median and the lower quartile is shorter than between the median and the upper quartile. Similarly, the distance between the median and the minimum is shorter than between the median and the maximum. That is 50 % of values, those which are greater than the median, belong to a longer interval than 50 % of the values which are smaller than the median. The box and whisker plot is 'longer' on the right side. The variable has **right skewed** or **positive skewed** distribution. An analogical plot but 'longer' on the left side would represent the **left-skewed** or **negative skew** distribution. The comparison between the box and whisker plot of variable X_C and variable X_A (Fig. 3.5) reveals another aspect of probability distribution. The inter-quartile range in the case of variable X_C is smaller as compared to X_A, although by definition in both cases 50 % of observations fall into that interval. The distribution of variable X_C is more 'peaked' as compared to X_A. The indicator of 'peakedness' is a quantity called **kurtiosis**. A larger kurtosis indicates a more peaked distribution.

3.4.1 EXAMPLE

Problem. Measurements of daily concentrations of NO_x, performed in June 2009 by the air pollution monitoring station located in Wrocław at Wiśniowa Street are given in Table 3.2. Characterize the level of pollution regarding NO_x at this location in Wrocław in June 2009 based on the provided data set.

Table 3.2 Daily concentration of NO_x measured by the air pollution monitoring station located in Wrocław, at Wiśniowa Street, in June 2009.

day	NO_x/ µg/m^3	Day	NO_x/ µg/m^3	day	NO_x/ µg/m^3
1	194	11	180	21	195
2	196	12	110	22	175
3	79	13	79	23	183
4	167	14	224	24	192
5	151	15	275	25	192
6	96	16	166	26	139
7	214	17	181	27	98
8	185	18	175	28	230
9	202	19	144	29	259
10	152	20	131	30	231

Solution. The basic components of the statistical description of the data set are the measures of center and spread. Nothing is known in advance about the kind of distribution of the variable: daily concentration of NO_x at Wiśniowa Street in Wrocław. Therefore, we are going to use the median in order to indicate the center and percentiles (minimum, maximum, upper and lower quartiles) for the representation of spread. The numerical values of these measures are given in Table 3.3.

Table 3.3 Measures of center and spread for the data set given in Table 3.2.

Median	180.5
Minimum	79
Maximum	259
lower quartile	144
upper quartile	196

Also, the graphical representation of major features of the data set is shown in Fig. 3.6 using a box and whisker plot. Additionally, the relative frequency histogram is displayed in Fig. 3.7.

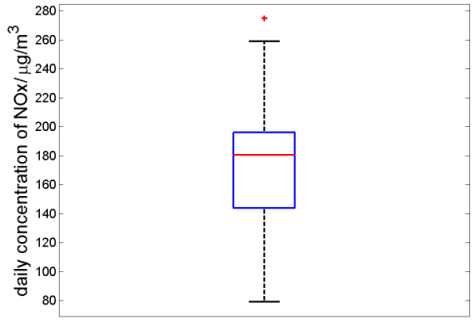


Figure 3.6 Box and whisker plot for the data set shown in Table 3.2.

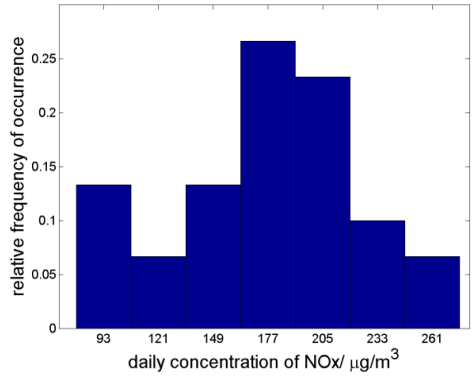


Figure 3.7 Relative frequency histogram for the data set shown in Table 3.2.

Based on the provided descriptors, the following may be concluded about the level of NO_x pollution in June 2009 at Wiśniowa Street in Wrocław:

- the daily concentration of NO_x varied around the level of 180.5 µg/m^3,
- 50 % of the time the concentration remained in a range between 144 and 195 µg/m^3,
- the minimum observed concentration was 79 µg/m^3 and the maximum concentration was 259 µg/m^3,
- the observed maximum concentration of 275 µg/m^3 was considered as an outlier, which may indicate faulty measurement,
- the distribution of data around the center is not clearly symmetric, but also a definite asymmetry was not observed.

4 DISCRETE VARIABLES AND THEIR PROBABILITY DISTRIBUTIONS

4.1 DISCRETE VARIABLES

An important group of variables encountered in engineering practice have discrete character. Statistics provides a description for a number of theoretical discrete variables, in particular regarding their probability distributions. Theoretical discrete variables are actually formalized representations of certain categories of real discrete variables. The most commonly encountered categories of real discrete variables, which have their theoretical counterparts, represent

(1) the number of elements which have a particular attribute in a sample drawn from a population, for example the number of faulty pumps in the sample from the production lot;
(2) the size of a sample in which a defined fraction of elements has a particular attribute, for example the size of a sample of students in which there are two students with the best grade;
(3) the number of times that a particular event occurs, for example the number of car crashes on the crossing during the average weekend; the number of times the engine starts before it fails to start for the first time; the number of times a batch of microprocessors has to be sampled before the first wrong microprocessor is found.

Discrete variable X takes values x_i from a finite $i = 1, 2, \dots, n$ or countably infinite $i = 1, 2, \dots$ set.

Each value x_i has a probability of occurrence assigned to it. The probability of occurrence is denoted by $p(x_i)$.

Discrete variable X has its probability distribution function, $P(X) = p(X = x_i)$.

The probability $p(X = x_i)$ fulfills the following conditions:

- for a finite set of values n

$$(\forall x)\, p(X = x_i) \geq 0 \text{ and } \sum_{i=1}^{n} p(x_i) = 1$$

- for an infinite set of values

$$(\forall x)\, p(X = x_i) \geq 0 \text{ and } \sum_{i=1}^{\infty} p(x_i) = 1$$

Discrete variable X has its cumulative distribution function, $F(X) = p(X <= x)$.

Graphical representations of the probability distribution function and the cumulative distribution function of discrete variable are shown in Fig. 4.1. The stem plot is used for plotting the probability distribution function of a discrete variable (Fig. 4.1 a). The height of each stem indicates the probability of occurrence for a single value of X. The stair-like plot is used for plotting the cumulative distribution function of a discrete variable (Fig. 4.1 b). Stairs climb from zero, which indicates zero probability that X is smaller than the minimum value, up to one, which indicates that all values of X are smaller or equal to its maximum.

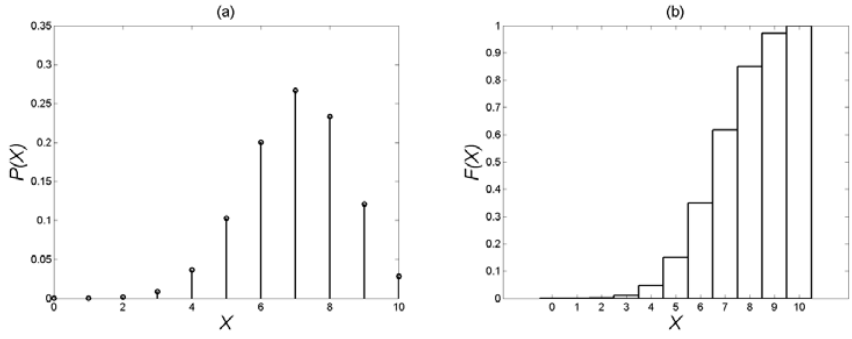


Figure 4.1 Graphical representation of (a) probability distribution function and (b) cumulative distribution function of a discrete variable.

The mean of a random variable which has a discrete character is calculated using the following formula:

$$\mu = \sum_i x_i p(x_i)$$

The variance of a discrete random variable is calculated by the formula:

$$\sigma^2 = \sum_i (x_i - \mu)^2 p(x_i).$$

The following distributions of theoretical discrete variables were selected for presentation in this book: Binomial, Poisson, Negative Binomial, Geometric, and Multinomial. The choice was guided by their applicability to solving practical engineering problems.

4.2 BINOMIAL DISTRIBUTION

Variable X which has binomial distribution may be described in the following way:

- the number of successes in a defined number of trails,
- the number of elements which have a particular attribute in the sample of defined size.

The probability distribution of a binomial variable $P(X)$ is described by the following formula:

$$P(X) = \binom{n}{x} p^x q^{n-x}$$

where: n is the number of trails/size of the sample; p is the probability of success in one trail/probability that a single element in a set has a particular attribute; $q = 1 - p$ is the probability of failure in one trail/ probability that a single element does not have the certain attribute.

The mean μ and variance σ^2 of variable X are calculated as follows:

$$\mu = np$$
$$\sigma^2 = npq.$$

4.2.1 EXAMPLE 1

Problem. The supplier is allowed to provide no more than 2 % defective parts per lot. Lots are huge and consist of 1000 items each. Every lot is randomly sampled for testing. It is technically possible to take a sample which consists of 10 elements. If the number of defective parts in the sample is 0, the lot is passed. Otherwise the lot it is rejected. Find the probability that a lot which contains: (a) 2 % defective parts is accepted, (b) 10 % defective parts is accepted.

Solution. Let us consider a theoretical discrete random variable X described as the number of elements in the sample which have a particular attribute. This well represents a real random variable which is encountered in our problem, namely the number of defective parts in the sample. The variable X has binomial distribution. Based on the problem description, the size of the sample is $n = 10$, the probability p that a single part is defective is: (a) $p = 0.02$ and (b) $p = 0.1$, respectively. The probability of accepting a lot of parts is equivalent to the probability that $X = 0$ in the sample of size $n = 10$. The following are calculations for cases (a) and (b).

(a) $P(X) = \binom{n}{x} p^x q^{n-x} = \binom{10}{0} 0.02^0 0.98^{10} = 0.82$

It is quite unlikely to reject a lot which contains 2 % faulty parts based on a 10 element random sample. The probability of lot rejection is $p = 1 - 0.82 = 0.18$.

(b) $P(X) = \binom{n}{x} p^x q^{n-x} = \binom{10}{0} 0.1^0 0.9^{10} = 0.35$

It is quite likely to reject a lot which contains 10 % faulty parts based on a 10 element random sample. The probability of rejection is $p = 1 - 0.35 = 0.75$.

The reader is encouraged to investigate how the size of a sample influences $P(X)$ by calculating solutions for $n = 5$ and $n = 20$.

4.2.2 EXAMPLE 2

Problem. The installation is equipped with 10 pumps. Based on the information from the producer, the probability that a single pump fails in one year of operation is approximately 0.05. Answer the following questions:

(a) What is the probability that none of the pumps fail during one year?
(b) What is the probability that all 10 pumps fail during one year?
(c) What is the probability that a single pump does not fail during 10 years?

(d) What is the probability that a single pump fails once every year during 10 years?

Solution.

- Let us consider a theoretical discrete random variable X described as the number of elements in a sample which have a particular attribute. This well represents a real random variable which is encountered in our problem, in case (a) and (b), namely the fraction of pumps which fail during one year of operation. The variable X has binomial distribution. The size of the sample is $n = 10$ and the probability that a single pump fails in one year is $p = 0.05$. We search for the probability that X takes a defined value: (a) $X = 0$, (b) $X = 10$. The relevant calculations are the following:

(a) $P(X) = \binom{n}{x} p^x q^{n-x} = \binom{10}{0} 0.05^0 0.95^{10-0} = 0.60$

The probability that none of the 10 pumps fail during year 1 is 0.60.

(b) $P(X) = \binom{n}{x} p^x q^{n-x} = \binom{10}{10} 0.05^{10} 0.95^{10-10} = 9.76 \cdot 10^{-14}$

The probability that 10 of 10 pumps fail during 1 year is $9.76 \cdot 10^{-14}$.

- Let us consider a theoretical discrete random variable X described as the number of successes in a defined number of trails. This well represents a real random variable which is encountered in our problem, in cases (c) and (d), namely the number of times a single pump fails during 10 years of operation. The variable X has binomial distribution. The size of the sample is $n = 10$ and the probability that a single pump fails in one year is p = 0.05. We search for the probability that X takes a defined value: (a) $X = 0$, (b) $X = 10$. The relevant calculations are the following:

(c) $P(X) = \binom{n}{x} p^x q^{n-x} = \binom{10}{0} 0.05^0 0.95^{10-0} = 0.60$

The probability that a single pump does not fail during 10 years is 0.60.

(d) $P(X) = \binom{n}{x} p^x q^{n-x} = \binom{10}{10} 0.05^{10} 0.95^{10-10} = 9.76 \cdot 10^{-14}$

The probability that a single pump fails ten times in course of 10 years is $9.76 \cdot 10^{-14}$.

Please note that we ignore the possibility of a single pump failing more often than once a year.

As shown by the obtained results, identical probabilities were obtained in cases (a) and (c), as well as in cases (b) and (d), although the paired cases represent conceptually different problems.

4.3 POISSON DISTRIBUTION

The Poisson distribution is a special case of the Binomial distribution. The Poisson distribution shall be employed when the sample is large and the probability of success in a single trail is very small.

Variable X which has binomial distribution may be described as

- the number of successes,
- the number of elements which have a particular attribute.

The probability distribution of a Poisson variable $P(X)$ is described by the following formula:

$$P(X) = e^{-\lambda}\frac{\lambda^x}{x!}, \qquad x = 0, 1, 2, \ldots$$

where: λ is the parameter of the distribution.

The mean μ and variance σ^2 of variable X are calculated as follows:

$$\mu = \lambda$$
$$\sigma^2 = \lambda$$

4.3.1 EXAMPLE

Problem. There are $10\,000$ joints in a very complicated installation. The probability that a single joint fails in two years time is 0.1 %. The producer gives a 2 year guarantee for the installation. Calculate the probability that a) none of the joints, b) no more than 10 joints fail in that period of time.

Solution. Let us consider a theoretical discrete random variable X described as the number of elements which have a particular attribute. This well represents a real random variable encountered in our problem, namely the number of joints which fail during two years of installation life. The variable X has Poisson distribution. In order to utilize the probability distribution of a Poisson variable, the parameter λ has to be calculated. Using the formula for the mean, which holds for Binomial distribution (§4.2.1), the mean number of parts which fail during two years is

$$\mu = np = 10000 \cdot 0.001 = 10$$

The requested parameter λ of Poisson distribution is $\lambda = \mu = 10$.

(a) The probability that none of the parts fail during two years of installation life is the probability that $X = 0$:

$$P(X) = e^{-\lambda}\frac{\lambda^x}{x!} = e^{-10}\frac{10^0}{0!} = e^{-10} = 4.54 \cdot 10^{-5}$$

The probability that none of the parts fail during two years of installation life is $4.54 \cdot 10^{-5}$. Such a situation is very unlikely.

(b) The probability that no more than 10 parts fail during two years of installation life is the probability that 0 or 1 or 2, …, or 10 parts fail. That is $X = 0$ or $X = 1$,…, or $X = 10$.

$$X = 1,\ P(X) = e^{-\lambda}\frac{\lambda^x}{x!} = e^{-10}\frac{10^1}{1!} = 4.54 \cdot 10^{-4}$$

$$X = 2,\ P(X) = e^{-\lambda}\frac{\lambda^x}{x!} = e^{-10}\frac{10^2}{2!} = 2.27 \cdot 10^{-3}$$

…

$$X = 10,\ P(X) = e^{-\lambda}\frac{\lambda^x}{x!} = e^{-10}\frac{10^{10}}{10!} = 1.25 \cdot 10^{-1}$$

$$\begin{aligned}\mathrm{P(X = 0, X = 1, \dots, X = 10)} &= \mathrm{P(X = 0) + P(X = 1) + \cdots + P(X = 10)} \\ &= 4.54 \cdot 10^{-5} + 4.54 \cdot 10^{-4} + 2.27 \cdot 10^{-3} + \cdots + 1.25 \cdot 10^{-1} \\ &= 0.583\end{aligned}$$

The probability that no more than 10 parts fail during two years of installation life is 0.583.

The reader is invited to perform the additional calculations and to plot the probability distribution of variable X, $P(X)$ for $X = 0, 1, \dots, 30$.

4.4 NEGATIVE BINOMIAL DISTRIBUTION

Variable X which has negative binomial distribution may be described in the following way:

- the number trials which are needed to obtain a success r-times,
- the size of a sample needed to find r elements which have a particular attribute.

The probability distribution of a negative binomial variable $P(X)$ is described by the following formula:

$$P(X) = \binom{x-1}{r-1} p^r q^{x-r}$$

where: r is the number of successes requested in x trails (number of elements which have a particular attribute); p is the probability of success in one trail/probability that a single element has the attribute; and $q = 1 - p$ is the probability of failure in one trail/probability that a single element does not have the attribute.

The mean μ and variance σ^2 of variable X are calculated as follows:

$$\mu = \frac{r}{p}$$

$$\sigma^2 = \frac{r(1-p)}{p^2}$$

A special case of Negative Binomial distribution is the Geometric distribution. The variable X, which has Geometric distribution describes the number of trails needed to obtain success for the first time (the size of the sample needed to find 1 element which has a certain attribute). Therefore, the Geometric distribution is the Negative Binomial distribution with $r = 1$. The Reader is invited to develop the formulas describing $P(x)$, μ and σ for the Geometric distribution.

4.4.1 EXAMPLE

Problem. The supplier is allowed to provide no more than 2 % defective parts per lot. Lots are huge and consist of 1000 items each. The delivered lot is randomly sampled for testing. Answer the following questions:

(a) What is the average size of the test sample which contains one faulty element?
(b) What is the average size of the test sample which contains three faulty elements?
(c) What is the probability that the first faulty element is found in the 10^{th} trial?
(d) What is the probability that the third faulty element is found in the 10^{th} trial?

Solution:

- Let us consider a theoretical discrete random variable X described as the size of the sample needed to find r elements which have certain attribute. This well represents a real random variable encountered in our problem, in cases (a) and (b), namely the size of the sample needed to find a defined number of faulty parts. The variable X has negative binomial distribution. The probability that a single element is faulty is $p = 0.02$ while the requested number of faulty parts r is (a) $r = 1$ and (b) $r = 3$, respectively. Calculations for the average value of variable X in cases (a) and (b) are given as follows.

(a) $\mu = \frac{r}{p} = \frac{1}{0.02} = 50$

The average sample size containing 1 faulty element is 50.

(b) $\mu = \frac{r}{p} = \frac{3}{0.02} = 150$

The average sample size containing 3 faulty elements is 150.

- Let us consider a theoretical discrete random variable X described as the number of trials which are needed to obtain success r-times. This well represents a real random variable encountered in our problem, in cases (c) and (d), namely the ordinal trial number in which the r^{th} faulty element is found. The variable X has negative binomial distribution. The probability of finding a faulty element in one trial is $p = 0.02$ while the success expected is (c) $r = 1$ and (d) $r = 3$ times,

respectively in the course of 10 trials. The relevant probability calculations are given as follows:

(c) $P(X) = \binom{x-1}{r-1} p^r q^{x-r} = \binom{10-1}{1-1} 0.02^1 0.98^{10-1} = 0.017$

The probability that the first wrong part is drawn in the 10th draw is 0.017.

(d) $P(X) = \binom{x-1}{r-1} p^r q^{x-r} = \binom{10-1}{3-1} 0.02^3 0.98^{10-3} = 0.000025$

The probability that the third wrong part is drawn in the 10th draw is 0.000025. Such a situation is very unlikely.

4.5 MULTINOMIAL DISTRIBUTION

The binomial distribution is the special case of multinomial distribution. Multinomial distribution refers to m variables $X_1, X_2, \dots, X_m$.

With multinomial distribution the probability is calculated that $X_1 = x_1$, and $X_2 = x_2$, …, and $X_m = x_m$. This may be described in the following way.

- The event of the 1^{st} type occurs x_1 times, and the event of the 2^{nd} type occurs x_2 times, …, and the event of the m^{th} type occurs x_m times. There are n events in total.
- There are x_1 elements of the 1^{st} type and there are x_2 elements of the 2^{nd} type, …, and there are x_m elements of the m^{th} type. The sample consist of n-elements.

Multinomial probability distribution $P(X)$ is described by the following formula:

$$P(X_1 = x_1, \dots, X_m = x_m) = \frac{n!}{x_1! \cdot \dots \cdot x_m!} p_1^{x_1} \cdot \dots \cdot p_m^{x_m}, \quad \text{and} \quad \sum_{k=1}^{m} x_k = n$$

where: x_k is the number of times the k^{th} event occurs during n trials, p_k is the probability that the k^{th}event occurs in a single trial, $k = 1 \dots m$.

4.5.1 EXAMPLE.

Problem. A construction element is produced which has 2 delicate holders. Based on experience, there is a 75 % chance that a randomly selected user will not destroy any holder, a 15 % chance that the user will destroy one holder, and a 10 % chance that the user will break two holders while fixing the element during construction.

(a) What is the probability that among 20 randomly selected users there are 15 who fixed the element successfully, 3 who broke 1 holder and 2 who damaged 2 holders?

(b) Is the probability calculated in case (a) different from the one associated with the following conditions: the sample consists of 100

users and we expect 75 successful users, 15 users who broke 1 holder and 10 users who broke 2 holders?

Solution. Let us consider multinomial distribution referring to the following case: there are x_1 elements of the 1^{st} type, x_2 elements of the 2^{nd} type, …, and there are x_m elements of the m^{th} type in the n-element sample. This well represents our problem if the following assignment is performed: X_1 is the number of users who did not do any harm to the holders, X_2 is number of users who broke 1 holder, X_3 is number of users who damaged 2 holders, p_1 is the probability that a randomly selected user will mount the element successfully and p_2 and p_3 are probabilities that the user will damage 1 and 2 holders, respectively. Based on the problem formulation, the probabilities are the following: $p_1 = 0.75$, $p_2 = 0.15$ and $p_3 = 0.1$. The following calculations for cases (a) and (b) are provided.

(a) In this case the probability is calculated for $X_1 = 15, X_1 = 3, X_3 = 2$ and $n = 20$.

$$P(X_1, X_2, X_3) = \frac{n!}{x_1! \cdot x_2! \cdot x_3!} p_1^{x_1} \cdot p_2^{x_2} \cdot p_3^{x_3} = \frac{20!}{15!\,3!\,2!} 0.75^{15} 0.15^3 0.10^2 = 0.070$$

The probability that the proportions of users who break none of the holders, one holder and two holders are 15:3:2 in a 20 element sample of users is 0.07.

(b) In this case, the probability is calculated for $X_1 = 75$, $X_2 = 15$, $X_3 = 10$ and $n = 100$.

$$P(X_1, X_2, X_3) = \frac{n!}{x_1! \cdot x_2! \cdot x_3!} p_1^{x_1} \cdot p_2^{x_2} \cdot p_3^{x_3} = \frac{100!}{75!\,15!\,10!} 0.75^{75} 0.15^{15} 0.10^{10} = 0.015$$

The probability that the proportions of users who break none of the holders, one holder and two holders are 15:3:2 in a 100 element sample of users is 0.015. The probabilities calculated in cases (a) and (b) are different.

5 CONTINUOUS VARIABLES AND THEIR PROBABILITY DISTRIBUTIONS

5.1 CONTINUOUS VARIABLES

A substantial group of variables encountered in engineering practice have continuous character. Considering their applicability, the most commonly used continuous variables represent physical and chemical properties of physical objects. Their examples are the following: temperature, humidity, concentration, content, age, speed, height and many others.

Continuous variable X takes values from an infinite set.

In the case of continuous variables, a probability of occurrence is not assigned to a single value of variable X. The probability is instead assigned to an interval of values of variable X. This is a so called interval estimation.

A continuous variable has a probability density function $f(x)$, with the following properties:

$$(\forall x) f(x) > 0$$

$$\int_a^b f(x)dx = P(a < X \leq b) \text{ , for any } a < b$$

$$\int_{-\infty}^{\infty} f(x)dx = P(-\infty < X \leq \infty) = 1$$

A continuous variable has a cumulative distribution function $F(X)$, with the following properties:

$$F(x) = P(X < x) = \int_{-\infty}^{x} f(x)dx$$

Graphical representations of the probability density function (*PDF*) and cumulative distribution function (*CPDF*) of continuous variable are shown in Fig. 5.1.

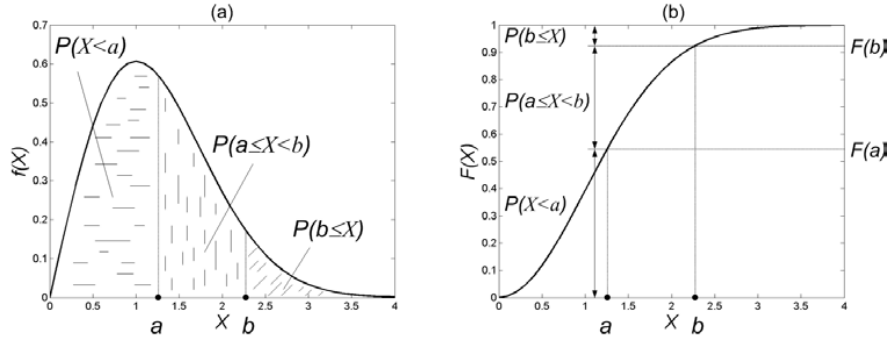


Figure 5.1 Graphical representations of (a) probability density function, (b) cumulative distribution function, of a continuous variable.

Both the *PDF* and *CPDF* of a variable are useful for finding the probability that values of the variable belong to a defined interval.

The following features of *PDF* are most frequently exploited in practice:

- $\int_{-\infty}^{a} f(x)dx$, i.e. the surface under the *PDF*, between $X = -\infty$ and $X = a$, (Fig. 5.1a) is the probability $P(X \leq a)$ that variable X has values smaller or equal to a;
- $\int_{a}^{b} f(x)dx$, i.e. the surface under the *PDF*, between $X = a$ and $X = b$, (Fig. 5.1a) is the probability $P(X \in \langle a, b \rangle)$ that variable X has values in the interval $\langle a, b \rangle$;
- $\int_{b}^{\infty} f(x)dx$, i.e. the surface under the *PDF*, between $X = b$ and $X = \infty$, (Fig. 5.1a) is the probability $P(X \geq b)$ that variable X has values greater than or equal to b.

The following features of *CPDF* are most frequently exploited in practice:

- $F(a)$, i.e. the value of *CPDF*, for $X = a$, (Fig. 5.1b) is the probability $P(X \leq a)$ that variable X has values smaller or equal to a
- $F(b) - F(a)$, i.e. the difference between values of *CPDF*, for $X = b$ and $X = a$, (Fig. 5.1b) is the probability $P(X \in (a, b\rangle$ that variable X has values in the interval $(a, b\rangle$
- $1 - F(b)$, i.e. the difference between one and the value of *CPDF*, for $X = b$, (Fig. 5.1b) is the probability $P(X > b)$ that variable X has values greater than b.

The mean of a continuous random variable is calculated by the following formula:

$$\mu = \int_{-\infty}^{\infty} x f(x)dx$$

The variance of a continuous random variable is calculated as follows:

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2$$

There are a number of theoretical continuous variables which have well defined probability density functions. Their $PDFs$ are known as equations, but they are also available in the form of statistical tables (see Appendix 1-5, 7).

The following theoretical $PDFs$ of continuous variables were selected for presentation in this book: normal, *t*-Student, Chi^2 and *F*-Snedecore. This choice was guided by their practical applicability in solving engineering problems.

5.2 NORMAL DISTRIBUTION

The Normal distribution is the most desired distribution of the observed random variable.

Variable X, which has the probability distribution described by the following probability density function:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \text{ for } x \in \langle -\infty, \infty \rangle$$

where: μ is the mean of X, σ is the standard deviation of X, is considered as having normal distribution.

The PDF of normal distribution has two parameters: μ and σ. This fact is represented using the following notation: $N(\mu, \sigma)$.

A selection of probability density functions for normal variables is presented in Fig. 5.2.

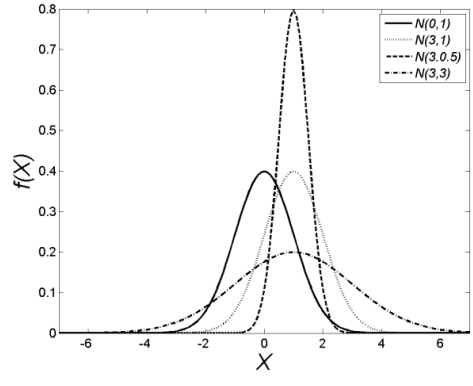


Figure 5.2 Examples of normal probability density functions.

The probability density function of normal distribution has a bell shape, as shown in Fig. 5.2. This shape is also called Gaussian. The normal PDF function is symmetric. The location of the function maximum is determined by μ, whereas its flatness depends on σ.

The special case of normal distribution is the standardized normal distribution, $N(0,1)$. It is the normal distribution with the mean $\mu = 0$ and the standard deviation $\sigma = 1$. The variable having standardized normal distribution is called Z. The Z variable is obtained by transforming the X variable, which has normal distribution $N(\mu, \sigma)$, in the following way:

$$Z = \frac{X - \mu(X)}{\sigma(X)}$$

The Z variable is very useful in practical applications of statistics.

When using normal distribution for describing the distribution of the observed variable X, $\bar{X}$ is used as the estimate of the mean μ, and s^2 is used as the estimate of variance σ^2 of variable X (see §3.1 and §3.2).

Statistical tables of normal distribution refer to the Z variable. The most commonly used form of Z distribution tables is provided in Appendix 1. Due to the symmetric character of the distribution, just the right part of it, i.e. for $z \in < 0, \infty)$ is described in Z tables.

It is very convenient to deal with a variable having normal distribution. Many statistical methods require that the analyzed variable has normal distribution and fulfilling this assumption is required for the valid use of such methods. There are a number of statistical tests available for checking the normality of variables (see §7.7).

5.2.1 EXAMPLE

Problem. It is known that variable X has normal distribution $N(150, 5)$. What is the probability that values of variable X

(a) are greater than 157?
(b) are less than 146?
(c) belong to the following intervals: 150 ± 5; 150 ± 10; 150 ± 15.

Solution. Considering that variable X has normal distribution, Z distribution may be used to solve the problem. First, the normal variable X has to be converted to the standardized variable Z. In the next step, Z statistical tables shall be used (Appendix 1). Solutions for cases (a), (b) and (c) are given as the following.

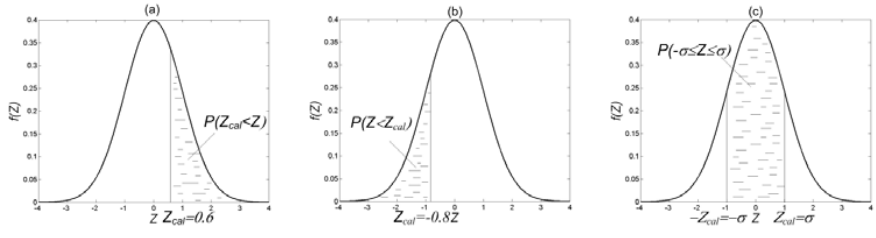


Figure 5.3 Graphical illustration of problem 5.2.1.

(a) $z_{cal} = \frac{x - \mu(X)}{\sigma(X)} = \frac{157 - 150}{5} = 0.6$

In order to use the Z tables, the calculated value of $z_{cal} = 0.6$ shall be substituted for Z_α. In the Z distribution tables one finds $P(Z \leq z_\alpha) = P(Z \leq 0.6) = 0.7257$. Therefore, the requested probability is

$$P(Z > 0.6) = 1 - P(Z \leq 0.6) = 1 - 0.7257 = 0.2743$$

The probability that the value of variable X, is greater than 157 is 0.2743.

The graphical interpretation of the probability $P_a = P(X > 157) = P(Z > 0.6)$ is shown in Fig. 5.3a.

(b) $z_{cal} = \frac{x-\mu(X)}{\sigma(X)} = \frac{146-150}{5} = -0.8$

In order to use Z tables, the negative value $Z = -0.8$ shall be reflected in order to produce a positive value $-Z = 0.8$. This is allowed due to the symmetry of normal distribution. Next, the calculated value $-z_{cal} = 0.8$ is substituted for z_α. From the table of Z distribution one reads $P(Z \leq z_\alpha) = P(Z \leq 0.8) = 0.7881$. Therefore, the requested probability is

$$P(Z < -0.8) = P(Z > 0.8) = 1 - P(Z \leq 0.8) = 1 - 0.7881 = 0.2119$$

The probability that the value of variable X is less than 146 is 0.2119.

The graphical interpretation of the probability $P_b = P(X < 146) = P(Z < -0.8)$ is shown in Fig. 5.3b.

(c) Two limits between which the X variable is supposed to fall are (i) $\langle 150 - 5, 150 + 5\rangle$, (ii) $\langle 150 - 10, 150 + 10\rangle$ and (iii) $\langle 150 - 15, 150 + 15\rangle$. Please note that intervals (i), (ii) and (iii) represent the so called 1σ, 2σ and 3σ intervals (see §6.1).

To make use of Z tables, right limits of the intervals of the X variable are transformed into Z. Next, the calculated values z_{cal} are substituted for z_α in order to read the probability $P(Z \leq z_\alpha)$. The following are calculations for cases (i), (ii) and (iii).

(i) $$Z_{cal} = \frac{x-\mu(X)}{\sigma(X)} = \frac{150+5-150}{5} = 1$$

$$P(-Z_\propto \leq Z < Z_\propto) = 2(P(Z < Z_\propto) - 0.5) = 2(0.8413 - 0.5) = 0.6826$$

The probability that variable X belongs to the interval $\langle 150 - 5, 150 + 5\rangle$ is 68.26 %. In general, the probability that variable X, which has normal distribution, belongs to the interval $\langle \mu - \sigma, \mu + \sigma\rangle$ is 0.6826.

(ii) $$z_{cal} = \frac{x-\mu(X)}{\sigma(X)} = \frac{150+2\cdot5-150}{5} = 2$$

$$P(-Z_\propto \leq Z < Z_\propto) = 2(P(Z < Z_\propto) - 0.5) = 2(0.97725 - 0.5) = 0.9545$$

The probability that variable X belongs to the interval $\langle 150 - 2\cdot5, 150 + 2\cdot5\rangle$ is 0.9545. In general, the probability that variable X, which has normal distribution, belongs to the interval $\langle \mu - 2\sigma, \mu + 2\sigma\rangle$ is 0.9545.

(iii) $$z_{cal} = \frac{x-\mu(X)}{\sigma(X)} = \frac{150+3\cdot5-150}{5} = 3$$

$$P(-z_\propto \leq Z < z_\propto) = 2(P(Z < z_\propto) - 0.5) = 2(0.99865 - 0.5) = 0.9973$$

The probability that variable X belongs to the interval $\langle 150 - 3 \cdot 5, 150 + 3 \cdot 5 \rangle$ is 0.9973. In general, the probability that variable X, which has normal distribution, belongs to the interval $\langle \mu - 3\sigma, \mu + 3\sigma \rangle$ is 0.9973.

The graphical interpretation of the probability $P_c = P(-x_\alpha \leq X \leq x_\alpha) = P(-z_\alpha \leq Z \leq z_\alpha)$ is shown in Fig. 5.3c for case (i).

5.3 t-STUDENT DISTRIBUTION

The t-Student distribution is mainly applied for reasoning about the mean.

If variable X has normal distribution $N(\mu, \sigma)$, and an n-element sample is drawn from the population of values of X, the variable:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

has t-Student distribution, shortly, t distribution with $\nu = n - 1$ degrees of freedom.

The probability distribution of variable t is described by the following probability density function:

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \text{ for } t \in \langle -\infty, \infty \rangle$$

where: Γ is the gamma function, ν are the degrees of freedom.

The PDF of the t-Student distribution has one parameter ν. This fact is represented using the following notation $t(\nu)$.

Examples of probability density functions of the t variable are shown in Fig. 5.4 for selected degrees of freedom $\nu = 1$, 15, and 35 together with the normal distribution $N(0,1)$ as a reference.

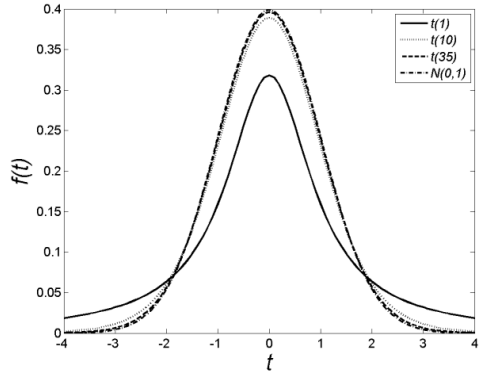


Figure 5.4 Examples of t-Student probability density functions.

The probability density function of t-Student distributions has a bell shape, as shown in Fig. 5.4 with the function being symmetric. The location of the function maximum is fixed, whereas its flatness depends on ν. With increasing degrees of freedom, the t-Student distribution approaches standard normal distribution. It is usually assumed that for $\nu > 30$, normal distribution shall be used instead of t-Student distribution.

The mean μ and variance σ^2 of variable t are calculated using the following formulas:

$$\mu = 0$$

$$\sigma^2 = \frac{\nu}{\nu - 2}$$

There are statistical tables available for t-Student distributions (see Appendix 2).

5.3.1 EXAMPLE

Problem. A variable has t-Student distribution with $\nu = 7$ degrees of freedom. What is the probability that the variable takes values which are

(a) greater than or equal to 2.365,
(b) belong to the interval $(-2.365, 2.365)$.

Solution. In order to solve the problem, statistical tables of t-Student distribution are needed (Appendix 2). The following are solutions for cases (a) and (b).

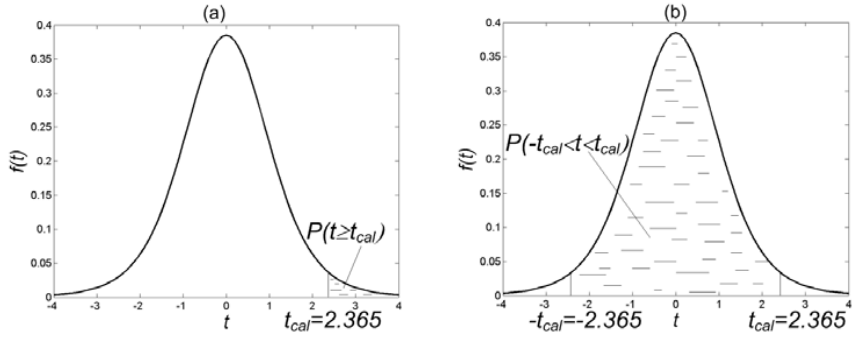


Figure 5.5 Graphical illustration of problem 5.3. 1.

(a) In order to use t tables, $t_{cal} = 2.365$ shall be substituted for $t_{\alpha,\nu}$. One reads the probability α, associated with $t_{\alpha,\nu} = 2.365$, i.e. $P(|t| \geq t_{\alpha,\nu})$ and the following is calculated:

$$P\left(t \geq t_{\propto,\nu}\right) = 0.5P\left(|t| \geq t_{\propto,\nu}\right) = 0.5P(|t| \geq 2.365) = 0.5 \cdot 0.05 = 0.025$$

The probability that t is greater than or equal to 2.365 is 0.025.
The graphical representation of the requested probability is shown in Fig. 5.5a.

(b) Knowing that $P(|t| \geq t_{\alpha,\nu})$ for $t_{\alpha,\nu} = 2.365$, the following is calculated:

$$P\left(|t| < t_{\propto,\nu}\right) = 1 - P\left(|t| \geq t_{\propto,\nu}\right) = 1 - P(|t| \geq 2.365) = 1 - 0.05 = 0.95$$

The probability that t belongs to the interval $(-2.365, 2.365)$ is 0.95.
The graphical representation of the requested probability is shown in Fig. 5.5b.

5.4 CHI SQUARE DISTRIBUTION

The χ^2 distribution is mainly applied for reasoning about the variance.

If variable X has normal distribution $N(\mu, \sigma)$ and an n-element sample is drawn from the population of values of X, the variable:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

has Chi square distribution with $\nu = n - 1$ degrees of freedom. The s^2 is the estimate of σ^2 based on the sample.

The probability distribution of variable χ^2 is described by the following probability density function:

$$f(\chi^2) = 0 \text{ , for } \chi^2 \leq 0$$

$$f(\chi^2) = \frac{\left(\frac{1}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)} (\chi^2)^{\frac{\nu}{2}-1} e^{-\frac{\chi^2}{2}}, \text{ for } \chi^2 > 0$$

where: Γ is the gamma function, ν are degrees of freedom.

The PDF of the Chi-square distribution has one parameter ν. This fact is represented using the following notation: $\chi^2(\nu)$.

Examples of the probability density function for the χ^2 variable are shown in Fig. 5.6 for the selected degrees of freedom $\nu = 5, 10$ and 35.

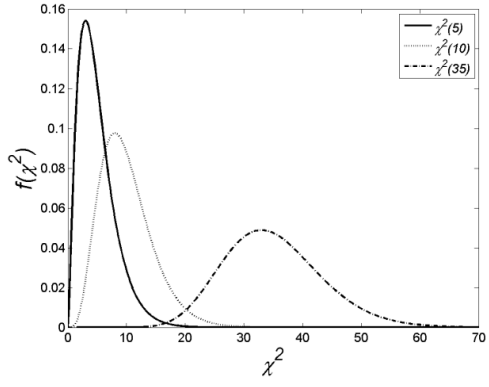


Figure 5.6 Examples of χ^2 probability density functions.

The probability density function of the χ^2 variable is asymmetric for small degrees of freedom, as shown in Fig. 5.6. With increasing degrees of freedom, the distribution loses its asymmetric character. It becomes quite well represented by the normal distribution for $\nu > 30$.

The mean μ and variance σ^2 of variable χ^2 are calculated by the following formulas:

$$\mu = \nu$$

$$\sigma^2 = 2\nu$$

There are statistical tables available for χ^2 distributions (Appendix 3).

5.4.1 EXAMPLE

Problem. A variable has χ^2 distribution with $\nu = 20$ degrees of freedom. What is the probability that values of the variable are

(a) greater than or equal to 10.851,

(b) less than 7.434.

Solution. In order to solve the problem, statistical tables of χ^2 distribution are needed (Appendix 3). The following are calculations for cases (a) and (b).

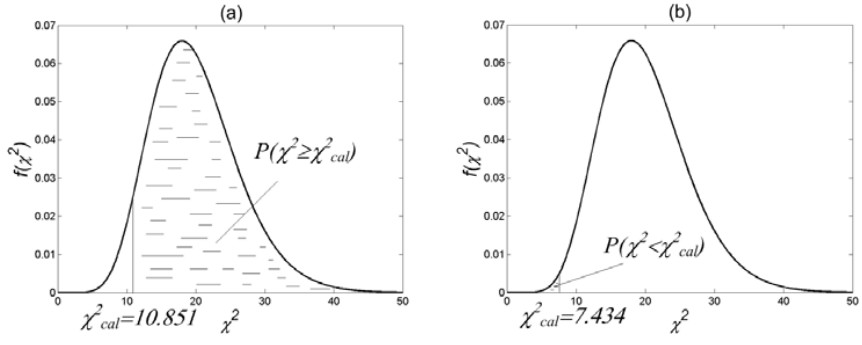


Figure 5.7 Graphical interpretation of problem 5.4.1.

(a) In order to use Chi^2 distribution tables, $\chi^2_{cal} = 10.851$ shall be substituted for $\chi^2_{\alpha,\nu}$. Then, one reads the probability α, associated with $\chi^2_{\alpha,\nu} = 10.851$, i.e. the probability $P(\chi^2 \geq \chi^2_{\alpha,\nu})$:

$$P(\chi^2 \geq \chi^2{}_{\alpha,\nu}) = P(\chi^2 \geq 10.851\ ) = 0.95$$

The probability that the value of the variable is greater than or equal to 10.851 is 0.95.

The graphical interpretation of the requested probability is shown in Fig. 5.7a.

(b) In order to use Chi^2 distribution tables, $\chi^2_{\alpha,\nu}$ is substituted with $\chi^2_{cal} = 7.434$. Next, one reads the probability α, associated with $\chi^2_{\alpha,\nu} = 7.434$, i.e. the probability $P(\chi^2 \geq \chi^2_{\alpha,\nu})$, and the following is calculated:

$$P(\chi^2 < \chi^2{}_{\alpha,\nu}) = 1 - P(\chi^2 \geq \chi^2{}_{\alpha,\nu}) = 1 - P(\chi^2 \geq 7.434\ ) = 1 - 0.995 = 0.005$$

The probability that the value of the variable is less than 7.434 is 0.005.

The graphical interpretation of the requested probability is shown in Fig. 5.7b.

5.5 F-SNEDECORE DISTRIBUTION

F-Snedecore distribution is mainly applied for comparing variances.

Let variable X have normal distribution $N(\mu_1, \sigma_1)$ in one population and normal distribution $N(\mu_2, \sigma_2)$ in another population. If a sample consisting of n_1 elements is drawn from the first population and a sample consisting of n_2 elements is drawn from the second population, the variable:

$$F = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}}$$

has F-Snedecore distribution with the following degrees of freedom $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$.

The probability distribution of variable F is described by the following probability density function:

$$f(F) = 0, \text{ for } F \leq 0$$

$$f(F) = \frac{\nu_1^{\frac{\nu_1}{2}} \nu_2^{\frac{\nu_2}{2}} \Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \frac{F^{\frac{\nu_1}{2}-1}}{(\nu_1 F+\nu_2)^{\frac{1}{2}(\nu_1+\nu_2)}}, \text{ for } F > 0$$

where: Γ is the gamma function, ν_1 and ν_2 are degrees of freedom.

The PDF of F-Snedecore distribution has two parameters ν_1 and ν_2. This fact is represented using the following notation: $F(\nu_1, \nu_2)$.

Plots of exemplary F-Snedecore distributions are shown in Fig. 5.8 for the selected pairs of degrees of freedom $F(5,5)$, $F(5,35)$, $F(35,5)$.

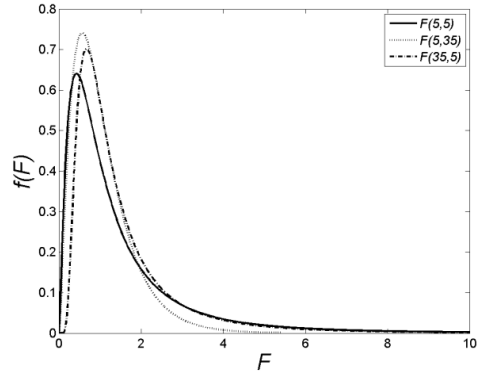


Figure 5.8 Examples of F-Snedecore probability density functions.

The probability density function of variable F is asymmetric for small degrees of freedom. With the increasing degrees of freedom the distribution loses its asymmetric character, as shown in Fig. 5.8.

The mean μ and variance σ^2 of variable F are calculated using the following formulas:

$$\mu = \frac{\nu_2}{\nu_2 - 2}$$

$$\sigma^2 = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}$$

There are statistical tables available for the F distribution (Appendix 4, 5). A single table refers to a fixed probability $\alpha = P(F \geq F_{\alpha,\nu_1,\nu_2})$ and all different pairs of degrees of freedom ν_1 and ν_2.

5.5.1 EXAMPLE

Problem. A variable has F-Snedecore distribution with $\nu_1 = 15$ and $\nu_2 = 23$ degrees of freedom. What value of variable F is neither reached nor exceeded with the probability $p = 0.95$.

Solution. The graphical interpretation of the problem is shown in Fig. 5.9.

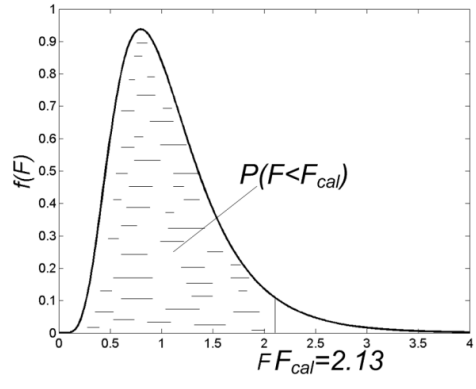


Figure 5.9 Graphical interpretation of problem 5.5.1.

In order to solve the problem, statistical tables of the F distribution are needed (Appendix 4, 5).

Tables of F_{α,ν_1,ν_2} distribution are constructed for the probability α that F exceeds or is equal to a certain value F_{α,ν_1,ν_2}, that is $\alpha = p(F \geq F_{\alpha,\nu_1,\nu_2})$.

To solve the problem, we are going to use the probability that a certain value of the F variable is neither reached nor exceeded $p\left(F < F_{\alpha,\nu_1,\nu_2}\right)$ in the following way:

$$\alpha = p\left(F \geq F_{\alpha,\nu_1,\nu_2}\right) = 1 - p\left(F < F_{\alpha,\nu_1,\nu_2}\right) = 1 - 0.95 = 0.05.$$

The obtained value of α indicates that one shall refer to F-Snedecore distribution which was constructed for $\alpha = 0.05$. In the table one reads $F_{0.05,15,23} = 2.13$, for $\nu_1 = 15$ and $\nu_2 = 23$ degrees of freedom.

The value of variable F, which is neither reached nor exceeded with the probability 0.05 is 2.13.

6 CONFIDENCE INTERVAL AND TOLERANCE INTERVAL

One of the basic engineering problems is to evaluate the confidence in the information about objects which is obtained by means of measurement/observation. Examples of simple problems of that kind are

- What is the probability that the observed/measured value of a variable, the mean of the variable or its spread do not deviate from their real values by more than a certain limit?
- What is the range of values that contains the true value of the variable, true mean, or true spread with the defined probability?

The above engineering problems may be translated into statistical problems of defining tolerance level and tolerance interval for a variable or defining a confidence level and confidence interval for the parameters of probability distribution of a random variable.

6.1 CONFIDENCE INTERVAL

Confidence level for variable V is the probability that values of the variable fall into the interval (a, b), which is called the confidence interval. The confidence level is usually denoted by P_α and the following holds:

$$P_\alpha = P(a < V < b) \quad \text{and} \quad P_\alpha = 1 - \alpha$$

where α is the significance level, as explained in §7.2. Confidence level refers to the parameters of statistical distribution of the observed variable X, e.g. the mean, the variance or standard deviation, and it does not refer to values of variable X. Those parameters are also random variables and have their probability distributions. They are represented by V in the above definition of confidence level.

The graphical interpretation of the confidence level and the confidence interval is shown in Fig. 6.1.

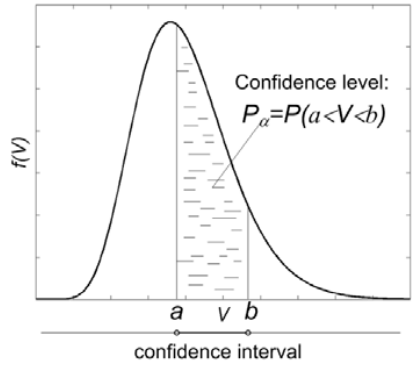


Figure 6.1 Graphical interpretation of the confidence level and the confidence interval.

Based on its definition, confidence level is the surface under the PDF between the left and right limit of the confidence interval. If the confidence interval is extended, the associated surface, i.e. the confidence level increases. The confidence level and confidence interval are related to each other. The confidence interval defines the confidence level in an unequivocal manner, but the opposite is not true. There are many possible confidence intervals for a defined confidence level. Statistics is interested in the narrowest of all confidence intervals at a particular confidence level.

In statistics, less confidence is associated with a narrower interval and more confidence is associated with a wider interval. This is counterintuitive, as one tends to associate confidence with something precise (narrow interval) rather than with something vague (wide interval). However, statistical confidence is a probability. For a defined PDF, a larger probability (surface under the PDF) is associated with longer intervals (range of X) while a smaller probability is associated with shorter intervals.

The most commonly used confidence levels are $P_\alpha = 0.95$, and $P_\alpha = 0.99$.

In engineering applications, there are two ways of using confidence level and confidence interval. Either confidence level is known and confidence interval is asked for or the interval is known and the confidence level is in question.

This chapter presents methods of calculating the

- confidence interval on the mean,
- confidence interval on the variance and standard deviation,
- tolerance interval.

6.2 CONFIDENCE INTERVAL ON THE MEAN

There are three possibilities of calculating confidence interval on the mean of variable X:

1. Variable X has normal distribution and variance σ^2 is known.

In this case the following variable is utilized for confidence interval calculation:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

As Z distribution is symmetric, the confidence interval stretches symmetrically around the mean. The confidence level is the probability P_α that

$$P_\alpha = P\left(-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}\right)$$

Therefore, the following transformations are allowed:

$$-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}$$

$$-Z_{\frac{\alpha}{2}} < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < Z_{\frac{\alpha}{2}}$$

$$\bar{X} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

The confidence interval on the mean of variable X is $\left(\bar{X} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$, at the confidence level P_α.

2. Variable X has normal distribution and variance σ^2 is unknown. It has to be estimated using s^2, based on a sample drawn from the population.

In this case, the following variable is utilized for confidence interval calculation:

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

As t -Student distribution is symmetric, the confidence interval stretches symmetrically around the mean. The confidence level is the probability P_α that

$$P_\alpha = P\left(-t_{\frac{\alpha}{2},\nu} < t < t_{\frac{\alpha}{2},\nu}\right)$$

Therefore, the following transformations are allowed:

$$-t_{\frac{\alpha}{2},\nu} < t < t_{\frac{\alpha}{2},\nu}$$

$$-t_{\frac{\alpha}{2},\nu} < \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} < t_{\frac{\alpha}{2},\nu}$$

$$\bar{X} - t_{\frac{\alpha}{2},\nu}\frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{\frac{\alpha}{2},\nu}\frac{s}{\sqrt{n}}$$

The confidence interval on the mean of variable X is $\left(\bar{X} - t_{\frac{\alpha}{2},\nu}\frac{s}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2},\nu}\frac{s}{\sqrt{n}}\right)$ at the confidence level P_α.

For big samples, i.e. $n > 30$, the solution converges with the solution described in possibility 1 since the t-Student distribution converges with the normal distribution for $n > 30$.

3. The variable X has unknown distribution and either variance σ^2 or its estimate s^2 is known.

In this case the following variable is utilized for confidence interval calculation:

$$Z = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

where s is the estimate of standard deviation based on an n element sample.

The calculation of confidence interval is identical as in the case when variable X has normal distribution and variance σ^2 is known. The confidence interval on the mean of variable X is $\left(\bar{X} - Z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \bar{X} + Z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right)$, at the confidence level P_α; however, the obtained solution has an approximate character. The size of the sample used for estimating $\bar{X}$, and s shall be big $(n > 120)$.

Examples of confidence interval calculation are given regarding cases 2 and 3. Case 1 occurs quite rarely in practice and the solution strategy is identical with case 3; therefore, it was not analyzed.

6.2.1 EXAMPLE

Problem. It is known that the monthly concentration of NO_x in city A has normal distribution. The monthly concentration of NO_x was measured in the city over one year and it was found that the average monthly concentration was 100 μg/m^3 with a standard deviation of 50 μg/m^3. What is the confidence interval on the mean monthly concentration of NO_x at the confidence level of 0.98?

Solution. The considered variable has normal distribution, but the parameters of the distribution are unknown. They were estimated based on an $n = 12$ element sample (12 monthly averages) and they are $\bar{X} = 100$ and $s = 50$. The problem falls into the category: the confidence interval on the mean of variable X which has a normal distribution in population and the variance σ^2 is unknown (case 2). Therefore, the formula describing the confidence interval on the mean is the following:

$$\bar{X} - t_{\frac{\alpha}{2},\nu}\frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{\frac{\alpha}{2},\nu}\frac{s}{\sqrt{n}}$$

The only missing value in the formula is $t_{\frac{\alpha}{2},\nu}$. It is found in the t-Student distribution table (Appendix 2) for $\alpha = 1 - P_\alpha = 1 - 0.98 = 0.02$, $\frac{\alpha}{2} = 0.01$ and $\nu = n - 1 = 12 - 1 = 11$. The missing value is $t_{0.01,11} = 3.106$. As a result of substitution

$$100 - 3.106\frac{50}{\sqrt{12}} < \mu < 100 + 3.106\frac{50}{\sqrt{12}}$$

$$55.17 < \mu < 144.83.$$

the confidence interval on the mean monthly concentration of NO_x in the city is $(53.17, 144.83)$ μg/m^3 at the confidence level $P = 0.98$.

6.2.2 EXAMPLE

Problem. The probability distribution of the daily SO_2 concentration in the city in winter is not known. The daily concentration of SO_2 was measured in the city over a period of 5 winter months. It was found that the average daily concentration was

60 μg/m^3 with a standard deviation of 20 μg/m^3. What is the confidence interval on the mean monthly concentration of SO_2 at the confidence level of 0.99.

Solution. The considered variable has unknown distribution. The mean value of variable $\bar{X} = 60$ and its standard deviation $s = 20$ were estimated based on an $n = 150$ element sample (30 daily average concentrations for 5 months). The problem falls within the category: the variable X has unknown distribution in population and the estimate of variance s^2 is available, the size of sample is big (case 3). The following formula describes the confidence interval on the mean:

$$\bar{X} - Z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}} < \mu < \bar{X} + Z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$$

The only missing value in the formula is $Z_{\frac{\alpha}{2}}$. It is found in the Z table (Appendix 1) for $\Phi(Z) = 0.5 + 0.5P_\alpha = 0.5 + 0.5 \cdot 0.99 = 0.995$. The value is $Z_{0.005} = 2.58$. As a result of substitution

$$60 - 2.58\frac{20}{\sqrt{150}} < \mu < 60 + 2.58\frac{20}{\sqrt{150}}$$

$$55.79 < \mu < 64.21$$

the confidence interval on the average daily concentration of SO_2 is $(55.79, 64.21)$ μg/m^3 at the confidence level of $P_\alpha = 0.99$.

6.3 CONFIDENCE INTERVAL ON THE VARIANCE

The assumption about the normality of variable X is required for calculating the confidence interval on the variance.

The following variable is utilized for the confidence interval calculation:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

As the χ^2 distribution is asymmetric, the confidence interval is not symmetric with respect to the mean. The confidence level is the probability P_α that

$$P_\alpha = P\left(\chi^2_{1-\frac{\alpha}{2},\nu} < \chi^2 < \chi^2_{\frac{\alpha}{2},\nu}\right)$$

Therefore, the following transformations are allowed:

$$\chi^2_{1-\frac{\alpha}{2},\nu} < \chi^2 < \chi^2_{\frac{\alpha}{2},\nu}$$

$$\chi^2_{1-\frac{\alpha}{2},\nu} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{\frac{\alpha}{2},\nu}$$

$$\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},\nu}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},\nu}}$$

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},\nu}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},\nu}}}$$

The confidence interval on the variance of variable X is $\left(\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},\nu}}, \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},\nu}}\right)$ at the confidence level P_α.

The confidence interval on the standard deviation of variable X is $\left(\sqrt{\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},\nu}}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},\nu}}}\right)$ at the confidence level P_α.

6.3.1 EXAMPLE

Problem. The temperature control system is expected to stabilize the temperature around 50 ± 1 °C. In order to evaluate the performance of the system, the temperature was measured in the course of $n = 15$ independent measurements. The obtained results are provided in Table 6.1. What is the confidence interval on the spread of temperature at the confidence level of 0.95? Does the confidence interval satisfy the requirements? It is correct to assume that the temperature has normal distribution?

Table 6.1 Results of temperature measurements.

Measurement	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Temperature	50	50.1	50.3	49.8	50	50.6	48.7	49.1	50.4	50.1	51	49.9	50.7	49	50.3

Solution. The standard deviation may be used as the measure of spread of temperature values. Based on the data provided in Table 6.1 and the relevant formula (see §3.2), the estimate of standard deviation is

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x_i})}{n-1}} = 0.64$$

As the considered variable X (here, temperature) has normal distribution, the following formula describes the confidence interval for standard deviation:

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},\nu}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},\nu}}}$$

The missing values $\chi^2_{\frac{\alpha}{2},\nu}$ and $\chi^2_{1-\frac{\alpha}{2},\nu}$ are found in the Chi square distribution table (Appendix 3) for $\frac{\alpha}{2} = 0.5(1 - P_\alpha) = 0.025$, $1 - \frac{\alpha}{2} = 0.975$, and $\nu = n - 1 = 14$. They are $\chi^2_{0.025,14} = 26.119$ and $\chi^2_{0.975,14} = 5.629$.

$$\sqrt{\frac{(15-1)0.64^2}{26.119}} < \sigma < \sqrt{\frac{(15-1)0.64^2}{5.629}}$$

$$0.47 < \sigma < 1.1$$

The confidence interval on the spread of temperature is $(0.47,1.1)$, at the confidence level 0.95. It does not fully satisfy the requirements because the accepted value of spread was 1, which is less than the right limit of the confidence interval.

In fact, the assumption concerning normality of temperature should be confirmed using the normality test (see §7.7).

6.4 TOLERANCE INTERVAL

Tolerance level and tolerance interval are calculated for variable X. These are notions corresponding to confidence level and confidence interval which refer to parameters of the statistical distribution of variable X. Tolerance level q is a probability described by the following formula:

$$q = P(F(X2) - F(X1) \geq Q)$$

where $F(X2)$ and $F(X1)$ are values of the cumulative distribution function of variable X. $X1$ is the lower tolerance limit while $X2$ is the upper tolerance limit.

The tolerance interval $\langle X1, X2 \rangle$ hosts at least $Q \cdot 100$ % values of variable X with the probability q. Q is the smallest fraction of values of X which fall into the tolerance interval with the probability q.

In practical applications, it is most frequently assumed that variable X has normal distribution and in such case the tolerance interval for a single value of variable X is the following:

$$\bar{x} - K_{n,q,Q}s \leq X \leq \bar{x} + K_{n,q,Q}s$$

where $\bar{x}$ is the estimate of the mean of X, based on an n-element sample, s is the estimate of the standard deviation of X, based on an n-element sample and $K_{n,q,Q}$ is available in statistical tables. Tables are available for the most frequently used values of n, q and Q (see Appendix 6).

6.4.1 EXAMPLE

Problem. It is known that the length X of screws delivered by the production line follows a normal distribution. Based on a randomly selected sample of 70 screws,

the average screw length is $\bar{X} = 10$ mm and the standard deviation of screw length is $s = 0.2$ mm. What is the tolerance interval which hosts at least $99\ \%$ of randomly selected screws at the tolerance level $q = 0.9$?

Solution. The tolerance interval is to be found for a single value of the variable which is known to have normal distribution. The estimates of parameters of distribution are known based on the sample of known size. The tolerance interval is given by the following formula:

$$\bar{x} - K_{n,q,Q}s \leq X \leq \bar{x} + K_{n,q,Q}s$$

The value of K is found in K-value tables (Appendix 6) for $n = 70$, $q = 0.9$ and $Q = 0.99$. $K_{70,0.9,0.99} = 2.92$. After substitution:

$$10 - 2.92 \cdot 0.2 \leq X \leq 10 + 2.92 \cdot 0.2$$

$$9.42 \leq X \leq 10.58$$

The $99\ \%$ tolerance interval for the screw length is $\langle 9.42, 10.58 \rangle$ at the tolerance level 0.9.

In other words, one can be $90\ \%$ sure that $99\ \%$ of screws have their lengths in the tolerance interval $\langle 9.42, 10.58 \rangle$ mm.

7 STATISTICAL HYPOTHESES AND THEIR TESTING

A number of engineering tasks consist of comparing objects including comparing an object with a reference, comparing a single object with itself in different conditions and comparing different objects.

Various kinds of comparisons may be translated into statistical hypotheses. Especially useful hypotheses include

- hypothesis on the mean value of a variable; this is applicable for comparing average states of objects,
- hypothesis on the variance of a variable; this may be used for comparing the variability of object states.

The methodology of statistical hypothesis testing is presented in this chapter regarding

- test on one mean,
- test on two means,
- test on the variance,
- test on two variances,
- normality test.

7.1 STATISTICAL HYPOTHESIS

A **statistical hypothesis** is a supposition concerning the statistical distribution of a variable. There are two main types of suppositions. Suppositions of the first type are called **parametric hypotheses.** They refer to parameters of distribution for the observed variable, e.g. the mean and the variance. These hypotheses require the preliminary assumption about the kind of distribution of the original variable. Suppositions of the second type are called **nonparametric hypotheses**. Most important classes of nonparametric hypotheses refer to the randomness of a sample, the independence of variables or the kind of variable distribution. These hypotheses do not require any assumption about the kind of distribution of the original variable.

A particular statistical hypothesis actually consists of a pair of complementary hypotheses. These are the null hypothesis and the alternative hypothesis. The supposition called the **null hypothesis** is indicated by H_0. It usually states that two entities are equal. The contradictory supposition, denoted with H_A, is called the **alternative hypothesis**. It usually states that two entities are unequal or that one entity is greater than the other. These two cases represent two–sided and one–sided alternative hypotheses, respectively.

7.2 STATISTICAL HYPOTHESIS TESTING

Statistical hypotheses are subject to testing. The tools designed for testing statistical hypotheses are called **statistical tests**. There are parametric and nonparametric tests corresponding to the types of statistical hypotheses. The main groups of parametric statistical tests are tests on the mean, tests of variance and tests for proportion. The main groups of nonparametric statistical tests are tests of the randomness of a sample, tests on the independence of variables and the goodness of fit test between the probability distribution of a variable and another distribution.

The basis for statistical hypothesis testing is a random sample of variable values drawn from the population. The statistical test is a set of rules which allow for the acceptance or rejection of a hypothesis for a particular sample. In reality, hypotheses are either false or true. However, statistical testing is not able to provide such judgment. With statistical methodology, one can either reject the null hypothesis or accept it.

The decision concerning the null hypothesis is not absolute. It takes into account the possibility that a null hypothesis which is actually true is rejected. The individual testing the hypothesis has to decide about the acceptable probability of rejecting a true hypothesis. This probability is called the **significance level** and it is denoted with α. The rejection of a true hypothesis is called a **I type error**. This error shall be low. Therefore, the typically used values of α are 0.01, 0.05 and 0.1. The significance level is selected arbitrarily for testing particular hypotheses.

The **test statistic** is used for testing statistical hypotheses. The test statistic is a variable V, which has a known distribution $f(V)$, if the null hypothesis is true. The value V_{cal} of the test statistic is calculated for a random sample of the original variable X. The general rule is that the more extreme is V_{cal} regarding the statistical distribution of V, the more likely that the null hypothesis is rejected. This is because the null hypothesis assumes that V_{cal} well represents the V distribution. Therefore, if the null hypothesis is true, V_{cal} is expected to match well the V distribution and not be too extreme.

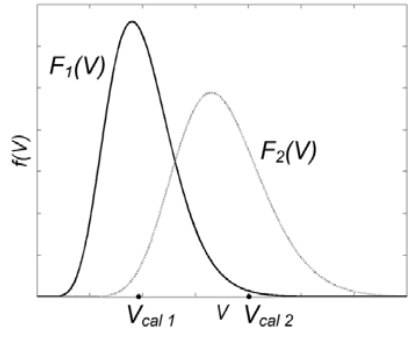


Figure 7.1 Illustration of the conception of a test statistic and its distribution in statistical hypothesis testing.

This idea is illustrated in Fig. 7.1. It is quite likely that if the test statistic V_{cal} has the distribution $f_1(V)$, the calculated value of the test statistic is V_{cal1}. Contrarily, it is quite unlikely that if the test statistic V_{cal} has the distribution $f_1(V)$, the calculated value of the test statistic is V_{cal2}. Therefore, if the test statistic takes the value V_{cal1}, the hypothesis that V_{cal} originates from the distribution $f_1(V)$ would rather be accepted. On the other hand, if the test statistic takes the value V_{cal2}, the hypothesis that V_{cal} originates from the distribution $f_1(V)$ would rather be rejected. The reader is invited to carry out the analogue reasoning concerning $f_2(V)$.

There are two possible approaches in the domain of statistical hypothesis testing. The classical approach utilizes the conception of **critical interval**. It is elegant and well suited for manual calculations. The second approach utilizes the conception of the ***p*-value**, which is also called the **critical significance level**. The possibility of using this approach is thanks to the development of computing.

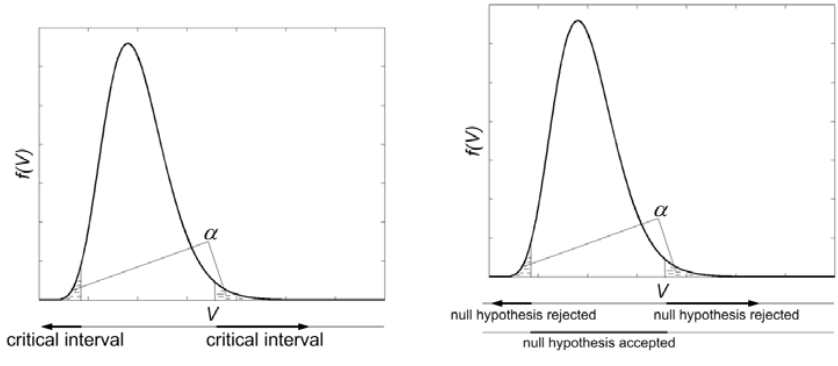


Figure 7.2 Illustration of the conception of the critical interval in statistical hypothesis testing.

In the framework of the first approach, the status of the hypothesis is judged based on checking whether the value V_{cal} belongs to the critical interval. The critical interval is the interval of extreme values of the test statistic V. The probability that values of variable V belong to the critical interval is equal to the significance level. Therefore, the size of the critical interval depends on α. If the calculated value of the test statistic V_{cal} belongs to the critical interval, the null hypothesis is rejected at the significance level α. If the calculated value of the test statistic V_{cal} remains outside the critical interval, the null hypothesis is accepted at the significance level α. The concept of using a critical interval for testing a statistical hypothesis is illustrated in Fig. 7.2.

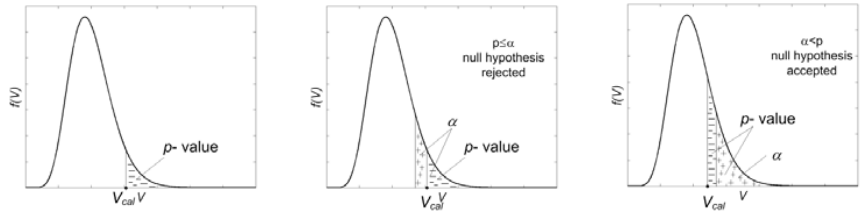


Figure 7.3 Illustration of the conception of the *p*-value in statistical hypothesis testing.

In the framework of the second approach, the status of the null hypothesis is judged based on a comparison between the p-value and the significance level α. The p-value is the probability that variable V takes values at least as extreme as the calculated value of the test statistic V_{cal}. In other words, the p-value is the smallest probability of null hypothesis rejection. The significance level α is actually the largest acceptable probability of null hypothesis rejection. In case the p-value is smaller than α, the probability that variable V takes values at least as extreme as V_{cal} is lower than acceptable. In other words, V_{cal} is too extreme. Therefore, the null hypothesis is rejected. Contrarily, if the p-value is larger than α, the probability that values of variable V are at least as extreme as V_{cal} is greater than acceptable. In other words, V_{cal} is not too extreme. In such case the null hypothesis is accepted. The concept of using the p-value for testing a statistical hypothesis is illustrated in Fig. 7.3.

7.3 TEST ON ONE MEAN

The null hypothesis H_0 in the case of a test on one mean states that the mean μ of variable X in the general population is equal to a defined reference value μ_0. The formal notation of the null hypothesis is the following:

$$H_0\colon\ \mu = \mu_0$$

The null hypothesis is tested versus one of three different alternative hypotheses:

I.	$H_a\colon\ \mu \neq \mu_0$	two–sided hypothesis
II.	$H_a\colon\ \mu > \mu_0$	one–sided hypothesis
III.	$H_a\colon\ \mu < \mu_0$	one–sided hypothesis

The form of the test statistic depends on the assumption concerning the distribution of variable X. In this book two cases are considered: (1) variable X has normal distribution $N(\mu, \sigma)$ and σ is unknown(§7.3.1), (2) variable X has unknown distribution (§7.3.2).

The criteria of null hypothesis rejection depend on the kind of alternative hypothesis which is considered together with the null hypothesis.

7.3.1 VARIABLE X HAS NORMAL DISTRIBUTION AND σ IS UNKNOWN

The test statistic is the following:

$$t_{cal} = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

where: $\bar{X}$ is the estimate of μ and it is calculated based on a random sample, μ_0 is the reference value, s is the estimate of σ and it is calculated based on a random sample, n is the number of elements in the random sample.

If the null hypothesis is true, the test statistic t_{cal} has t-Student distribution, with $\nu = n - 1$ degrees of freedom.

I. Criterion of null hypothesis rejection on one mean versus $H_a\colon\ \mu \neq \mu_0$.

- Criterion of the critical interval

The criterion of null hypothesis rejection is the following:

$$P\left(t_{\frac{\alpha}{2},\nu} \leq |t|\right) = \alpha$$

$$P\left(\ t \leq -t_{\frac{\alpha}{2},\nu}\ V\ \ t_{\frac{\alpha}{2},\nu} \leq t\right) = \alpha$$

Therefore, the critical interval is $t \in \left(-\infty, -t_{\frac{\alpha}{2},\nu}\right\rangle \cup \left\langle t_{\frac{\alpha}{2},\nu}, \infty\right)$.

The null hypothesis is rejected at the significance level α if the calculated value of the test statistic t_{cal} belongs to the critical interval, i.e. if the following holds: $t_{cal} \in \left(-\infty, -t_{\frac{\alpha}{2},\nu}\right\rangle \cup \left\langle t_{\frac{\alpha}{2},\nu}, \infty\right)$.

The null hypothesis is accepted at the significance level α if the calculated value of the test statistic t_{cal} falls outside the critical interval, i.e. if the following is true: $t_{cal} \in \left(-t_{\frac{\alpha}{2},\nu}, t_{\frac{\alpha}{2},\nu}\right)$.

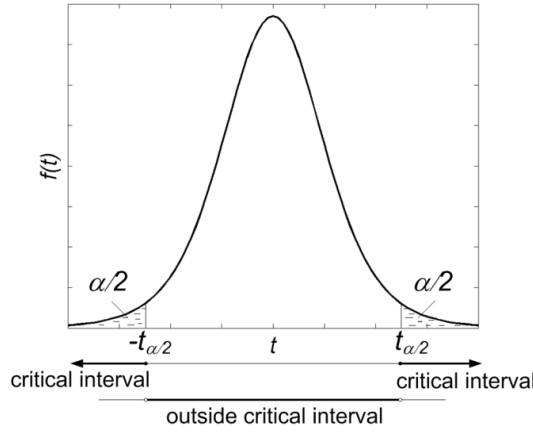


Figure 7.4 Graphical interpretation of null hypothesis rejection based on the critical interval criterion. Test on one mean. $\boldsymbol{H_0}$: $\boldsymbol{\mu = \mu_0}$. $\boldsymbol{H_a}$: $\boldsymbol{\mu \neq \mu_0}$.

The graphical interpretation of null hypothesis rejection based on the critical interval criterion is shown in Fig. 7.4.

- Criterion of the p-value

The criterion of null hypothesis rejection is the following: $p \leq \alpha \equiv \frac{p}{2} \leq \frac{\alpha}{2}$, where $p = P(|t| \geq t_{cal})$ is the probability that the absolute value of the t variable, which has t-Student distribution with $\nu = n - 1$ degrees of freedom, is greater than or equal to the calculated value of the test statistic t_{cal}.

The null hypothesis is rejected if the significance level α is greater than or equal to p.

The null hypothesis is accepted if the significance level α is less than p.

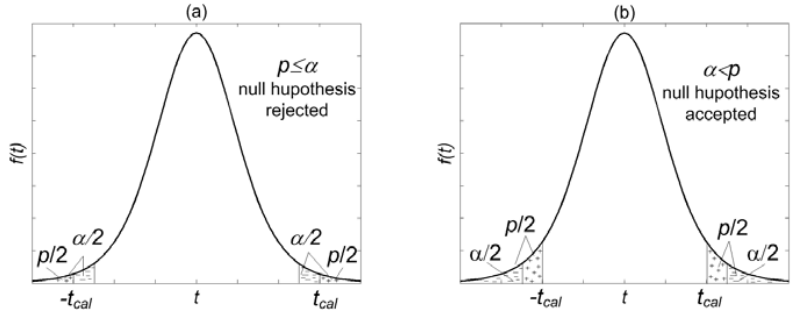


Figure 7.5 Graphical interpretation of null hypothesis rejection based on the $\boldsymbol{p}$-value criterion. Test on one mean. $\boldsymbol{H_0}$: $\boldsymbol{\mu = \mu_0}$. $\boldsymbol{H_a}$: $\boldsymbol{\mu \neq \mu_0}$.

The graphical interpretation of the p-value criterion of null hypothesis rejection is shown in Fig. 7.5.

II. Criterion of null hypothesis rejection on one mean versus H_a: $\mu > \mu_0$.

- Criterion of the critical interval

The criterion of null hypothesis rejection is $P\left(t \geq t_{\alpha,\nu}\right) = \alpha$. Therefore, the critical interval is $t \in \langle t_{\alpha,\nu}, \infty\rangle$.

The null hypothesis is rejected at the significance level α if the calculated value of test statistic t_{cal} belongs to the critical interval, i.e. if the following holds: $t_{cal} \in \langle t_{\alpha,\nu}, \infty\rangle$.

There is no reason for rejecting the null hypothesis at the significance level α, if the calculated value of test statistic t_{cal} remains outside the critical interval, i.e. if the following is true: $t_{cal} \in \langle 0, t_{\alpha,\nu}\rangle$.

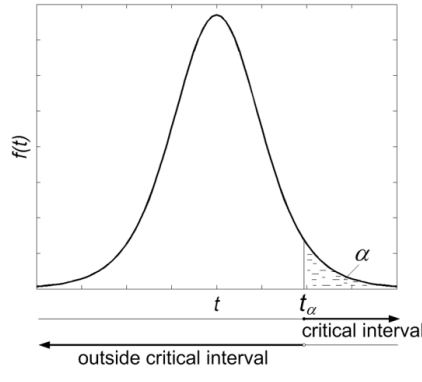


Figure 7.6 Graphical interpretation of null hypothesis rejection based on the critical interval criterion. Test on one mean. $\boldsymbol{H_0}$: $\boldsymbol{\mu = \mu_0}$. $\boldsymbol{H_a}$: $\boldsymbol{\mu > \mu_0}$.

The graphical interpretation of the critical interval criterion of null hypothesis rejection is shown in Fig. 7.6.

- Criterion of the p-value

The criterion of null hypothesis rejection is the following: $p \leq \alpha$, where $p = P(t \geq t_{cal})$ is the probability that the t variable, which has t-Student distribution with $\nu = n - 1$ degrees of freedom, is greater than or equal to the calculated value of the test statistic t_{cal}:

The null hypothesis is rejected if the significance level α is greater than or equal to p.

The null hypothesis is accepted if the significance level α is less than p.

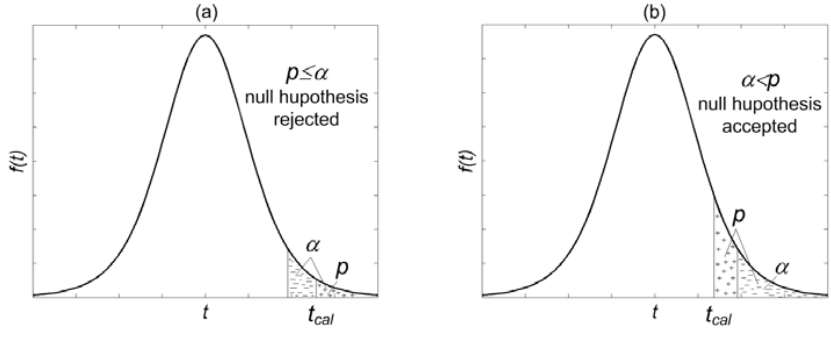


Figure 7.7 Graphical interpretation of null hypothesis rejection based on the $\boldsymbol{p}$-value criterion. Test on one mean. $\boldsymbol{H_0}$: $\boldsymbol{\mu = \mu_0}$. $\boldsymbol{H_a}$: $\boldsymbol{\mu > \mu_0}$.

The graphical interpretation of the p-value criterion of null hypothesis rejection is shown in Fig. 7.7.

III. Criterion of null hypothesis rejection on one mean versus H_a: $\mu < \mu_0$.

- Criterion of the critical interval

The criterion of null hypothesis rejection is $P(t \leq -t_{\alpha,\nu}) = \alpha$. Therefore, the critical interval is $t \in (-\infty, -t_{\alpha,\nu}\rangle$.

The null hypothesis is rejected at the significance level α if the calculated value of test statistic t_{cal} belongs to the critical interval, i.e. if the following holds: $t_{cal} \in (-\infty, -t_{\alpha,\nu}\rangle$.

There is no reason for rejecting the null hypothesis at the significance level α, if the calculated value of test statistic t_{cal} remains outside the critical interval, i.e. if the following is true: $t_{cal} \in (-t_{\alpha,\nu}, 0\rangle$.

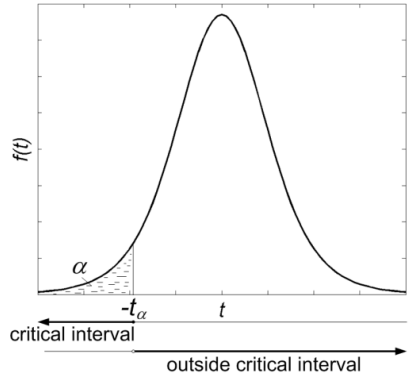


Figure 7.8 Graphical interpretation of null hypothesis rejection based on the critical interval criterion. Test on one mean. $\boldsymbol{H_0}$: $\boldsymbol{\mu = \mu_0}$. $\boldsymbol{H_a}$: $\boldsymbol{\mu < \mu_0}$.

The graphical interpretation of the critical interval criterion of null hypothesis rejection is shown in Fig. 7.8.

- Criterion of the p-value

The criterion of null hypothesis rejection is the following: $p \leq \alpha$, where $p = P(t \leq t_{cal})$ is the probability that the t variable, which has t-Student distribution with $\nu = n - 1$ degrees of freedom, is less than the calculated value of test statistic t_{cal}.

The null hypothesis is rejected if the significance level α is greater than or equal to p.

The null hypothesis is accepted if the significance level α is less than p.

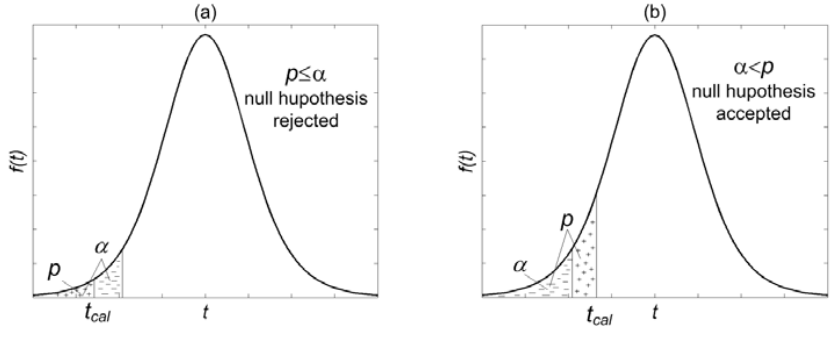


Figure 7.9 Graphical interpretation of null hypothesis rejection based on the $\boldsymbol{p}$-value criterion. Test on one mean. $\boldsymbol{H_0}$: $\boldsymbol{\mu = \mu_0}$. $\boldsymbol{H_a}$: $\boldsymbol{\mu < \mu_0}$.

The graphical interpretation of the p-value criterion of null hypothesis rejection is shown in Fig. 7.9.

7.3.1.1 EXAMPLE

Problem. There were $n = 7$ measurements of pressure inside the combustion chamber of a rocket engine. The measurement results are shown in Table 7.1.

Table 7.1 Results of pressure measurements inside the combustion chamber of an engine.

number	1	2	3	4	5	6	7
pressure/ kPa	3123.41	3075.35	2973.36	3030.24	3108.70	3177.34	3098.89

It is known that the pressure has normal distribution. Is the mean pressure inside the chamber equal to 3000 kPa at the significance level $\alpha = 0.01$?

Solution. The problem may be solved using a test on one mean regarding variable X, which is the pressure inside the combustion chamber of an engine. It is worth to consider the null hypothesis which states that the average pressure is equal to 3000 kPa, $H_0: \mu = 3000$. The null hypothesis is tested versus the two-sided alternative hypothesis that the mean pressure is different than 3000 kPa, $H_\alpha: \mu \neq 3000$. The distribution of variable X is normal and parameters of the distribution are estimated based on measurement results in the following way:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 3083.90$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = 66.21$$

The corresponding test statistic is the following:

$$t_{cal} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{3083.90 - 3000}{\frac{66.21}{\sqrt{7}}} = 3.35$$

- The criterion of null hypothesis rejection based on the critical interval.

Considering the two-sided alternative hypothesis, the critical interval is described by the following formula: $\left(-\infty, -t_{\frac{\alpha}{2},\nu}\right\rangle \cup \left\langle t_{\frac{\alpha}{2},\nu}, \infty\right)$. $t_{\frac{\alpha}{2},\nu}$ is found in statistical tables of t -Student distribution (Appendix 2) for $\alpha = 0.01$ and $\nu = n - 1 = 6$. Numerically, the critical interval is $(-\infty, -3.707\rangle \cup \langle 3.707, \infty)$.

The value of test statistic $t_{cal} = 3.35$ is located outside the critical interval; therefore, the null hypothesis is accepted at the significance level $\alpha = 0.01$.

- The criterion of null hypothesis rejection based on the p-value.

The p-value was calculated using the T.DISTRIBUTION function available in Excel. Considering $t_{cal} = 3.35$, the associated $p = 0.0154$.

The value of $p = 0.0154$ is greater than the value of $\alpha = 0.01$; therefore, the null hypothesis is accepted at the significance level 0.01.

Based on the obtained results of hypothesis testing, the engineer has good reason to claim that the average pressure in the combustion chamber is equal 3000 kPa at the significance level $\alpha = 0.01$.

7.3.2 VARIABLE X HAS UNKNOWN DISTRIBUTION

If the probability distribution of variable X is unknown and one needs to test the hypothesis on one mean, the size of the random sample should be big. It is recommended that the number of sampled values of X exceeds $20 - 30$ elements.

The test statistic is the following:

$$Z_{cal} = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

where: $\bar{X}$ is the estimate of μ and is calculated based on a random sample, μ_0 is the reference value, s is the estimate of σ and is calculated based on a random sample, n is the number of elements in the random sample.

If the null hypothesis is true, the test statistic Z_{cal} has normal distribution $N(0,1)$ with the $\mu = 0$ and the $\sigma = 1$.

I. Criterion of rejection of null hypothesis on one mean versus H_a: $\mu \neq \mu_0$.

- Criterion of critical interval

The criterion of null hypothesis rejection is the following:

$$P\left(z_{\frac{\alpha}{2}} \leq |Z|\right) = \alpha$$

$$P\left(Z \leq -z_{\frac{\alpha}{2}} \; V \; z_{\frac{\alpha}{2}} \leq Z\right) = \alpha$$

Therefore, the critical interval is $Z \in \left(-\infty, -z_{\frac{\alpha}{2}}\right\rangle \cup \left\langle z_{\frac{\alpha}{2}}, \infty\right)$.

The null hypothesis is rejected at the significance level α, if the calculated value of test statistic Z_{cal} belongs to the critical interval, i.e. if the following holds:

$$Z_{cal} \in \left(-\infty, -z_{\frac{\alpha}{2}}\right\rangle \cup \left\langle z_{\frac{\alpha}{2}}, \infty\right).$$

The null hypothesis is accepted at the significance level α if the calculated value of test statistic Z_{cal} falls outside the critical interval, i.e. if the following is true:

$$Z_{cal} \in \left(-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}}\right).$$

For the graphical interpretation of the critical interval criteria of null hypothesis rejection refer to Fig. 7.4. While analyzing replace t with Z and ignore v. The

principle of interpretation is identical in case of t-Student distribution and normal distribution as both are symmetric.

- Criterion of the p-value

The criterion of null hypothesis rejection is

$$p \leq \alpha \equiv \frac{p}{2} \leq \frac{\alpha}{2}, \text{ where } p = P(|Z| \geq z_{cal})$$

is the probability that the absolute value of the Z variable is greater than or equal to the calculated value of the test statistic z_{cal}.

The null hypothesis is rejected if the significance level α is greater than or equal to p.

The null hypothesis is accepted if the significance level α is less than p.

For the graphical interpretation of the p-value criterion of null hypothesis rejection refer to Fig. 7.5. While analyzing replace t with Z and ignore v. The principle of interpretation is identical in case of t-Student distribution and normal distribution as both are symmetric.

II. Null hypothesis rejection criterion on one mean versus H_a: $\mu > \mu_0$.

- Criterion of the critical interval

The criterion of null hypothesis rejection is $P(Z \geq z_\alpha) = \alpha$. Therefore, the critical interval is$Z \in \langle z_\alpha, \infty)$.

The null hypothesis is rejected at the significance level α, if the calculated value of test statistic Z_{cal} belongs to the critical interval, i.e. if the following holds: $Z_{cal} \in \langle z_\alpha, \infty)$.

The null hypothesis is accepted at the significance level α, if the calculated value of test statistic z_{cal} falls outside the critical interval, i.e. if the following is true: $Z_{cal} \in (0, z_\alpha)$.

For the graphical interpretation of the critical interval criterion of null hypothesis rejection refer to Fig. 7.6. While analyzing replace t with Z and ignore v. The principle of interpretation is identical in case of t-Student distribution and normal distribution as both are symmetric.

- Criterion of the p-value

The criterion of null hypothesis rejection is the following:

$$p \leq \alpha, \text{ where } p = P(Z \geq z_{cal})$$

is the probability that the Z variable is greater than or equal to the calculated value of test statistic z_{cal}.

The null hypothesis is rejected if the significance level α is greater than or equal to p.

The null hypothesis is accepted if the significance level α is less than p.

For the graphical interpretation of the critical interval criteria of null hypothesis rejection refer to Fig. 7.7. While analyzing replace t with Z and ignore v. The

principle of interpretation is identical in case of t-Student distribution and normal distribution as both are symmetric.

III. Criterion of null hypothesis rejection on one mean versus H_a: $\mu < \mu_0$.

- Criterion of the critical interval

The criterion of null hypothesis rejection is $P(Z \leq -z_\alpha) = \alpha$. Therefore, the critical interval is $Z \in (-\infty, -z_\alpha\rangle$.

The null hypothesis is rejected at the significance level α, if the calculated value of test statistic Z_{cal} belongs to the critical interval, i.e. if the following holds: $Z_{cal} \in (-\infty, -z_\alpha\rangle$.

There is no reason for rejecting the null hypothesis at the significance level α if the calculated value of test statistic Z_{cal} falls outside the critical interval, i.e. if the following is true: $Z_{cal} \in (-z_\alpha, 0)$.

For the graphical interpretation of the critical interval criteria of null hypothesis rejection refer to Fig. 7.8. While analyzing replace t with Z and ignore v. The principle of interpretation is identical in case of t-Student distribution and normal distribution as both are symmetric.

- Criterion of the p-value

The criterion of null hypothesis rejection is the following:

$$p \leq \alpha, \text{ where } p = P(Z \leq z_{cal})$$

is the probability that the Z variable is less than or equal to the calculated value of test statistic z_{cal}.

The null hypothesis is rejected if the significance level is greater than or equal to p.

The null hypothesis is accepted if the significance level is less than p.

For the graphical interpretation of the critical interval criteria of null hypothesis rejection refer to Fig. 7.9. While analyzing replace t with Z and ignore v. The principle of interpretation is identical in case of t-Student distribution and normal distribution as both are symmetric.

7.3.2.1 EXAMPLE

Problem. The absence of workers was investigated in a huge production factory. A random sample of $n = 100$ people was selected for the study. It was found that the average leave duration was $\bar{x} = 35$ days and the standard deviation of leave duration was $s = 17$ days in that sample. Is it allowed to conclude that the average leave duration in the considered factory was longer than 1 month (31 days) at the significance level $\alpha = 0.05$?

Solution. It is possible to solve the problem using a test on one mean regarding variable X, which is the leave duration. It is worth to consider the null hypothesis, which states that the average leave duration is equal 35 days, H_0: $\mu = 35$. The null

hypothesis shall be tested versus the one-sided alternative hypothesis that the leave duration is greater than 31 days, $H_a: \mu > 31$. The distribution of variable X is unknown, but the size of the sample is big. Estimates of μ and σ are available.

The corresponding test statistic is the following:

$$z_{cal} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{35 - 31}{\frac{17}{\sqrt{100}}} = 2.35$$

- The criterion of null hypothesis rejection based on the critical interval.

Considering the one-sided alternative hypothesis, the critical interval is $\langle z_\alpha, \infty)$. Based on statistical Z tables (Appendix 1), $z_\alpha = 1.64$ for $\alpha = 0.05$. Therefore, numerically the critical interval is $\langle 1.64, \infty)$.

The value of test statistic $z_{cal} = 2.35$ is located inside the critical interval. Therefore, the null hypothesis is rejected in favor of the alternative hypothesis at the significance level $\alpha = 0.05$.

- The criterion of null hypothesis rejection based on the p-value.

The p-value was calculated using the Z.DISTRIBUTION function available in Excel. Considering $z_{cal} = 2.35$, the associated $p = 0.0094$.

The value of $p = 0.0094$ is less than the value of $\alpha = 0.05$; therefore, the null hypothesis is rejected in favor of the alternative hypothesis at the significance level 0.05.

Based on the obtained results of hypothesis testing, a manager at the factory has good reason to claim that the average leave duration in the company was longer than 1 month, at the significance level $\alpha = 0.05$.

7.4 TEST ON TWO MEANS

The null hypothesis H_0 in the case of the test on two means states that the mean μ_1 of variable X_1 is equal to the mean μ_2 of variable X_2. The formal notation of the null hypothesis is the following:

$$H_0: \mu_1 = \mu_2$$

The null hypothesis is tested versus one of three different alternative hypotheses:

I.	$H_A: \mu_1 \neq \mu_2$	two–sided hypothesis
II.	$H_A: \mu_1 > \mu_2$	one–sided hypothesis
III.	$H_A: \mu_1 < \mu_2$	one–sided hypothesis

The selection of the test statistic depends on the assumption concerning the distribution of variable X. In this book two cases are considered: (1) variable X_1 has normal distribution $N(\mu_1, \sigma_1)$, variable X_2 has normal distribution $N(\mu_2, \sigma_2)$

and variances σ_1 and σ_2 are unknown (§7.4.1), (2) variables X_1 and X_2 have unknown distributions (§7.4.2).

The criteria of null hypothesis rejection depend on the kind of alternative hypothesis considered together with the null hypothesis.

7.4.1 VARIABLE X_1 HAS DISTRIBUTION $N(\mu_1, \sigma_1)$, VARIABLE X_2 HAS DISTRIBUTION $N(\mu_2, \sigma_2)$ AND VARIANCES σ_1 AND σ_2 ARE UNKNOWN.

Test statistic for the hypothesis on two means is the following:

$$t_{cal} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where: $\bar{X}_1$ is the estimate of μ_1 and is calculated based on a random sample from population 1, $\bar{X}_2$ is the estimate of μ_2 and is calculated based on a random sample from population 2, s is the standard deviation calculated for both random samples considered together, n_1 is the number of elements in the random sample from population 1, n_2 is the number of elements in the random sample from population 2.

If the null hypothesis is true, the test statistic t_{cal} has t-Student distribution with $\nu = n_1 + n_2 - 1$ degrees of freedom.

The criteria of null hypothesis rejection in the considered case are identical with those referring to the hypothesis on one mean in case the variable X has normal distribution $N(\mu, \sigma)$ with unknown parameters. Please refer to §7.3.1 for more detailed information.

7.4.2 VARIABLE X_1 AND VARIABLE X_2 HAVE UNKNOWN DISTRIBUTIONS

In case the probability distributions of variables X_1 and X_2 are unknown, their random samples should be big. It is recommended that the number of elements in each sample exceed $20 \div 30$.

The test statistic for the hypothesis on two means is the following:

$$Z_{cal} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

where: $\bar{X}_1$ is the estimate of μ_1 and s_1 is the estimate of σ_1, both are calculated based on a random sample from population 1, $\bar{X}_2$ is the estimate of μ_2 and s_2 is the

estimate of σ_2, both are calculated based on a random sample from population 2, n_1 is the number of elements in the random sample from population 1 and n_2 is the number of elements in the random sample from population 2.

If the null hypothesis is true, the test statistic Z_{cal} has normal distribution $N(0,1)$.

The criteria of null hypothesis rejection in the considered case are identical with those referring to the hypothesis on one mean in case the variable X has unknown distribution. Please refer to §7.3.2 for more detailed information.

7.4.2.1 EXAMPLE

Problem. It was hypothesized that the exchange of a cutting tool for a different kind shortens the time of workpiece tooling with a lathe. Is this hypothesis justified at the significance level 0.01? In order to answer the question, the durations of tooling 10 workpieces with an old cutting tool and the durations of tooling 10 workpieces with a cutting tool of different kind were measured. The obtained measurement data is shown in Table 7.2. It may be assumed that both times have normal distribution.

Table 7.2 Time of workpiece tooling with an old cutting tool and with a cutting tool of a different kind/ min.

old cutting tool (I)	58	58	56	38	70	38	42	75	68	67
cutting tool of different kind (II)	57	55	63	24	67	43	33	68	56	54

Solution. It is possible to solve the problem using a test on two means. The two means are the mean of variable X_1, which is the time of tooling with cutting tool I and the mean of variable X_2, which is the time of tooling with cutting tool II. It is worth considering the null hypothesis which states that the average time of tooling with cutting tool I is equal to the average time of tooling with cutting tool II, namely $H_0: \mu_1 = \mu_2$. The null hypothesis is tested versus the one-sided alternative hypothesis that the mean time of tooling with tool I is longer than the mean time of tooling with tool II, $H_a: \mu_1 > \mu_2$. Variables X_1 and X_2 have normal distribution and their parameters are estimated based on the measurement results in the following way:

$$\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i} = 57$$

$$\bar{x}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} x_{2j} = 52$$

$$s = \sqrt{\frac{1}{n_1 + n_2 - 1} \sum_{k=1}^{n_1+n_2} (x_k - \bar{x})^2} = 13.90$$

The corresponding test statistic is the following:

$$t_{cal} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{57 - 52}{\sqrt{13.90^2\left(\frac{1}{10} + \frac{1}{10}\right)}} = 0.804$$

- The criterion of null hypothesis rejection based on the critical interval.

The critical interval is $\langle t_\alpha, \infty)$. Based on t-Student tables (Appendix 2), $t_\alpha = 2.539$ for $\alpha = 0.01$ and $\nu = n_1 + n_1 - 1 = 19$. Therefore, numerically the critical interval is $\langle 2.539, \infty)$.

The value of test statistic $t_{cal} = 0.804$ is located outside the critical interval; therefore, the null hypothesis is accepted at the significance level $\alpha = 0.01$.

- The criterion of null hypothesis rejection based on the p-value.

The p-value was calculated using the T.DISTRIBUTION function available in Excel. Considering $t_{cal} = 0.804$, the associated $p = 0.2156$.

The value of $p = 0.2156$ is greater than the value of $\alpha = 0.01$; therefore, the null hypothesis is accepted at the significance level 0.01.

Based on the obtained results of hypothesis testing, an engineer infers that the times of tooling with the two cutting tools are the same at the significance level $\alpha = 0.01$. Therefore, the exchange of cutting tool I for cutting tool II does not shorten the time of tooling.

7.5 TEST ON THE VARIANCE

The null hypothesis H_0 in the case of the test on the variance sates that the variance σ^2 of variable X in the general population is equal to a certain reference value σ_0^2. The formal notation of the null hypothesis is the following:

$$H_0\colon\ \sigma^2 = \sigma_0^2$$

In practice, one is usually interested in a small variance because a large variance is disadvantageous. Therefore, the null hypothesis on the variance is typically tested versus the alternative hypothesis that the variance is greater than the reference value:

$$H_A\colon\ \sigma^2 > \sigma_0^2 \qquad \text{one–sided hypothesis.}$$

In order to test the hypothesis, the assumption is required that the variable X has normal distribution $N(\mu, \sigma)$, but it is allowed that μ and σ are unknown. The assumption shall be confirmed by a normality test.

The test statistic for the hypothesis on the variance is the following:

$$\chi^2_{cal} = \frac{(n-1)s^2}{\sigma_0}$$

where s^2 is the estimate of the variance of variable X, based on a random sample which consists of n elements.

If the null hypothesis is true, the test statistic χ^2_{cal} has a χ^2 distribution with $\nu = n - 1$ degrees of freedom.

One may use the test on variance, σ^2 in order to actually perform the test on standard deviation, σ.

- The criterion of null hypothesis rejection based on the critical interval.

The criterion of null hypothesis rejection is $P\left(\chi^2 \geq \chi^2_{\alpha,\nu}\right) = \alpha$.

Therefore, the critical interval is $\chi^2 \in \langle \chi^2_{\alpha,\nu}, \infty \rangle$.

The null hypothesis is rejected at the significance level α if the calculated value of test statistic χ^2_{cal} belongs to the critical interval, i.e. if the following holds: $\chi^2_{cal} \in \langle \chi^2_{\alpha,\nu}, \infty \rangle$.

There is no reason for rejecting the null hypothesis at the significance level α if the calculated value of test statistic χ^2_{cal} remains outside the critical interval, i.e. if the following is true: $U_{cal} \in \langle 0, \chi^2_{\alpha,\nu} \rangle$.

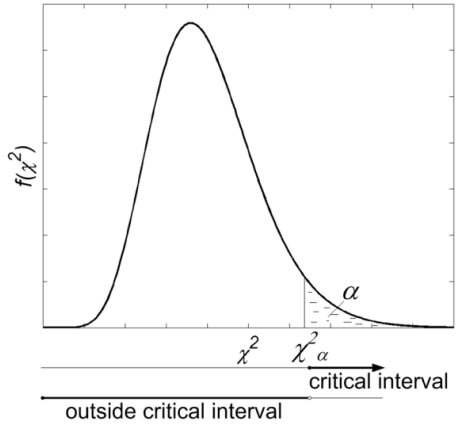


Figure 7.10 Graphical interpretation of null hypothesis rejection based on the critical interval criterion. Test on the variance. $\boldsymbol{H_0}$: $\sigma^2 = \boldsymbol{\sigma_0^2}$. $\boldsymbol{H_A}$: $\sigma^2 > \boldsymbol{\sigma_0^2}$.

The graphical interpretation of the critical interval criterion of null hypothesis rejection is shown in Fig. 7.10.

- The criterion of null hypothesis rejection based on the p-value.

The criterion of null hypothesis rejection is the following:

$$p \leq \alpha, \text{ where } p = P\left(\chi^2 \geq \chi^2_{cal}\right)$$

is the probability that χ^2 variable, which has χ^2 distribution with $\nu = n - 1$ degrees of freedom, is greater than or equal to the calculated value of test statistic χ^2_{cal}.

The null hypothesis is rejected if the significance level α is greater than or equal to p.

The null hypothesis is accepted if the significance level α is less than p.

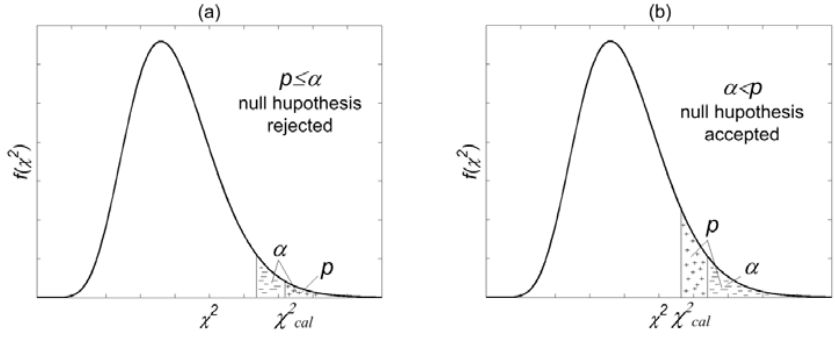


Figure 7.11 Graphical interpretation of null hypothesis rejection based on the $\boldsymbol{p}$-value criterion. Test on the variance. $\boldsymbol{H_0}$: $\sigma^2 = \boldsymbol{\sigma_0^2}$. $\boldsymbol{H_A}$: $\sigma^2 > \boldsymbol{\sigma_0^2}$.

The graphical interpretation of the p-value criterion of null hypothesis rejection is shown in Fig. 7.11.

7.5.1.1 EXAMPLE

Problem. The quality assurance standard requires that the variance of the diameter of molded pipes is not larger than 4 mm. As a routine check, $n = 11$ measurements were performed on the diameter of molded pipes. The obtained data is shown in Table 7.3.

Table 7.3 Results of measurement on the diameter of molded pipes/ mm.

50.2	50.4	50.6	50.5	49.9	50.0	50.3	50.1	50.0	49.6	50.6

It is known that the diameter of molded pipes has normal distribution. Is the quality assurance standard met by the tested production lot at the significance level 0.01?

Solution. It is possible to solve the problem using a test of variance regarding variable X, which is the diameter of molded pipes. It is worth considering the null hypothesis which states that the variance of the diameter of molded pipes is equal to 0.04 mm, namely $H_0: \sigma = 0.04$. The null hypothesis is tested versus the one-sided alternative hypothesis that the variance of pipe diameter is greater than 0.04 mm, $H_a: \sigma > 0.04$. The distribution of variable X is normal. The estimate of σ is calculated as follows:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0.1$$

The corresponding test statistic is the following:

$$\chi^2_{cal} = \frac{(n-1)s^2}{\sigma_0} = \frac{(11-1)0.1^2}{0.04} = 2.5$$

- The criterion of null hypothesis rejection based on the critical interval.

Considering the one-sided alternative hypothesis, the critical interval is $\langle \chi^2_{\alpha,\nu}, \infty \rangle$. Based on χ^2 tables (Appendix 3), $\chi^2_{\alpha,\nu} = 23.209$ for $\alpha = 0.01$ and $\nu = n - 1 = 10$ degrees of freedom. Therefore, numerically the critical interval is $\langle 23.209, \infty \rangle$.

The value of test statistic $\chi^2_{cal} = 2.5$ is located outside the critical interval. Therefore, the null hypothesis is accepted at the significance level $\alpha = 0.01$.

- The criterion of null hypothesis rejection based on the p-value.

The p-value was calculated using the CHI2.DISTRIBUTION function available in Excel. Considering $\chi^2_{cal} = 2.5$, the associated $p = 0.991$.

The value $p = 0.991$ is greater than the value of $\alpha = 0.01$; therefore, the null hypothesis is accepted at the significance level 0.01.

Based on the obtained results of hypothesis testing, the quality assurance engineer has good reason to claim that the quality assurance standard concerning the variation of pipe diameter is met at the significance level $\alpha = 0.01$.

7.6 TEST ON TWO VARIANCES

The null hypothesis H_0 in case of a test of two variances states that the variance σ_1^2 of variable X_1 is equal to the variance σ_2^2 of variable X_2.

$$H_0: \sigma_1^2 = \sigma_2^2$$

The null hypothesis is tested versus one of three different alternative hypotheses:

I. $H_A: \sigma_1^2 \neq \sigma_2^2$ two–sided hypothesis

II. $H_A: \sigma_1^2 > \sigma_2^2$ one–sided hypothesis

III. H_A: $\sigma_1^2 < \sigma_2^2$ one–sided hypothesis.

For the sake of testing the hypothesis on two variances, the assumption is required that variable X_1 has normal distribution $N(\mu_1, \sigma_1)$ and variable X_2 has normal distribution $N(\mu_2, \sigma_2)$. It is allowed that the parameters of these distributions μ_1, σ_1 and μ_2, σ_2 are unknown. The assumption shall be confirmed by a normality test.

The test statistic for the hypothesis on two variances is the following:

$$F_{cal} = \frac{s_1^2}{s_2^2}$$

where: s_1^2 is the estimate of σ_1^2, based on a random sample consisting of n_1 elements and s_2^2 is the estimate of σ_2^2, based on a random sample consisting of n_2 elements. The indices 1 and 2 are assigned in a way that $s_1^2 > s_2^2$ and consequently $F_{cal} \geq 1$.

If the null hypothesis is true, the test statistic F_{cal} has an F-Snedecore distribution with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ degrees of freedom.

The criteria of null hypothesis rejection depend on the kind of alternative hypothesis which is considered together with the null hypothesis.

One may use a test on two variances, σ_1^2 and σ_2^2, in order to actually perform a test on two standard deviations, σ_1 and σ_2.

I. Criterion of null hypothesis rejection on two variances versus H_A: $\sigma_1^2 \neq \sigma_2^2$.

- The criterion of null hypothesis rejection based on the critical interval.

The criterion of null hypothesis rejection is the following:

$$P\left(F_{1-\frac{\alpha}{2},\nu_1,\nu_2} \leq F \vee F \leq F_{\frac{\alpha}{2},\nu_1,\nu_2}\right) = \alpha.$$

Therefore, the critical interval is $F \epsilon \langle 0, F_{1-\frac{\alpha}{2},\nu_1,\nu_2} \rangle \cup \langle F_{\frac{\alpha}{2},\nu_1,\nu_2}, \infty \rangle$.

Due to the fact that $F_{cal} \geq 1$, only the right part of the critical interval is used.

Therefore, the actual critical interval is: $F_{cal} \epsilon \langle F_{\frac{\alpha}{2},\nu_1,\nu_2}, \infty \rangle$.

The null hypothesis is rejected at the significance level α if the calculated value of test statistic F_{cal} belongs to the critical interval, i.e. if the following holds:

$F_{cal} \epsilon \langle F_{\frac{\alpha}{2},\nu_1,\nu_2}, \infty \rangle$.

The null hypothesis is accepted at the significance level α if the calculated value of test statistic F_{cal} remains outside the critical interval, i.e. if the following holds:

$F_{cal} \epsilon \left(F_{1-\frac{\alpha}{2},\nu_1,\nu_2}, F_{\frac{\alpha}{2},\nu_1,\nu_2}\right)$.

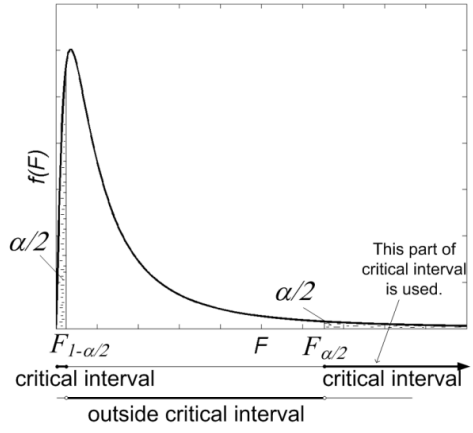


Figure 7.12 Graphical interpretation of null hypothesis rejection based on the critical interval criterion. Test on two variances. $\boldsymbol{H_0}$: $\boldsymbol{\sigma_1^2 = \sigma_2^2}$. $\boldsymbol{H_A}$: $\boldsymbol{\sigma_1^2 \neq \sigma_2^2}$.

The graphical interpretation of the critical interval criterion of null hypothesis rejection is shown in Fig. 7.12.

- The criterion of null hypothesis rejection based on the p-value.

The criterion of null hypothesis rejection is $p \leq \frac{\alpha}{2}$, where $p = P(F \geq F_{cal})$ is the probability that the F variable, which has F-Snedecore distribution with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ degrees of freedom is greater than or equal to the calculated value of test statistic F_{cal}.

The null hypothesis is rejected if half of the significance level α is greater than or equal to p.

The null hypothesis is accepted if half of the significance level α is less than p.

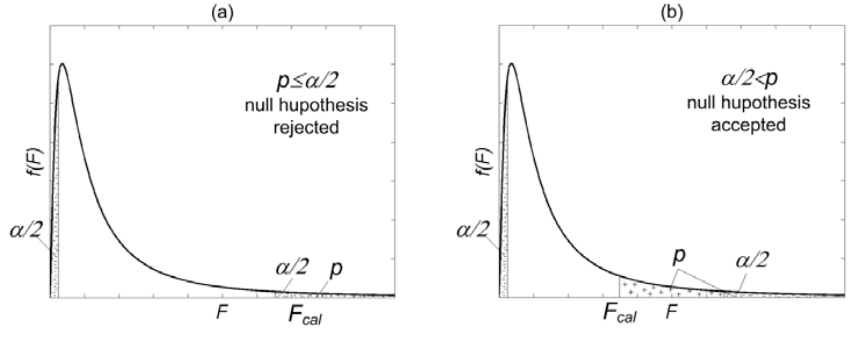


Figure 7.13 Graphical interpretation of null hypothesis rejection based on the $\boldsymbol{p}$-value critical interval criterion. Test on two variances. $\boldsymbol{H_0}$: $\boldsymbol{\sigma_1^2 = \sigma_2^2}$. $\boldsymbol{H_A}$: $\boldsymbol{\sigma_1^2 \neq \sigma_2^2}$.

The graphical interpretation of the p-value criterion of null hypothesis rejection is shown in Fig. 7.13.

II. Criterion of null hypothesis rejection for two variances versus the one-sided alternative hypothesis (case II of H_a).

- Criterion of the critical interval

The criterion of null hypothesis rejection is the following:

$$P\left(F_{\alpha,\nu_1,\nu_2} \leq F\right) = \alpha$$

Therefore, the critical interval is $F\epsilon\langle F_{\alpha,\nu_1,\nu_2}, \infty\rangle$.

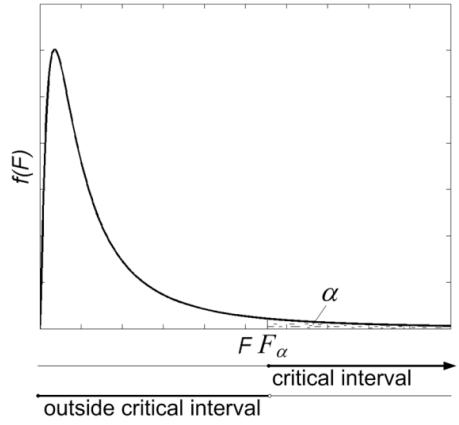


Figure 7.14 Graphical interpretation of null hypothesis rejection based on the critical interval criterion. Test on two variances. $\boldsymbol{H_0}$: $\boldsymbol{\sigma_1^2 = \sigma_2^2}$. $\boldsymbol{H_A}$: $\boldsymbol{\sigma_1^2 > \sigma_2^2}$.

The graphical interpretation of the critical interval criterion of null hypothesis rejection is shown in Fig. 7.14.

- Criterion of the p-value

The criterion of null hypothesis rejection is $p \leq \alpha$, where $p = P(F \geq F_{cal})$ is the probability that F variable, which has F-Snedecore distribution with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ degrees of freedom is greater than or equal to the calculated value of test statistic F_{cal}.

The null hypothesis is rejected if the significance level is greater than or equal to p.

The null hypothesis is accepted if the significance level is less than p.

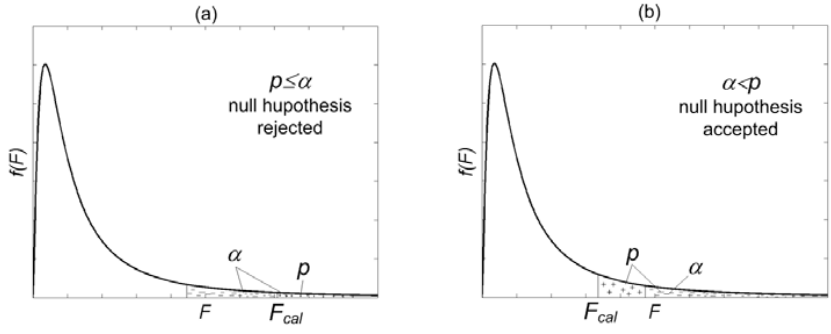


Figure 7.15 Graphical interpretation of null hypothesis rejection based on the $\boldsymbol{p}$-value criterion. Test on two variances. $\boldsymbol{H_0}$: $\boldsymbol{\sigma_1^2 = \sigma_2^2}$. $\boldsymbol{H_A}$: $\boldsymbol{\sigma_1^2 > \sigma_2^2}$.

The graphical interpretation of the p-value criterion of null hypothesis rejection is shown in Fig. 7.15.

III. Criterion of null hypothesis rejection of the mean $\mathrm{H_0}$ for the one-sided alternative hypothesis (case III of $\mathrm{H_a}$).

The hypothesis shall be tested as shown for case II. However, prior to testing, the variables shall be renumbered, i.e. X_1 shall take the index 2 and X_2 shall take the index 1 so that the ratio of variances is greater than one.

7.6.1.1 EXAMPLE

Problem. In order to check the precision of current measurement with two different measuring devices, measurements of 7 A current were performed. The obtained results are shown in Table 7.4.

Table 7.4 Measurement results of 7 A current with two different measuring devices.

Device 1	7.2	6.7	6.9	6.9	7.2	7.0	7.1
Device 2	7.4	6.8	7.4	6.6	6.3	7.5	

Is the measurement precision of the two devices equal at the significance level 0.05? It is correct to assume that the measurement results have normal distribution in each case.

Solution. Precision is indicated by the spread of replicate measurement results. It is possible to solve the problem using a test of two variances regarding variable X_1, which is the result of measuring with device I and variable X_2, which is the result of measuring with device II. It is worth considering the null hypothesis, which states that the variance of measurements performed with device I is the same as the variance of measurements performed with device II, namely $H_0: \sigma_1 = \sigma_2$. The null hypothesis is tested versus the two-sided alternative hypothesis that the variance of measurements performed with device I is different from the variance of

measurements performed with device II, $H_a: \sigma_1 \neq \sigma_2$. The distributions of both variables X_1 and X_2 are normal. The estimates of σ_1^2 and σ_2^2 are calculated as follows:

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2 = 0.033$$

$$s_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}\left(x_j - \bar{x}\right)^2 = 0.252$$

As s_1^2 is greater than s_2^2, we renumber our variables.

The corresponding test statistic is the following:

$$F_{cal} = \frac{s_1^2}{s_2^2} = \frac{0.252}{0.033} = 7.56$$

- Criterion of null hypothesis rejection based on the critical interval.

Considering the two-sided alternative hypothesis, the critical interval is $\langle F_{\frac{\alpha}{2},\nu_1,\nu_2}, \infty\Big)$. Based on F tables (Appendix 5), $F_{\alpha,\nu_1,\nu_2} = 4.387$ for $\alpha = 0.05$ and $\nu_1 = n_1 - 1 = 5$ and $\nu_2 = n_2 - 1 = 6$ degrees of freedom. Therefore, numerically the critical interval is $\langle 4.387\ , \infty)$.

The value of test statistic $F_{cal} = 7.56$ is located inside the critical interval. Therefore, the null hypothesis is rejected at the significance level $\alpha = 0.05$.

- Criterion of null hypothesis rejection based on the p-value.

The p-value was calculated using the F.DISTRIBUTION function available in Excel. Considering $F_{cal} = 7.56$, the associated $p = 0.0143$.

The value $p = 0.0143$ is less than the value of $\alpha = 0.05$; therefore, the null hypothesis is rejected at the significance level 0.05.

Based on the obtained results of hypothesis testing, an engineer is allowed to conclude that the two devices offer different measurement precision at the significance level $\alpha = 0.05$. Device I offers measurements with significantly higher precision.

7.7 NORMALITY TESTS

For testing normality any test which belongs to a group of goodness–of–fit tests may be used. Goodness–of–fit tests are a class of nonparametric tests. They are used for testing two kinds of hypotheses. One kind refers to suppositions that two variables have the same statistical distribution. The other kind refers to suppositions that a variable has the defined statistical distribution. Normality tests belong here.

A number of tests are available for testing normality. The tests which were designed for testing normality are, for example, Saphiro-Wilk test and Epps-Pulley test. The tests were designed for testing goodness of fit between two distributions including Normal are, for example, the Kolmogorov test, Kolmogorov-Smirnov test, χ^2 test. Tests belonging to the second group are weaker when testing normality compared to tests in the first group.

Goodness-of-fit tests are calculation-intensive and therefore they are usually performed using statistical software. This refers in particular to normality tests. However, they are not widely available in common access software. Therefore, the method of performing the λ Kolmogorov test is presented in this book.

It may occur that the hypothesis about the normality of the probability distribution of variable X is rejected by the statistical test. Such a result is disadvantageous because many statistical methods and tools require the assumption about the normality of the variable. Therefore, the transformation of the original variable is recommended for obtaining the variable X_T which has normal distribution instead of the variable which does not have normal distribution X. The new variable X_T is then statistically analyzed. The most popular, although not always successful, transformation of a 'non-normal' variable X into a 'normal' variable X_T is $X_T = ln(X)$.

7.7.1 λ KOLMOGOROV TEST

λ Kolmogorov test is a goodness-of-fit test.

The null hypothesis states that the empirical cumulative distribution $F(X)$ of variable X is equal to a hypothetical, reference continuous cumulative distribution $F_0(X)$. In particular, $F_0(X)$ is a standardized normal distribution Z.

$$H_0: F(x) = F_0(x)$$

The null hypothesis is tested versus the alternative hypothesis:

$$H_A: F(x) \neq F_0(x)$$

The test statistic is the following:

$$\lambda_{cal} = D\sqrt{n}$$

where: n is the number of elements in a sample (it should be several dozen at least). D is represented by the following formula:

$$D = sup\left|F_j(x) - F_{0j}(x)\right|$$

where: $F_j(x)$ is the value of the empirical cumulative distribution function for the j^{th} interval of values of variable X, and $F_{0j}(x)$ is the value of the reference cumulative distribution function calculated for the right limit of the j^{th} interval of values of variable X.

If the null hypothesis is true, λ_{cal} has λ Kolmogorov distribution which is independent from the empirical distribution $F(x)$.

All n values of variable X are grouped into j intervals between the minimum and maximum value of X.

- Criterion of null hypothesis rejection based on the critical interval criterion

The criterion of null hypothesis rejection is the following:

$$P(\lambda_\alpha \leq \lambda) = \alpha$$

where: λ_α comes from the tables of λ Kolmogorov distribution (Appendix 7).

Therefore, the critical interval is $\lambda \epsilon (\lambda_\alpha, \infty)$. The null hypothesis is rejected if the following is true $\lambda_{cal} \epsilon (\lambda_\alpha, \infty)$.

- Criterion of the p-value

The criterion of null hypothesis rejection is $p \leq \alpha$, where $p = P(\lambda \geq \lambda_{cal})$ is the probability that λ variable, which has λ Kolmogorov distribution, is greater than or equal to the calculated value of the test statistic λ_{cal}.

The null hypothesis is rejected if the significance level is greater than or equal to p.

The null hypothesis is accepted if the significance level is less than p.

7.7.1.1 EXAMPLE

Problem. 200 sardines were caught in the Atlantic Ocean. Their size was measured and the results are shown in Table 7.5. Does the size of the sardines have normal distribution at the significance level 0.05.?

Table 7.5 Empirical data concerning sardines caught in the Atlantic Ocean.

Length of sardine/ cm	Number of fish
10-12	10
12-14	26
14-16	56
16-18	64
18-20	30
20-22	14

Solution. It is possible to solve the problem using the λ Kolmogorov goodness–of–fit test. The considered variable X is the size of the sardines. The reference cumulative distribution is a normal distribution. The parameters of normal distribution are estimated based on the random sample in the following manner:

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{r} x_j^0 n_j = 16.2$$

$$s = \sqrt{\frac{1}{n}\sum_{j=1}^{r}\left(x_j^0 - \bar{x}\right)^2 n_j} = 2.47$$

where: x_j^0 is the mean value of X for the j^{th} interval, n_j is the number of elements inside j^{th} interval.

The values of X, which represent the right limits of intervals for sardine length, x_j, are standardized using the formula:

$$z_j = \frac{x_j - \bar{x}}{s}$$

so that the reference normal distribution $N(\bar{x}_j, s)$ is converted into the standardized normal distribution $Z(0,1)$. Values of the cumulative Z distribution function $F(z_j)$ are read out from the statistical tables (Appendix 1) for all standardized values z_j. In this way, the reference cumulative distribution function $F_{0j}(x)$ is calculated.

Values of the cumulative empirical distribution are calculated for each interval of sardine length using the formula:

$$F_j(x) = \frac{n_{cum}}{\sum_{j=1}^{k} n_j}$$

where n_j is the number of sardines with their length belonging to the j^{th} interval, $k = 1 \ldots j$.

The comparison of two cumulative distributions: empirical and normal is shown in Table 7.6.

Table 7.6 The comparison of cumulative empirical distribution of the length of sardines $F_{emp}(x)$ and normal distribution $F(z_j)$.

j	x_j	z_j	$F(z_j) = F_{0j}(x)$	n_j	n_{cum}	$F_j(x)$	$\lvert F_j(x) - F_{0j}(x)\rvert$
1	12	-1.70	0.037	10	10	0.05	0.0054
2	14	-0.89	0.187	26	36	0.18	0.0067
3	16	-0.08	0.468	56	92	0.46	0.0081
4	18	0.73	0.767	64	156	0.78	**0.0127**
5	20	1.54	0.938	30	186	0.93	0.0082
6	22	2.35	0.991	14	200	1	0.0094

Based on Table 7.6 the following is true:

$$D = sup\left|F_j(x) - F_{0j}(x)\right| = 0.0127$$

The corresponding value of the test statistic is

$$\lambda_{cal} = D\sqrt{n} = 0.0127 \cdot \sqrt{200} = 0.180$$

- Criterion of null hypothesis rejection based on the critical interval.

The critical interval for λ is $\langle \lambda_\alpha, \infty)$. Based on λ Kolmogorov distribution tables (Appendix 7), $\lambda_\alpha = 1.358$ for $\alpha = 0.05$. Therefore, numerically the critical interval is $\langle 1.358\,, \infty)$.

The value of the test statistic $\lambda_{cal} = 0.180$ is located outside the critical interval. Therefore, the null hypothesis is accepted at the significance level $\alpha = 0.05$.

- The criterion of null hypothesis rejection based on the p-value.

Based on λ Kolmogorov distribution tables (Appendix 7), the p-value associated with $\lambda_{cal} = 0.180$ is greater than 0.999. The p-value is greater than the value $\alpha = 0.05$; therefore, the null hypothesis is accepted at the significance level 0.05.

Based on the obtained results of hypothesis testing, one can assume that the size of sardines has normal distribution at the significance level $\alpha = 0.05$.

8 ANALYSIS OF VARIANCE

The problem of indicating change in objects as a result of being influenced by different factors is daily encountered in engineering practice.

A good example is a product. Its features are influenced by the parameters of a production process, e.g. temperature, concentration of ingredients, type of additive, intensity and/or duration of mixing and the like.

An engineer may be interested in securing reproducible products that requires process parameters to remain within certain limits which do not cause significant variability of the object. It is also possible that an engineer is interested in modifying a product, e.g. improving its quality. This requires process parameters to be changed in a way that causes significant and desirable change of the object.

The sensitivity of objects to nonrandom factors which act on them is statistically analyzed with the analysis of variance (ANOVA). The main idea of the analysis of variance consists of studying the variability in a response variable regarding factors which are responsible for it.

The total variability of the response variable is decomposed into parts. Part of the variability is attributed to random factors, another part is assigned to nonrandom factors and yet another part is considered as resulting from interactions between nonrandom factors. The significance of variation caused by nonrandom factors and their interactions is judged versus the variability which has random origin.

The analysis of variance shall be employed to the measurement data collected in an active manner (Charter 2).

8.1 ONE WAY ANALYSIS OF VARIANCE (ANOVA)

The simplest kind of analysis of variance, the so called one-way analysis of variance, is dedicated to one-factor problems. The aim of this analysis is to find out whether the investigated object is sensitive to one selected nonrandom factor. The feature of the object, which is expected to be influenced by the factor, is represented by a measurable response variable Y.

8.1.1 PREPARATION OF MEASUREMENT DATA FOR ONE-WAY ANOVA

The main idea of the experiment providing data for one–way ANOVA is to expose the object to several different levels of the considered factor, X_A and to measure values of the response variable Y several times for each level of the factor. All the other known and controllable factors shall remain at a constant level during the course of the experiment. The recommended form of the data table is shown in Table 8.1.

There are n levels of factor X_A considered. The i^{th} level of the factor is denoted by X_A^i and $i = 1 \dots n$. r replicate measurements of Y at each i^{th} level of factor X_A are performed. The k^{th} replicate measurement is denoted by y_k^i, where $k = 1 \dots r$.

Table 8.1 The table of measurement data prepared for the one-way analysis of variance. Values of response variable Y correspond to different levels of factor $\boldsymbol{X_A}$.

Level of factor X_A	Replicate measurement of response variable Y				
	1	…	k	…	r
X_A^1	y_1^1	…	y_k^1	…	y_r^1
…					
X_A^i	y_1^i	…	y_k^i	…	y_r^i
…					
X_A^n	y_1^n	…	y_k^n	…	y_r^n

It is important to randomize the levels of the factor X_A and to apply them to the object in randomized order. The object shall never be exposed to increasing or decreasing levels of the factor in sequence.

8.1.2 DECOMPOSITION OF VARIANCE IN ONE-WAY ANOVA

Two sources of variation of the response variable Y in the one-way analysis of variance are considered. These are random factors and the nonrandom factor X_A. Their contribution to the variation of Y is represented by the associated variances.

In ANOVA, the total variation of variable Y is decomposed into two parts. The first part is the so called within-level or within-group variation. It is attributed to random factors. The second part is the so called cross-level or between group variation. It is attributed to the factor X_A.

8.1.2.1 MEANS OF THE RESPONSE VARIABLE

The overall mean μ of the response variable Y, associated with object exposure to factor X_A is represented by the average $\bar{y}$ of all values y_k^i recorded during r replicate measurements at each of the n levels of the factor.

$$\bar{y} = \frac{1}{nr} \sum_{i=1}^{n} \sum_{k=1}^{r} y_k^i$$

The average response of the object μ^i to the i^{th} level of factor X_A is represented by the average $\bar{y}^i$ of values recorded during repeated measurements while exposing the object to the i^{th} level of the factor.

$$\bar{y}^i = \frac{1}{r}\sum_{k=1}^{r} {y_k}^i$$

It is expected that values of variable Y which are recorded in the course of repeated measurements are different, i.e. $y_i^1 \neq \ldots \neq y_i^k \neq \ldots \neq y_i^r$, despite the fact that the object is exposed to a constant level of factor X_A. The spread of values is caused by random factors.

8.1.2.2 TOTAL VARIATION OF THE RESPONSE VARIABLE

In the one-way analysis of variance, the total variation of response variable Y is represented by the sum of squares SS_T of differences between the total mean and every single value of this variable.

$$SS_T = \sum_{i=1}^{n}\sum_{k=1}^{r}\left(\bar{y} - y_k^i\right)^2$$

The total variability of Y, represented by the sum of squares SS_T, is the algebraic sum of the variability of Y attributed to random factors, which is represented by the sum of squares SS_E, and the variability of Y attributed to a controlled factor X_A, which is represented by the sum of squares SS_E.

$$SS_T = SS_E + SS_A$$

8.1.2.3 VARIATION ATTRIBUTED TO RANDOM FACTORS

The within level variation of the response variable Y is observed when the level of factor X_A is fixed. This variation is attributed to random factors. In the one-way analysis of variance, the within level variation of Y is represented by a sum of squares SS_E. This is a sum of the squared differences between the mean value of the response variable associated with the i^{th} level of factor $\bar{y}^i$ and every single value of this variable ${y_k}^i$ recorded at this level of the factor.

$$SS_E = \sum_{i=1}^{n}\sum_{k=1}^{r}\left(\bar{y}^i - y_k^i\right)^2$$

Referring to Table 8.1, SS_E describes the variation of Y inside a single cell of the table. It is aggregated for all the cells.

There are ν_E degrees of freedom associated with the within level variance:

$$\nu_E = n \cdot r - n = n(r-1)$$

The within level variance of Y is given by the following mean square:

$$s^2(y)_E = MS_E = \frac{SS_E}{\nu_E}$$

8.1.2.4 VARIATION ATTRIBUTED TO A NON-RANDOM FACTOR

The cross-level variation of the response variable Y is observed when levels of factor X_A are changed. This variation is attributed to factor X_A. In the one-way analysis of variance, the between-level variation of Y is represented by a sum of squares SS_A. This is the sum of squared differences between the overall mean value of the response variable $\bar{y}$ and the mean values of the response variable $\bar{y}^i$ associated with each level of factor X_A.

$$SS_A = \sum_{i=1}^{n} \left(\bar{y} - \bar{y}^i\right)^2$$

Referring to Table 8.1, SS_A describes the variation of Y among rows of the table.

There are ν_A degrees of freedom associated with the cross-level variance:

$$\nu_A = n - 1$$

The cross-level variance is given by the following mean square:

$$s^2(y)_A = MS_A = \frac{SS_A}{\nu_A}$$

8.1.3 NULL HYPOTHESIS IN ONE-WAY ANOVA

The null hypothesis in the one-way analysis of variance states that the average response of the object to different levels of factor X_A is the same. In other words, on average the object responds in the same way to each level of factor X_A. The object is insensitive to the changes of the factor. The formal representation of the null hypothesis is the following:

$$H_0: \mu^1 = \cdots = \mu^i = \cdots = \mu^n$$

The null hypothesis is tested versus the alternative hypothesisstating that the average responses of the object are different for at least two different levels of factor X_A. In other words, the object is sensitive to the change between at least two levels of the factor. The formal representation of the alternative hypothesis is the following:

$$H_a: \exists\, \mu^i \neq \mu^l$$

where $i = 1..n$, $l = 1..n$ and $i \neq l$.

The following test statistic is used for testing the null hypothesis:

$$F_{cal} = \frac{s^2(y)_A}{s^2(y)_E}$$

If the null hypothesis is true, the variable F_{cal} has F-Snedecore distribution with the degrees of freedom ν_A and ν_E.

The critical interval criterion of null hypothesis rejection at the significance level α is

$$p\left(F \geq F_{\alpha,\nu_A,\nu_E}\right) = \alpha$$

where: F_{α,ν_A,ν_E} is the value of variable F, which comes the F -Snedecore distributions with the degrees of freedom ν_A and ν_E, for the assumed value of α.

The critical interval for F_{cal} is $F\epsilon\langle F_{\alpha,\nu_A,\nu_E}, \infty\rangle$. If $F_{cal}\epsilon\langle F_{\alpha,\nu_A,\nu_E}, \infty\rangle$, the null hypothesis is rejected.

The p-value criterion of null hypothesis rejection at the significance level α is

$$p = P\left(F_{\nu_A,\nu_E} \geq F_{cal}\right) \leq \alpha$$

The null hypothesis is rejected at the significance level α if α is greater than or equal to p.

For the graphical interpretation of the null hypothesis rejection criteria refer to Fig. 7.12 and Fig. 7.13.

Based on the presented reasoning, the null hypothesis is rejected when the test statistic F_{cal} reaches or exceeds F_{α,ν_A,ν_E}. The test statistic is the ratio between the variance of the response variable which comes from nonrandom factors, $s^2(y)_A$ and the variance of the response variable which is caused by random factors, $s^2(y)_E$. Therefore, the null hypothesis is rejected if the variation of Y caused by factor X_A is large enough when compared to the variation caused by random factors that the critical level F_{α,ν_A,ν_E} is reached. The rejection of the null hypothesis indicates that the considered factor X_A does significantly influence the object if represented by the response variable Y.

The null hypothesis is accepted on the condition that the ratio between the variance of response variable $s^2(y)_A$, which comes from factor X_A, and the variance of response variable $s^2(y)_E$, which is caused by random factors, does not exceed F_{α,ν_A,ν_E}. That is, the variation of Y caused by factor X_A is small enough when compared to its variation caused by random factors that the critical level F_{α,ν_A,ν_E} is not reached. The acceptance of the null hypothesis indicates that the considered factor X_A does not significantly influence the object if represented by the response variable Y.

8.1.4 EXAMPLE

Problem. Students were interested whether costs of dishwashing are influenced by the way the dishes are washed. They decided to carry out a relevant experiment. The response variable Y represented the costs of dishwashing. It was calculated as the sum of the following components: cost of electricity, cost of gas, cost of water, fee for the waste water, cost of the washing liquid, cost of dishwasher detergent and cost of dishwasher salt. The investigated factor X_A was the method of washing. Three methods of washing were considered as three levels of the factor: ordinary

manual dishwashing, i.e. washing and rinsing with the water running (X_A^1), economical manual dishwashing, i.e. washing in the sink and rinsing with running water (X_A^2) and washing with a dishwasher (X_A^3). A defined set of dishes was washed three times using each dishwashing method.

The results of the experiment are shown in Table 8.2.

Table 8.2 The measurement data for the experiment considered in Example 8.1.4.

Level of factor X_A	Replicate measurement of response variable Y [PLN]		
	1	2	3
X_A^1	0.34	0.41	0.8
X_A^2	0.13	0.19	0.14
X_A^3	0.827	0.837	0.827

The significance level $\alpha = 0.05$ was assumed.

Solution. It is possible to solve the problem using the one-way analysis of variance. The considered variable Y is the cost of dishwashing. It is worth testing the hypothesis that the cost of dishwashing using any considered method is the same, $H_0: \mu^1 = \mu^2 = \mu^3$ versus the alternative hypothesis that at least two methods produce different costs of dishwashing $H_\alpha: \exists\, \mu^i \neq \mu^l$, $i = 1..3$, $l = 1..3$. The relevant calculation help is offered by the DATA ANALYSIS TOOL in Excel. The results of the one-way ANOVA are shown in Table 8.3.

Table 8.3 ANOVA table for the measurement data shown in Table 8.2.

Source of variance	SS	ν	MS	F_{cal}	*p-value*	$F_{\alpha=0.05,\nu_A,\nu_E}$
X_A	$SS_A = 0.689$	$\nu_A = 2$	$MS_A = 0.344$	16.530	0.0036	5.143
random factor	$SS_E = 0.125$	$\nu_E = 6$	$MS_E = 0.021$			
Total	$SS_T = 0.8137$	$\nu_T = 8$				

The criteria of null hypothesis rejection are fulfilled: $F_{cal} \geq F_{\alpha,\nu_A,\nu_E} \equiv p \leq \alpha$, as shown by the results in the one-way analysis of variance presented in Table 8.3.

Based on the performed analysis, students were able to infer that the method of washing influenced the cost of washing in a statistically significant manner at the significance level $\alpha = 0.05$.

8.2 MULTI-WAY ANALYSIS OF VARIANCE (MANOVA)

A more complex version of the analysis of variance, the so called multi-way analysis of variance, is dedicated to multi-factor problems. The aim of the analysis is to find out whether the investigated object is sensitive to several selected nonrandom factors and possibly their interactions. The feature of the object, which is expected

to be influenced by the factors, is represented by a measurable response variable Y.

The simplest multi-factor problem is a two-factor problem and the corresponding analysis of variance is called two-way ANOVA. The analysis of two-factor problems is covered in this book.

8.2.1 PREPARATION OF MEASUREMENT DATA FOR TWO-WAY ANOVA

The main idea of the experiment providing data for two–way ANOVA is to expose an object to all combinations of different levels of factors X_A and X_B. The response variable values shall be measured several times for each combination of factor levels. All the other known and controllable factors shall remain at a constant level in the course of the experiment. The recommended form of data table is shown in Table 8.4.

In the analysis, n levels of factor X_A and m levels of factor X_B are considered. The i^{th} level of factor X_A is denoted as X_A^i , $i = 1 \dots n$. The j^{th} level of factor X_B is denoted as X_B^j, $j = 1 \dots m$. r replicate measurements of the response variable Y for each combination $\{X_A^i, X_B^j\}$ of levels of the considered factors are performed. The k^{th} replicate measurement is denoted by $y_k^{i,j}$, where $k = 1 \dots r$.

Table 8.4 Table of measurement data prepared for the two-way analysis of variance. Values of response variable $\boldsymbol{Y}$ correspond to combinations of different levels of factors $\boldsymbol{X_A}$ and $\boldsymbol{X_B}$.

Level of factor X_B / Level of factor X_A	X_B^1	…	X_B^j	…	X_B^m
X_A^1	$y_1^{1,1}, \dots, y_k^{1,1}, \dots, y_r^{1,1}$	…	$y_1^{1,j}, \dots, y_k^{1,j}, \dots, y_r^{1,j}$	…	$y_1^{1,m}, \dots, y_k^{1,m}, \dots, y_r^{1,m}$
…	…	…	…	…	…
X_A^i	$y_1^{i,1}, \dots, y_k^{i,1}, \dots, y_r^{i,1}$	…	$y_1^{i,j}, \dots, y_k^{i,j}, \dots, y_r^{i,j}$	…	$y_1^{i,m}, \dots, y_k^{i,m}, \dots, y_r^{i,m}$
…	…	…	…	…	…
X_A^n	$y_1^{n,1}, \dots, y_k^{n,1}, \dots, y_r^{n,1}$	…	$y_1^{n,j}, \dots, y_k^{n,j}, \dots, y_r^{n,j}$	…	$y_1^{n,m}, \dots, y_k^{n,m}, \dots, y_r^{n,m}$

It is important to randomize combinations $\{X_A^i, X_B^j\}$ of levels of the considered factors and to apply them to the object in a randomized order. The object shall not be subsequently exposed to combinations organized along the increasing or decreasing levels of factor X_A or X_B.

8.2.2 DECOMPOSITION OF VARIANCE IN TWO-WAY ANOVA

Four sources of variation of the measured variable Y in the two-way analysis of variance are considered. These are random factors, factor X_A, factor X_B and the interaction between factors X_A and X_B. Their contribution to the variation of Y is represented by the associated variances.

In two-way ANOVA, the total variation of variable Y is decomposed into four parts. The first part is the so called within level or within group variation. It is attributed to random factors. The second part is the so called cross-level or between group variation and it is attributed to factor X_A. The third part is the so called cross-level or between group variation and it is attributed to factor X_B. The fourth part is the variation attributed to the interaction between factor X_A and factor X_B. The interaction $X_A X_B$ can be understood as a 'virtual' factor resulting from the joint impact of two 'physical' factors X_A and X_B. It is a kind of added value due to the exposure of the object to two factors simultaneously. The statistical significance of the interaction is proof that one factor magnifies or reduces the impact of the other factor on the object as compared to the circumstances when only one factor acts on the object.

8.2.2.1 MEANS OF THE RESPONSE VARIABLE

The overall mean μ of the response variable Y, associated with object exposure to factors X_A and X_B, is represented by the average $\bar{y}$ of all values $y_k^{i,j}$ recorded during r replicate measurements of Y at each of $n \cdot m$ combinations of levels of factors X_A and X_B.

$$\bar{y} = \frac{1}{nmr} \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{r} y_k^{i,j}$$

The average response μ^i of the object to the i^{th} level of factor X_A is represented by the average $\bar{y}^i$ of values recorded during repeated measurements while factor X_A remained at the i^{th} level and factor X_B was changed.

$$\bar{y}^i = \frac{1}{mr} \sum_{j=1}^{m} \sum_{k=1}^{r} y_k{}^{i,j}$$

The average response μ^j of the object to the j^{th} level of factor X_B is represented by the average $\bar{y}^j$ of values recorded during repeated measurements while factor X_B remained at the j^{th} level and factor X_A was changed.

$$\bar{y}^j = \frac{1}{nr} \sum_{i=1}^{n} \sum_{k=1}^{r} y_k{}^{i,j}$$

Considering fixed combination $\{X_A^i, X_B^j\}$ of levels of factors X_A and X_B, it is expected that the object responds with slightly different values of variable Y in repeated measurements, namely $y_1^{i,j} \neq \ldots \neq y_k^{i,j} \neq \ldots \neq y_r^{i,j}$, for $i = const$ and $j = const$. The differences are caused by random factors. In the two-way analysis of variance, the average response $\mu^{i,j}$ of the object to the $\{i,j\}$ combination of levels of factor X_A and X_B is represented by the average $\bar{y}^{i,j}$ of values recorded during repeated measurements while factor X_A remained at the i^{th}level and factor X_B remained at the j^{th} level.

$$\bar{y}^{i,j} = \frac{1}{r}\sum_{k=1}^{r} y_k^{i,j}$$

8.2.2.2 TOTAL VARIATION OF THE RESPONSE VARIABLE

In the two-way analysis of variance, the total variation of response variable Y is represented by a sum of squares SS_T of differences between the total mean of the response variable $\bar{y}$ and every single value of this variable $y_k^{i,j}$ observed upon all replicate measurements at each combination of levels of factor X_A and X_B.

$$SS_T = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{r}\left(\bar{y} - y_k^{i,j}\right)^2$$

The total variability of Y represented by the sum of squares SS_T is the algebraic sum of the variability of Y attributed to random factors, which is represented by the sum of squares SS_E, the variability of Y attributed to factor X_A, which is represented by the sum of squares SS_A, the variability of Y attributed to factor X_B, which is represented by the sum of squares SS_B, and the variability of Y attributed to the interaction of factors X_A and X_B, which is represented by the sum of squares SS_{AB}.

$$SS_T = SS_E + SS_A + SS_B + SS_{AB}$$

8.2.2.3 VARIATION ATTRIBUTED TO RANDOM FACTORS

The within-level variation of the response variable Y is observed when the combination $\{X_A^i, X_B^j\}$ of levels of factors X_A and X_B is fixed. This variation is attributed to random factors. In the two-way analysis of variance, the within level variation of Y is represented by a sum of squares SS_E. This is a sum of differences between the mean value of the response variable $\bar{y}^{i,j}$ associated with the $\{i,j\}$ combination of levels of factors X_A and X_B and every single value of the response variable $y_k{}^{i,j}$ recorded upon repeated measurements associated with this combination of factor levels.

$$SS_E = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{r}\left(\bar{y}^{i,j} - y_k^{i,j}\right)^2$$

Referring to Table 8.4, the SS_E describes the variation of Y inside a single cell of the table. It is aggregated for all the cells.

There are ν_E degrees of freedom associated with the within-level variance:

$$\nu_E = nm(r-1)$$

The within-level variance is given by the following mean square:

$$s^2(y)_E = MS_E = \frac{SS_E}{\nu_E}$$

8.2.2.4 VARIATION ATTRIBUTED TO THE NONRANDOM FACTOR X_A

The cross-level variation of the response variable Y which is attributed to the factor X_A is observed when levels of factor X_A are changed. In the two-way analysis of variance, the cross-level variation of Y caused by X_A is represented by a sum of squares SS_A. This is a sum of square differences between the overall mean value of the response variable $\bar{y}$ and the mean value of response variable $\bar{y}^i$ associated with every single level of factor X_A.

$$SS_A = \sum_{i=1}^{n}\left(\bar{y} - \bar{y}^i\right)^2$$

Referring to Table 8.4, SS_A describes the variation of Y among the rows of the table.

There are ν_A degrees of freedom associated with the cross-level variance attributed to factor X_A:

$$\nu_A = n - 1$$

The cross-level variance attributed to factor X_A is given by the following mean square:

$$s^2(y)_A = MS_A = \frac{SS_A}{\nu_A}$$

8.2.2.5 VARIATION ATTRIBUTED TO THE NONRANDOM FACTOR X_B

The cross-level variation of the response variable Y, which is attributed to factor X_B, is observed when the levels of factor X_B are changed. In the two-way analysis of variance, the cross-level variation of Y, caused by X_B, is represented by the sum of squares SS_A. This is a sum of square differences between the overall mean value

of the response variable $\bar{y}$ and the mean value of response variable $\bar{y}^j$ associated with every single level of factor X_B.

$$SS_B = \sum_{j=1}^{m} \left(\bar{y} - \bar{y}^j\right)^2$$

Referring to Table 8.4, SS_B describes the variation of Y among the columns of the table.

There are ν_B degrees of freedom associated with the cross-level variance attributed to factor X_B:

$$\nu_B = m - 1$$

The cross-level variance attributed to factor X_B is given by the following mean square:

$$s^2(y)_B = MS_B = \frac{SS_B}{\nu_B}$$

8.2.2.6 VARIATION ATTRIBUTED TO THE INTERACTION BETWEEN TWO NONRANDOM FACTORS

The variation of the response variable Y attributed to the interaction between factors X_A and X_B, is observed when combinations $\{X_A^i, X_B^j\}$ of levels of factors X_A and X_B are changed. In the two-way analysis of variance, the cross-level variation of Y, caused by the interaction between factor X_A and X_B, is represented by a sum of squares SS_{AB}. This is a sum of squared differences between the mean value of response variable $\bar{y}^{i,j}$ associated with the $\{i,j\}$ combination of levels of the factors X_A and X_B increased by overall mean of the response variable $\bar{y}$, and the sum of the mean value of the response variable $\bar{y}^i$, associated with the i^{th} level of factor X_A increased by the mean value of the response variable $\bar{y}^j$, associated with the j^{th} level of factor X_B.

$$SS_{AB} = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{r} \left(\left(\bar{y}^{i,j} + \bar{y}\right) - \left(\bar{y}^i + \bar{y}^j\right)\right)^2$$

Referring to Table 8.4, SS_{AB} describes the variation of Y among the cells of the table.

There are ν_{AB} degrees of freedom associated with the cross-level variance attributed to the interaction between factors X_A and X_B:

$$\nu_{AB} = (n-1)(m-1)$$

The cross-level variance attributed to the combination of factors X_A and X_B is given by the following mean square:

$$s^2(y)_{AB} = MS_{AB} = \frac{SS_{AB}}{\nu_{AB}}$$

8.2.3 NULL HYPOTHESES IN TWO-WAY ANOVA

Three null hypotheses are considered in the two-way analysis of variance. The first is used for testing the influence of factor X_A on the response variable Y. The second is used for testing the influence of factor X_B on the response variable Y. The third is used for testing the influence of the interaction between factors X_A and X_B on the response variable Y. All three hypotheses are tested in parallel. The result of the two-way analysis of variance consists of the summarized results of their testing.

1. NULL HYPOTHESIS ON FACTOR X_A

The null hypothesis, which tests the influence of factor X_A on the response variable, states that the average responses of the object to different levels of factor X_A are the same in the whole range of variability of factor X_A. In other words, on average, the object responds in the same way to any level of factor X_A. It is insensitive to changes in this factor. The formal representation of the null hypothesis is the following:

$$H_0: \mu^1 = \cdots = \mu^i = \cdots = \mu^n$$

The null hypothesis is tested versus the alternative hypothesis which states that the mean responses of the object are different for at least two different levels of factor X_A. In other words, the object is sensitive to the change between at least two levels of factor X_A. The formal representation of the alternative hypothesis is the following:

$$H_a: \exists \mu^i \neq \mu^l$$

where $i = 1..n$, $l = 1..n$ and $i \neq l$.

The following test statistic is used for testing the null hypothesis which refers to the influence of factor X_A on the response variable:

$$F_{cal} = \frac{s^2(y)_A}{s^2(y)_E}$$

If the null hypothesis is true, the variable F_{cal} has F-Snedecore distribution with the degrees of freedom ν_A and ν_E.

The critical interval criterion of null hypothesis rejection at the significance level α is

$$P\left(F \geq F_{\alpha,\nu_A,\nu_E}\right) = \alpha$$

where: F_{α,ν_A,ν_E} is the value of variable F, which comes the F-Snedecore distributions with the degrees of freedom ν_A and ν_E, for the assumed value of α.

The critical interval for F_{cal} is $F\epsilon\langle F_{\alpha,\nu_A,\nu_E},\infty\rangle$. If $F_{cal}\epsilon\langle F_{\alpha,\nu_A,\nu_E},\infty\rangle$, the null hypothesis is rejected.

The p-value criterion of null hypothesis rejection at the significance level α is

$$p = P\left(F_{\nu_A,\nu_E} \geq F_{cal}\right) \leq \alpha$$

The null hypothesis is rejected at the significance level α if α is greater than or equal to p.

For the graphical interpretation of the null hypothesis rejection criteria refer to Fig. 7.12 and Fig. 7.13.

Based on the presented reasoning, the null hypothesis is rejected when the test statistic F_{cal} reaches or exceeds F_{α,ν_A,ν_E}. The test statistic is the ratio between the variance of response variable $s^2(y)_A$, which comes from factor X_A, and the variance of response variable $s^2(y)_E$ which is caused by random factors. Therefore, the null hypothesis is rejected if the variation of Y caused by factor X_A is large enough when compared to its variation caused by random factors that the critical level F_{α,ν_A,ν_E} is reached. The rejection of the null hypothesis indicates that the considered factor X_A does significantly influence the object if represented by the response variable Y.

The null hypothesis is accepted on the condition that the ratio between the variance of response variable $s^2(y)_A$, which comes from factor X_A, and the variance of response variable $s^2(y)_E$, which is caused by random factors, does not exceed F_{α,ν_A,ν_E}. That is, the variation of Y caused by factor X_A is small enough when compared to its variation caused by random factors that the critical level F_{α,ν_A,ν_E} is not reached. The acceptance of the null hypothesis indicates that the considered factor X_A does not significantly influence the object if represented by the response variable Y.

2. NULL HYPOTHESIS ON FACTOR X_B

The null hypothesis testing the influence of factor X_B on the response variable is the following:

$$H_0: \mu^1 = \cdots = \mu^j = \cdots = \mu^m$$

The null hypothesis claims that the average responses of the object to different levels of factor X_B are the same in the whole range of variability of factor X_B. In other words, on average, the object responds in the same way to any level of factor X_B. It is insensitive to changes in this factor.

The null hypothesis is tested versus the alternative hypothesis:

$$H_a: \exists \mu^j \neq \mu^l$$

where $i = 1..m$, $l = 1..m$ and $i \neq l$.

The alternative hypothesis states that the average responses of the object are different in case of at least two levels of factor X_B. In other words, the object is sensitive to the change between at least two levels of factor X_B.

The null hypothesis referring to the influence of factor X_B on the response variable is tested using the following test statistic:

$$F_{cal} = \frac{s^2(y)_B}{s^2(y)_E}$$

If the null hypothesis is true, the variable F_{cal} has F-Snedecore distribution with the degrees of freedom ν_B and ν_E.

The critical interval criterion of null hypothesis rejection at the significance level α is

$$P\left(F \geq F_{\alpha,\nu_B,\nu_E}\right) = \alpha$$

where: F_{α,ν_B,ν_E} is the value of variable F, which comes the F-Snedecore distributions with the degrees of freedom ν_B and ν_E, for the assumed value of α.

The critical interval for F_{cal} is $F \epsilon \langle F_{\alpha,\nu_B,\nu_E}, \infty \rangle$. If $F_{cal} \epsilon \langle F_{\alpha,\nu_B,\nu_E}, \infty \rangle$, the null hypothesis is rejected.

The p-value criterion of null hypothesis rejection at the significance level α is

$$p = P\left(F_{\nu_B,\nu_E} \geq F_{cal}\right) \leq \alpha$$

The null hypothesis is rejected at the significance level α if α is greater than or equal to p.

For the graphical interpretation of the null hypothesis rejection criteria refer to Fig. 7.12 and Fig. 7.13.

Based on the presented reasoning, the null hypothesis is rejected when the test statistic F_{cal} reaches or exceeds F_{α,ν_B,ν_E}. The test statistic is the ratio between the variance of response variable $s^2(y)_B$, which comes from factor X_B, and the variance of response variable $s^2(y)_E$, which is caused by random factors. Therefore, the null hypothesis is rejected if the variation of Y caused by factor X_B is large enough when compared to its variation caused by random factors that the critical value F_{α,ν_B,ν_E} is reached. The rejection of the null hypothesis indicates that the considered factor X_B does significantly influence the object if represented by the response variable Y.

The null hypothesis is accepted on the condition that the ratio between the variance of response variable $s^2(y)_B$, which comes from factor X_B, and the variance of response variable $s^2(y)_E$, which is caused by random factors, does not

exceed F_{α,ν_B,ν_E}. That is, the variation of Y caused by factor X_B is small enough when compared to its variation caused by random factors that the critical value F_{α,ν_B,ν_E} is not reached. The acceptance of the null hypothesis indicates that the considered factor X_B does not significantly influence the object if represented by the response variable Y.

3. NULL HYPOTHESIS ON THE INTERACTION BETWEEN FACTORS X_A AND X_B

The null hypothesis testing the influence of the interaction between factors X_A and X_B on the response variable is the following:

$$H_0: \mu^{1,1} = \cdots = \mu^{i,j} = \cdots = \mu^{n,m}$$

This is tested versus the alternative hypothesis:

$$H_a: \exists \mu^{i,j} \neq \mu^{l,o}$$

where: $i = 1..n, l = 1..n, j = 1..m, o = 1..m$ and $i \neq l$ or $j \neq o$.

The null hypothesis claims that the average responses of the object to different combinations of levels of factors X_A and X_B are the same in the whole range of variability of both factors. In other words, on average, the object responds in the same way to any combination of levels of factors X_A and X_B. The object is insensitive to the changes in the combination of levels for the two factors. The alternative hypothesis states that average responses of the object are not the same in case of at least two different combinations of levels of factors X_A and X_B. In other words, the object is sensitive to the change between at least two combinations of levels of factors X_A and X_B. This implies a sensitivity to the interaction between factors.

The null hypothesis referring to the influence of the interaction between factors X_A and X_B on the response variable is tested using the following test statistic:

$$F_{cal} = \frac{s^2(y)_{AB}}{s^2(y)_E}$$

If the null hypothesis is true, the variable F_{cal} has F-Snedecore distribution with the degrees of freedom ν_{AB} and ν_E.

The critical interval criterion of null hypothesis rejection at the significance level α is

$$P\left(F \geq F_{\alpha,\nu_{AB},\nu_E}\right) = \alpha$$

where: $F_{\alpha,\nu_{AB},\nu_E}$ is the value of variable F, which comes the F-Snedecore distributions with the degrees of freedom ν_{AB} and ν_E, for the assumed value of α.

The critical interval for F_{cal} is $F \epsilon \langle F_{\alpha,\nu_{AB},\nu_E}, \infty)$. If $F_{cal} \epsilon \langle F_{\alpha,\nu_{AB},\nu_E}, \infty)$, the null hypothesis is rejected.

The p-value criterion of null hypothesis rejection at the significance level α is

$$p = P\left(F_{\nu_{AB},\nu_E} \geq F_{cal}\right) \leq \alpha$$

The null hypothesis is rejected at the significance level α if α is greater than or equal to p.

For the graphical interpretation of the null hypothesis rejection criteria refer to Fig. 7.12 and Fig. 7.13.

Based on the presented reasoning, the null hypothesis is rejected when the test statistic F_{cal} reaches or exceeds $F_{\alpha,\nu_{AB},\nu_E}$. The test statistic is the ratio between the variance of response variable $s^2(y)_{AB}$, which comes from the interaction between factors X_A and X_B, and the variance of response variable $s^2(y)_E$, which is caused by random factors. Therefore, the null hypothesis is rejected if the variation of Y caused by the interaction between factors X_A and X_B is large enough when compared to its variation caused by random factors that the critical value $F_{\alpha,\nu_{AB},\nu_E}$ is reached. The rejection of the null hypothesis indicates that the interaction between factors X_A and X_B does significantly influence the object if represented by the response variable Y.

The null hypothesis is accepted on the condition that the ratio between the variance of response variable $s^2(y)_{AB}$, which comes from the interaction between factors X_A and X_B, and the variance of response variable $s^2(y)_E$, which is caused by random factors does not exceed $F_{\alpha,\nu_{AB},\nu_E}$. That is, the variation of Y caused by the interaction between factors X_A and X_B is small enough when compared to its variation caused by random factors that the critical value $F_{\alpha,\nu_{AB},\nu_E}$ is not reached. The acceptance of the null hypothesis indicates that the interaction between factors X_A and X_B does not significantly influence the object if represented by the response variable Y.

8.2.4 EXAMPLE

Problem. The owner of the greenhouse wants to buy soil and fertilizer in order to grow a new variety of plant. It is important to know whether the kind of soil and the kind of fertilizer influences the fruitage of the plant. Otherwise any soil and any fertilizer is good.

The owner of the greenhouse performed an agricultural experiment which could help him in selecting the right soil and fertilizer. Namely, he grew plants on different soils, he used different fertilizers and he observed the fruitage. The fruitage was indicated by the number of pieces of fruit delivered by a single plant. This was considered as the response variable Y. The fruitage was influenced by two factors. The first factor was the type of soil. It was denoted X_A. The factor had three levels X_A^1, X_A^2 and X_A^3, which were three different types of soil. The second factor was the type of fertilizer. It was denoted X_B. This factor had four levels X_B^1, X_B^2, X_B^3, and X_B^4, which were four different types of fertilizer. Seven plants were

grown for each combination of soil-fertilizer. The results of the experiment are shown in Table 8.5.

8.5 Table of the measurement data for the problem considered in Example 9.2.4.

X_A \ X_B	X_B^1					X_B^2					X_B^3					X_B^4				
X_A^1	33	15	31	24	34	36	34	39	34	29	26	29	26	30	34	26	25	24	24	31
X_A^2	29	25	29	19	36	26	15	27	27	27	25	25	21	29	28	29	33	36	25	23
X_A^3	33	28	29	25	31	43	38	31	26	47	34	30	33	27	37	43	30	28	32	39

Solution. It is possible to study the problem using the two-way analysis of variance. It is worth testing three null hypotheses:

1. The fruitage of the plant is the same irrespective of the soil used, $H_{01}: \mu^1 = \mu^2 = \mu^3$, versus the alternative hypothesis that at least two different soils provide different fruitage $H_{a1}: \exists\, \mu^i \neq \mu^l$, $i = 1 \dots 3$, $l = 1 \dots 3$.
2. The fruitage of the plant is the same irrespective of the fertilizer used, $H_{02}: \mu^1 = \mu^2 = \mu^3 = \mu^4$, versus the alternative hypothesis that at least two different fertilizers provide different fruitage $H_{a2}: \exists\, \mu^i \neq \mu^l$, $i = 1 \dots 4$, $l = 1 \dots 4$.
3. The fruitage of the plant is the same irrespective of the combination of the soil and fertilizer used, $H_{03}: \mu^{11} = \mu^{12} = \dots = \mu^{43} = \mu^{44}$, versus the alternative hypothesis that at least two different soils provide different fruitage, $H_{a3}: \exists \mu^{i,j} \neq \mu^{l,o}$, $i = 1 \dots 3$, $l = 1 \dots 3$, $j = 1 \dots 4$, $o = 1 \dots 4$.

The relevant calculation help is offered by the DATA ANALYSIS TOOL in Excel. The results of the two-way ANOVA are shown in Table 8.6.

Table 8.6 ANOVA table for the measurement data shown in Table 9.5.

Source of variance	SS	ν	MS	F_{cal}	*p-value*	$F_{\alpha=0.05}$
X_A	$SS_A = 436.156$	$\nu_A = 2$	$MS_A = 218.078$	7.644	0.0013	3.191
X_B	$SS_B = 130.775$	$\nu_B = 3$	$MS_B = 43.592$	1.528	0.2193	2.798
$X_A X_B$	$SS_{AB} = 328.430$	$\nu_{AB} = 6$	$MS_{AB} = 54.738$	1.919	0.0969	2.295
random factor	$SS_E = 1369.391$	$\nu_E = 48$	$MS_E = 28.529$			
total	$SS_T = 2264.751$	$\nu_T = 59$				

The obtained results of null hypotheses testing at the significance level $\alpha = 0.05$, based on ANOVA are shown in Table 8.6:

1. the criterion of rejection of the null hypothesis H_{01} was fulfilled because it was shown that $F_{cal} \geq F_{\alpha,\nu_A,\nu_E}$,
2. the criterion of rejection of the null hypothesis H_{02} was not fulfilled because it was shown that $F_{cal} < F_{\alpha,\nu_B,\nu_E}$,
3. the criterion of rejection of the null hypothesis H_{03} was not fulfilled because it was shown that $F_{cal} < F_{\alpha,\nu_{AB},\nu_E}$.

Based on the performed analysis, the owner of the greenhouse is able to infer that the type of soil influences the fruitage of the plant while the type of fertilizer does not at the significance level $\alpha = 0.05$. Also, a significant interaction between the soil and the fertilizer concerning the fruitage of the plant was not observed.

8.3 PAIRWISE COMPARISON - FISHER'S LEAST SIGNIFICANT DIFFERENCE (LSD) METHOD

The analysis of variance examines the change of an object as a result of being influenced by different factors. If results of ANOVA/MANOVA show that the object is sensitive to a factor, further and more detailed questions may be asked. For example: How big is the change of a factor which makes the object respond? Is the size of change independent from the initial level of the factor?

Pairwise comparison is a method useful for solving this kind of problem. It consists of comparing mean values of the response variable associated with various levels of the considered factor. The differences between the means are evaluated versus a certain reference regarding their statistical significance. The formula describing the reference depends on the selected method of pairwise comparison.

It is worth to employ pairwise comparison if the considered factor is a nominal or ordinal variable. Otherwise, a regression analysis may be attempted, which is still more informative (see Chapter 9).

Fisher's Least Significant Difference (LSD) method was selected for presentation in this book as an exemplary pairwise comparison method. In the framework of this approach, the reference is the least significant difference which is defined in the following way:

$$LSD = t_{\alpha,\nu}\sqrt{MS_E\left(\frac{1}{r_i}+\frac{1}{r_l}\right)}$$

where: t is the variable which has t-Student distribution, α is the significance level, ν are degrees of freedom associated with MS_E, MS_E is the mean square representing the within level variance of the collected measurement data, r_i is the number of replicate measurements at the i^{th} level of considered factor, r_l is the number of replicate measurements at the l^{th} level of the considered factor.

The difference between responses of the object to two different levels of the factor is compared with the LSD. The difference is considered significant at the significance level α if the following is true:

$$\left|\bar{y}^i - \bar{y}^l\right| \geq LSD$$

where: $\bar{y}^i$ is the average object response to the i^{th}level of the factor and $\bar{y}^l$ is the average object response to the l^{th} level of the factor.

The difference is considered insignificant, at the significance level α, if the following is true:

$$\left|\bar{y}^i - \bar{y}^l\right| < LSD$$

Calculations are done for each pair (i, l) of levels of the factor.

8.3.1 EXAMPLE

Problem. It was shown in the solution of Example 8.1.4 that the method of dishwashing significantly influenced the cost of dishwashing. It is interesting to find out which methods are really different in that respect.

Solution. It is possible to solve the problem using pairwise comparison. For example, the least significant difference method may be employed. The considered response variable Y is the cost of dishwashing while the considered factor X_A is the method of dishwashing.

The mean value of the response variable associated with each level of factor X_A, was calculated as shown in Table 8.7.

Table 8.7 Mean value of the response variable associated with each level of factor $\boldsymbol{X_A}$.

Level of factor X_A	Level description	mean value of Y
X_A^1	ordinary manual dishwashing	$\bar{y}^1 = 0.5167$
X_A^2	economic manual dishwashing	$\bar{y}^2 = 0.1533$
X_A^3	washing with a dishwasher	$\bar{y}^3 = 0.8303$

Assuming the significance level $\alpha = 0.05$, the LDS is:

$$LSD = t_{\alpha,v}\sqrt{MS_E\left(\frac{1}{r_i}+\frac{1}{r_l}\right)} = t_{0.05,6}\sqrt{0.0208\left(\frac{1}{3}+\frac{1}{3}\right)} = 2.447\sqrt{0.0208\left(\frac{1}{3}+\frac{1}{3}\right)}$$
$$= 0.2882$$

The value of $t_{\alpha,v} = t_{0.05,6}$ was found in the tables of the t-Student distribution. All the other values were found in the corresponding ANOVA table (Table 8.3). Due to the equal number of replicate measurements at each level of the factor X_A, the LSD is the same for all the pairs of compared levels of factor X_A.

The table of pairwise comparisons is presented in Table 8.8.

Table 8.8 Table of pairwise comparisons.

Pair of levels of factor X_A, i-l	$\left\|\bar{y}^i - \bar{y}^l\right\|$	LSD	Conclusion
1 - 2	|0.3643|	0.2882	$\bar{\mu}^1 \neq \bar{\mu}^2$
1 - 3	|-0.3136|	0.2882	$\mu^1 \neq \bar{\mu}^3$
2 - 3	|-0.677|	0.2882	$\bar{\mu}^2 \neq \bar{\mu}^3$

As shown in Table 8.8, the difference $\left|\bar{y}^i - \bar{y}^l\right|$ is greater than the LSD for any two levels $\{i, l\}$ of factor X_A. Therefore, changing between any two levels of the considered factor X_A caused significant change in the response variable Y, at the significance level $\alpha = 0.05$.

Based on the performed statistical analysis, it is inferred that changing between any two methods of dishwashing caused significant change in the costs of dishwashing. Additionally, by looking at values of differences $\bar{y}^i - \bar{y}^l$ and at their signs, one may notice that the most disadvantageous was the replacement of economic manual dishwashing by the dishwasher. Switching between the economic manual dishwashing and ordinary manual dishwashing increased the cost in a similar manner as changing ordinary manual dishwashing for the dishwasher. Surprisingly, the analysis has shown that machine dishwashing is the least beneficial (largest $\bar{y}^3$) .

9 REGRESSION ANALYSIS

The problem of objects changing as a result of being influenced by different factors may be studied in various aspects. The analysis of variance, which was introduced in Chapter 8, is adequate for detecting the statistically significant change of an object. However, in many cases such conclusions are insufficient. A more advanced approach consists of forming a quantitative description of object change resulting from the influence of factors.

The quantitative description of the relationship between the values of factors and the values of a response variable is of great practical importance. For example, knowing this relationship an engineer is able to predict the response variable based on values of factors. Also, the engineer may be able to identify the values of factors which make the response variable take a particular, desired value.

Regression analysis is used for the quantitative representation of the relationship between two or more random variables. Regarding their status in the relationship, variables are divided into two groups: independent, also called explanatory, or predictor variables and dependent, also called response variables. The main idea of the regression analysis is to explain the variability of the dependent variable using the variability of independent variables.

There are several types of regression regarding the number of independent variables. The most frequently used in engineering applications are:

- Simple or univariate regression. It is used for representing the relationship between one independent variable and one dependent variable.
- Multiple regression. It is used for representing the relationship between several independent variables and one dependent variable.

Considering the kind of mathematical relationship regarding model parameters:

- Linear regression. Observational data are modeled by a function which is a linear combination of the model parameters.
- Nonlinear regression. Observational data are modeled by a function which is a nonlinear combination of the model parameters.

Regression analysis consists of building a regression model and its diagnostics.

One shall distinguish two kinds of relationships described using regression analysis. Some relationships have a cause–response character while others represent only correlations. The difference is substantial from a practical point of view. The cause-response relationship is when the dependent variable is really influenced by independent variable(s), i.e. the change of an independent variable causes the change in a dependent variable. The relationship having a correlation character occurs when the dependent variable varies together in a synchronized manner with the independent variable(s). The change of independent variable(s) does not cause the change of the dependent variable, but there is another, third factor which

influences both variables in a cause-response manner and makes them change together in a correlated way.

The regression model which represents a cause-response relationship shall be built exclusively using the measurement data provided in the course of an active experiment (see Chapter 2). This model may be used for prediction purposes.

The regression model which has a correlation character may be built using the measurement data provided in course of a passive experiment (see Chapter 2). It is not allowed to use this model for prediction purposes unless the theoretical justification of the cause-response relationship is available.

9.1 REGRESSION MODEL

The general form of the **regression model** is the following:

$$Y = f(\vec{X}, \vec{\beta}) + \varepsilon$$

where: Y is the dependent variable, f indicates a mathematical function, $\vec{X} = [X_1, X_2, \dots, X_k]$ is the vector of k independent variables, $\vec{\beta}$ is the vector of coefficients in the regression equation, ε is a random component.

The regression model states that the total variability of the dependent variable Y is composed of two elements. The first element is the deterministic component $f(\vec{X}, \vec{\beta})$, which can be described using the mathematical function f. This element is also referred to as $\hat{Y}$, which indicates the part of variable Y accounted for by the deterministic part of the regression model. The second element of the regression model is the random component ε. It is the difference between the actual measured variable Y and its part which is calculated from the deterministic part of the regression model:

$$\varepsilon = Y - f(\vec{X}, \vec{\beta})$$

$$\varepsilon = Y - \hat{Y}$$

The random component is also referred to as a residual or an error.

Principal assumptions upon the regression analysis refer to a random component. These are: the mean of random component is zero; the variance or random component is constant across the observations and it is independent of $\vec{X}$; the random component is not autocorrelated. Another important assumption refers to the independent variables and it states they should be uncorrelated.

From the computational point of view, the regression analysis is aimed at calculating the vector of coefficients $\vec{\beta}$. The resulting deterministic component of the model shall allow for good separation of the variability of the dependent variable caused by the deterministic factors from the variability resulting from random component. The type of function f is either known or its convenient form is assumed, e.g. linear. It is required that the number of data points which are used

in calculations exceed the number of coefficients in the regression equation. Otherwise, there is not enough data to calculate all coefficients or the coefficients may be obtained directly using a set of algebraic equations.

The most commonly applied strategy aimed at calculating coefficients in a regression equation is called the ordinary least-squares method (*OLSM*). The main idea of this method is to minimize the sum of squared distances between the variable Y and the deterministic component of the regression model $\hat{Y}$ for all data points which are used for building the regression model.

For explaining the concept behind the LSM, the case of univariate linear regression is considered here. The simple regression model has the following form:

$$y = \beta_1 x + \beta_0 + \varepsilon$$

A scatter plot representing an example of the relationship between the random variables Y and X may be described using simple regression as shown in Fig. 9.1.

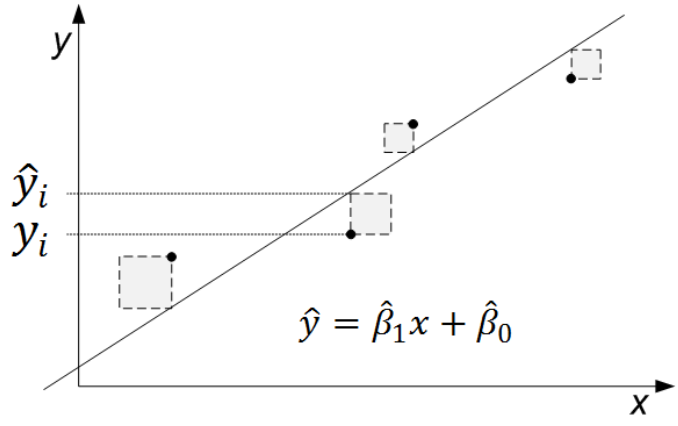


Figure 9.1 Scatter plot of the relationship between random variables **Y** and **X**, which may be described using simple regression.

In the case of simple regression the vector of model coefficients $\vec{\beta}$ consists of two elements: β_1 and β_0. As a result of using the *OLSM* for calculating the values of β_1 and β_0 for the regression line, their estimates $\hat{\beta}_1$ and $\hat{\beta}_0$ will have such values that the location of the regression line will be driven by the minimization of the sum of square areas, which are shown in Fig. 9.1. If the criterion of the minimum sum of squares, i.e. $\min(\sum_{i=1}^{n}(y_i - \hat{y}_i))$ is fulfilled, the estimates $\hat{\beta}_1$, $\hat{\beta}_0$ are calculated from the following equations:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 x$$

The ordinary least squares method is also applicable in the case of multiple linear regression. The multiple linear regression model is the following:

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k + \beta_0 + \varepsilon$$

where k is the number of independent variables in the regression model. It is recommended to use statistical software or adequate tollboxes for calculating the values of regression coefficients $\beta_0, \beta_1, \ldots, \beta_k$ in the case of multiple linear regression. Manual calculations are too complex and time consuming. The reader is referred to the relevant functionality available in Excel.

In the case that the ranges of independent variables are very different, e.g. they differ by one or more order of magnitude, it is recommended to standardize the variables (see §5.2) before including them in the regression model.

9.2 DIAGNOSTICS OF THE REGRESSION MODEL

A number of diagnostic tools are available for checking the quality of the regression model. Regression model diagnostic tools may be divided into two groups. The first group is used for checking if the particular regression model is the right selection for describing the relationship between the dependent variable and independent variable(s). This group includes statistical tests of significance for the entire model, statistical tests of significance of coefficients in the regression model and tests dedicated to verifying the assumptions which were made prior to model construction. The second group of tools is used for assessing the goodness-of-fit, i.e. how well the regression model explains the variability of the response variable Y. The most useful tools are the diagnostic plot, coefficient of determination and standard error.

9.2.1 SIGNIFICANCE OF THE REGRESSION MODEL

The significance of the regression model is investigated by testing the corresponding statistical hypothesis. The null hypothesis states that all the coefficients which stand next to the independent variables in the regression model are equal zero. The formal representation of the hypothesis is the following:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

If the null hypothesis is true, the regression model, for example multiple regression model:

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k + \beta_0 + \varepsilon$$

is reduced to the form:

$$y = \beta_0 + \varepsilon$$

which indicates that none of the independent variables contribute to explaining the variation of the dependent variable.

The null hypothesis is tested versus the alternative which states that at least one coefficient in the regression model is different from zero. The formal representation of the alternative hypothesis is the following:

$$H_A: \exists \beta_j \neq 0$$

The test statistic employed for null hypothesis testing is

$$F_{cal} = \frac{MS_R}{MS_E}$$

where MS_R is the mean square regression and it is calculated as follows:

$$MS_R = \frac{SS_R}{k} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{k}$$

and MS_E is the mean square error and it is calculated using the formula:

$$MS_E = \frac{SS_E}{n-k-1} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-k-1}$$

where: SS_R is the sum of square regression, SS_E is the sum of square error, n is the number of data points used for developing the regression model, k is the number of independent variables in the model, $\hat{y}_i$ is the i^{th} calculated value of the response variable, $\bar{y}$ is the mean of the measured values of the response variable, y_i is the i^{th} measured value of the response variable.

If the null hypothesis is true, the test statistic has F-Snedecore distribution with, $\nu_1 = k$ and $\nu_2 = n - k - 1$ degrees of freedom.

The mean square regression, MS_R indicates the variability of the response variable calculated from the regression model around the mean of the response variable $\bar{Y}$. The mean square error, MS_E indicates the discrepancy between measured values of the response variable Y and values calculated from the regression model $\hat{Y}$.

The criterion of null hypothesis rejection at the significance level α, is the following:

$$p\left(F \geq F_{\alpha,\nu_1,\nu_2}\right) = \alpha \equiv p = P\left(F_{\nu_1,\nu_2} \geq F_{cal}\right) \leq \alpha$$

The critical interval for F_{cal} is $F \epsilon \langle F_{\alpha,\nu_1,\nu_2}, \infty \rangle$. If $F_{cal} \epsilon \langle F_{\alpha,\nu_1,\nu_2}, \infty \rangle$, the null hypothesis is rejected. The same holds if the p-value is less than or equal to the significance level α.

For the graphical interpretation of the criteria of null hypothesis rejection see Fig. 7.12 and Fig. 7.13.

The rejection of the null hypothesis is synonymous with considering the regression model as significant, i.e. able to explain the variability of the response variable with a set of independent variables at the significance level α.

The acceptance of the null hypothesis is synonymous with considering the regression model as insignificant, i.e. unable to explain the variability of the response variable with a set of independent variables at the significance level α.

9.2.2 SIGNIFICANCE OF REGRESSION MODEL COEFFICIENTS

The test on the significance of a regression model refers to the entire model and does not offer any knowledge about the elements of the model. It may happen that the entire regression model is significant, but some of the model coefficients are statistically insignificant. If so, the independent variables which stand by these coefficients do not contribute much to the explanation of the variability of the dependent variable. One may consider removing them from the model, which results in model simplification.

The significance of a coefficient in the regression model is evaluated by testing the corresponding statistical hypothesis.

The null hypothesis on the significance of the k^{th} coefficient in the regression model states that this coefficient is equal to zero. The formal representation of the hypothesis is the following:

$$H_0: \beta_k = 0$$

The null hypothesis is tested versus the alternative hypothesis, which states that the coefficient is different from zero, as follows:

$$H_A: \beta_k \neq 0$$

The test statistic employed for null hypothesis testing is

$$t_{cal} = \frac{\hat{\beta}_k}{s_{\hat{\beta}_k}}$$

where: $\hat{\beta}_k$ is the estimate of coefficient β_k in the regression model, $s_{\hat{\beta}_k}$ is the standard error of estimation of β_k. The formula describing $s_{\hat{\beta}_k}$ is out of the scope of this book. The reader shall understand that a large value of $s_{\hat{\beta}_k}$ indicates that the estimate of β_k with $\hat{\beta}_k$ is unstable, which is unwanted.

Significant coefficients are clearly different from zero and the error of their estimation is low. For such coefficients the value of t_{cal} is relatively high. Insignificant coefficients are close to zero and/or the error of their estimation is high. For such coefficients the value of t_{cal} is relatively low.

If the null hypothesis is true, the test statistic t_{cal} has t-Student distribution with $\nu = n - k - 1$ degrees of freedom.

The criterion of null hypothesis rejection at the significance level α is the following:

$$P\left(|t| \geq t_{\frac{\alpha}{2},\nu}\right) = \alpha \equiv p = P(|t| \geq t_{cal}) \leq \alpha$$

The critical interval for t_{cal} is $t\epsilon\left(-\infty,-t_{\frac{\alpha}{2},\nu}\right\rangle \cup \left\langle t_{\frac{\alpha}{2},\nu},\infty\right)$. If $t_{cal}\epsilon\left(-\infty,-t_{\frac{\alpha}{2},\nu}\right\rangle \cup \left\langle t_{\frac{\alpha}{2},\nu},\infty\right)$, the null hypothesis is rejected. The same holds if the p-value is less than or equal to the significance level α.

For the graphical interpretation of the criteria of null hypothesis rejection see Fig. 7.4 and Fig. 7.5.

The rejection of the null hypothesis for a particular coefficient β_k in the regression model is synonymous with considering the coefficient as significant at the significance level α. In other words, the independent variable X_k, which stands next to the coefficient, is considered as significantly contributing to the explanation of the variability of the response variable Y.

The acceptance of the null hypothesis for a particular coefficient β_k in the regression model is synonymous with considering the coefficient as insignificant at the significance level α. In other words, the independent variable X_k, which stands next to the coefficient, is considered as insignificantly contributing to the explanation of the variability of the response variable Y. One shall consider removing this variable from the regression model and recalculating.

The relevant hypothesis shall be formulated and tested for every coefficient in the model.

9.2.3 DIAGNOSTIC PLOT

The simplest diagnostic tool which indicates the goodness-of-fit has a graphical character. The diagnostic plot is a scatter plot. Values of the response variable, Y are represented on the horizontal axis and values calculated from the regression model, $\hat{Y}$ are represented on the vertical axis. A single point in the plot has the coordinates $(y_i, \hat{y}_i)$. Both values are associated with the i^{th} set of values of the independent variables $[x_1^i, x_2^i, \dots, x_k^i]$.

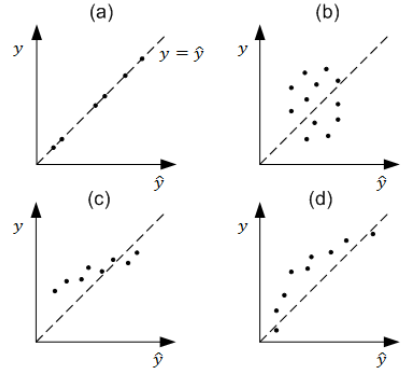


Figure 9.2 Examples of various diagnostic plots.

There are two extreme layouts of points in a diagnostic plot. The one shown in Fig. 9.2a indicates a perfect fit. The points in the diagnostic plot are located along the line $y = \hat{y}$. In this case, the regression model explains the entire variability of the response variable. This ideal case is unrealistic due to the existence of variability of

Y caused by random factors. If this identity is obtained in the course of regression model parameterization, it indicates that the data is overfitting. The other extreme layout is shown in Fig. 9.2b. This represents a lack of fit. The cloud of points in the diagnostic plot takes the form of a circular shape. In this case, the regression model is totally unable to explain the variability of the response variable. Between these two extremes there are scatter plots which show different degrees of goodness-of-fit. In general, the slim oval shape of the cloud of points along the line $y = \hat{y}$ indicates that the particular kind of regression model was a good selection. A smaller spread of points along the reference line $y = \hat{y}$, indicates a better fit. Also, more specific information is carried by the diagnostic plots. An example of a plot which indicates underestimation is presented in Fig. 9.2c. The cloud of points has a lower tilt than the reference line. The range of values of Y which are represented by the regression model is smaller than the entire range of the response variable. The inappropriateness of linear regression for representing the variability of Y is visible in the diagnostic plot shown in Fig. 9.2d. The bent form of the scatter indicates that there is a nonlinear component missing in the regression model.

9.2.4 COEFFICIENT OF DETERMINATION

The coefficient of determination is one of the basic diagnostic tools indicating the goodness-of-fit of experimental data by the regression model.

The coefficient of determination is calculated by the following formula:

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

The graphical representation of the idea behind the coefficient of determination is shown in Fig. 9.3 considering a single data point.

The denominator in the r^2 formula contains the difference $|y_i - \bar{y}|$. It is the key element in the formula describing the variance of variable Y. The difference tells the distance between the i^{th} value of the variable and the overall mean value of the variable (Fig. 9.3). Part of this distance is the difference $|\hat{y}_i - \bar{y}|$, which is placed in the nominator of the r^2 formula. It may be understood as the key element of variance of variable $\hat{Y}$. This part of the distance is accounted for by the regression model. The other part $|\hat{y}_i - y_i|$ remains unexplained. It comes from random factors and independent variables not included in the regression model.

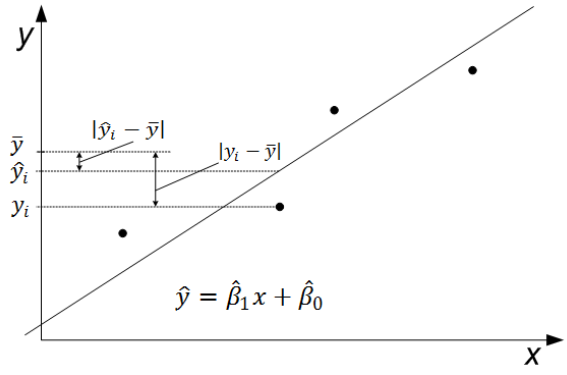


Figure 9.3 Illustration of the elements of the formula describing the coefficient of determination.

The coefficient of determination shows the fraction of variance of the response variable explained by the regression model. The coefficient takes values from the interval $r^2 \in \langle 0,1 \rangle$.

In the ideal case the entire variability of the response variable is explained by the regression model and $r^2 = 1$. In the worst imaginable case none of the variability of the response variable is explained by the regression model and $r^2 = 0$. The coefficient of determination indicates that the regression model well explains the variation of the response variable if its value is close to one. Such models are called adequate. Small values for the coefficient of determination are obtained for models which poorly explain the response variable. Close to zero values of r^2 indicate highly inadequate models.

In the case of multiple regression models, the basic formula for the coefficient of determination is slightly modified in order to obtain the corrected coefficient of determination, which shall be used for assessing the goodness-of-fit:

$$\bar{r}^2 = 1 - (1 - r^2)\frac{n-1}{n-k-1}$$

The correction prevents the increase of the value of this coefficient in case the number of independent variables is increased in the model while they do not contribute substantially to explaining the variance of the response variable. By using the significance test together with this coefficient, it is possible to point out redundant variables in the regression model and remove them.

9.3 PREDICTION WITH THE REGRESSION MODEL

One very useful application of the regression model is prediction. Prediction is the calculation of the value of the response variable for the set of values of independent variables. The principle restrictions to be obeyed when using the regression model for prediction concern the range of values of independent

variables. It is allowed to predict the response variable based on a regression model within the range of the values of independent variables considered while parameterizing the regression model. It is not allowed to predict the response variable with the regression model outside the range of the values of independent variables considered while parameterizing the regression model. The quality of prediction is quantifiable and the example of the relevant indicator is the standard error of prediction.

9.3.1 STANDARD ERROR

Standard error represents the distance between the real values of the response variable and its values obtained from the deterministic part of the regression model. The error formula is the following:

$$s = \sqrt{MS_E}$$

where: MS_E is the mean square error (see §9.2.1). The standard error is obtained in the units of the response variable.

Small values of standard error indicate good fit between the measured values of the response variable and their counterparts calculated from the regression model.

The standard error may be calculated for the pool of data which were used at the stage of model parameterization. In such case, it acts as the diagnostic tool. Also, standard error may be calculated for the pool of data which are different from those used at the stage of model parameterization. In such case, this measures the predictive ability of the model.

For the sake of obtaining a relative indicator, the standard error is referred to the average value of the response variable.

$$I = \frac{s}{\bar{y}} \cdot 100\%$$

Again, preferred indicator values are close to zero. Depending on their planned use, models characterized by up to 5 %, 10 % or even a 20% level of relative error may be considered satisfactory.

9.3.2 EXAMPLE

Problem. An engineer uses his car daily for driving to work, shopping and reaching many other destinations not far from home. He was interested in the relationship between fuel consumption and the distance driven as well as the number of stops encountered during travel. Stops are mainly enforced by traffic lights. He collected data concerning fuel consumption at various travel distances including the number of stops encountered during travel. They are shown in Table 9.1.

9.1 Experimental data concerning fuel consumption, travel distance and number of stops encountered during travel.

Fuel consumption (Y) / mL	Travel distance (X_1)/ km	Number of stops (X_2)	Fuel consumption (Y) / mL	Travel distance (X_1)/ km	Number of stops (X_2)
27	0.9	2	49	7.4	4
104.5	9.6	8	96.9	11.5	9
163.3	9.7	11	24.3	1	2
36	1.6	2	110	11.5	10
100	9.7	9	14.4	1.2	1
91.8	10.1	10	14.4	0.8	2
132.5	11	12	55.3	1.9	6
85	12	8	72	4.4	8
31.5	1.6	4	93.1	10.1	10
84.5	4.3	11	35.6	1.5	2
135	11.3	16	95.2	10.7	7
78	10.6	7	33.5	1.6	3
41	1.1	2	72	4.3	7
124.2	4.9	11	114.4	11.5	10
83.2	13.1	5	105	11.1	9
105	11.1	9			

Solution. It is possible to analyze the problem using multiple linear regression. One needs to assume that the relationship between the response variable - fuel consumption (Y) and the two independent variables: travel distance (X_1) and number of stops (X_2) is linear and it can be represented by the following equation:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_0 + \varepsilon$$

This is a good starting assumption as theoretical knowledge concerning the character of such a relationship is not available.

The relevant calculation help is offered by the DATA ANALYSIS TOOL in Excel. The results of the regression analysis are shown in Table 9.2 - 9.4.

Table 9.2 Regression analysis for the data shown in Table 9.1 - Regression statistics.

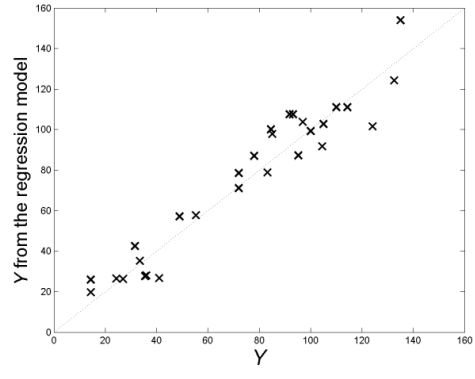
r^2	0.881
$\bar{r}^2$	0.872
MRE	14.126

Table 9.3 Regression analysis for the data shown in Table 9.1– Significance of the regression model.

	ν	SS	MS	F_{cal}	$F_{\alpha=0.05}$
Regression	2	$SS_R = 39725.97$	$MS_R = 19862.99$	99.55	$3.47E-13$
Random	27	$SS_E = 5387.34$	$MS_E = 199.53$		
Total	29	$SS_T = 45113.31$			

Table 9.4 Regression analysis for the data shown in Table 9.1 – Significance of model coefficients.

	$\hat{\beta}$	$s_{\hat{\beta}}$	$t_{\hat{\beta}}$	*p-value*
X_1	7.26	0.983	7.381	$6.1E-08$
X_2	2.53	0.840	3.013	0.00557
constant term	9.41	5.433	1.731	0.09478



9.4 Diagnostic plot for the regression model developed in example 10.3.2

Based on the results of the regression analysis shown in Table 9.2 - 9.4, a number of conclusions can be drawn about the considered regression model.

- The model is statistically significant at the significance level $\alpha = 0.05$. The condition of rejection of the null hypothesis about all model coefficients being zero is not fulfilled ($F_{cal} > F_{\alpha=0.05}$).
- The model offers high goodness-of-fit. The corrected coefficient of determination $\bar{r}^2 = 0.872$ has a high value.
- The multiple linear regression model was a good choice for representing the relationship between the considered variables. The points in the diagnostic plot do not retract systematically from the line $y = \hat{y}$.

- The two variables X_1 and X_2 contribute to the explanation of the response variable Y in a statistically significant manner. The associated p-values are smaller than $\alpha = 0.05$. The constant term $\hat{\beta}_0$ is not significant. The associated p-value is greater than α.
- The contribution of X_1 to the explanation of the response variable Y is over two times higher than the contribution of X_2.

Considering the real meaning of the variables included in the regression model, it is possible to infer about the relationship between the fuel consumption and travel distance together with the number of stops encountered during travel. Namely, there is a statistically significant relationship between these variables. The linear function is a good approximation of the relationship. Interestingly, the fuel consumption is more strongly influenced by the number of stops encountered during travel than by the travel distance. These conclusions are valid for the particular considered case, i.e. the car, the driver and the city. They do not have a general character.

APPENDICES

APPENDIX 1 NORMAL DISTRIBUTION

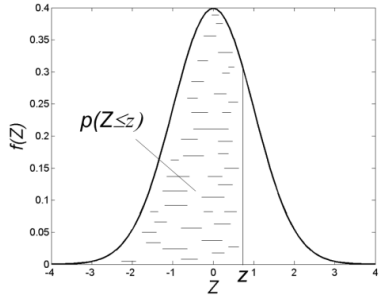


Table 1 $F(Z) = p(Z \leq z)$

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861

APPENDIX 2 T-STUDENT DISTRIBUTION

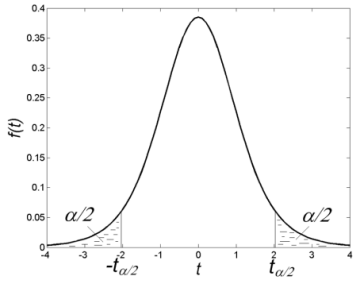


Table 2 $t_{\alpha,\nu}$ such that $p(t_{\alpha,\nu} \leq |t|) = \alpha$

ν	α										
	0.5	0.4	0.3	0.2	0.1	0.05	0.04	0.03	0.02	0.01	0.001
1	1.000	1.376	1.963	3.078	6.314	12.706	15.894	21.205	31.821	63.656	636.578
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	5.643	6.965	9.925	31.600
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	3.896	4.541	5.841	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.298	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.003	3.365	4.032	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	2.829	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.715	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.634	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.574	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.527	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.491	2.718	3.106	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.461	2.681	3.055	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.436	2.650	3.012	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.415	2.624	2.977	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.397	2.602	2.947	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.382	2.583	2.921	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.368	2.567	2.898	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.356	2.552	2.878	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.346	2.539	2.861	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.336	2.528	2.845	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.328	2.518	2.831	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.320	2.508	2.819	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.313	2.500	2.807	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.307	2.492	2.797	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.301	2.485	2.787	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.296	2.479	2.779	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.291	2.473	2.771	3.689
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.286	2.467	2.763	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.282	2.462	2.756	3.660
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.278	2.457	2.750	3.646

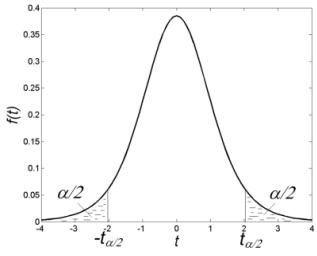


Table 2 continuation, $t_{\alpha,\nu}$ **such that** $p(t_{\alpha,\nu} \leq |t|) = \alpha$

ν	α										
	0.5	0.4	0.3	0.2	0.1	0.05	0.04	0.03	0.02	0.01	0.001
31	0.682	0.853	1.054	1.309	1.696	2.040	2.144	2.275	2.453	2.744	3.633
32	0.682	0.853	1.054	1.309	1.694	2.037	2.141	2.271	2.449	2.738	3.622
33	0.682	0.853	1.053	1.308	1.692	2.035	2.138	2.268	2.445	2.733	3.611
34	0.682	0.852	1.052	1.307	1.691	2.032	2.136	2.265	2.441	2.728	3.601
35	0.682	0.852	1.052	1.306	1.690	2.030	2.133	2.262	2.438	2.724	3.591
36	0.681	0.852	1.052	1.306	1.688	2.028	2.131	2.260	2.434	2.719	3.582
37	0.681	0.851	1.051	1.305	1.687	2.026	2.129	2.257	2.431	2.715	3.574
38	0.681	0.851	1.051	1.304	1.686	2.024	2.127	2.255	2.429	2.712	3.566
39	0.681	0.851	1.050	1.304	1.685	2.023	2.125	2.252	2.426	2.708	3.558
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.250	2.423	2.704	3.551
41	0.681	0.850	1.050	1.303	1.683	2.020	2.121	2.248	2.421	2.701	3.544
42	0.680	0.850	1.049	1.302	1.682	2.018	2.120	2.246	2.418	2.698	3.538
43	0.680	0.850	1.049	1.302	1.681	2.017	2.118	2.244	2.416	2.695	3.532
44	0.680	0.850	1.049	1.301	1.680	2.015	2.116	2.243	2.414	2.692	3.526
45	0.680	0.850	1.049	1.301	1.679	2.014	2.115	2.241	2.412	2.690	3.520
46	0.680	0.850	1.048	1.300	1.679	2.013	2.114	2.239	2.410	2.687	3.515
47	0.680	0.849	1.048	1.300	1.678	2.012	2.112	2.238	2.408	2.685	3.510
48	0.680	0.849	1.048	1.299	1.677	2.011	2.111	2.237	2.407	2.682	3.505
49	0.680	0.849	1.048	1.299	1.677	2.010	2.110	2.235	2.405	2.680	3.500
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.234	2.403	2.678	3.496
51	0.679	0.849	1.047	1.298	1.675	2.008	2.108	2.233	2.402	2.676	3.492
52	0.679	0.849	1.047	1.298	1.675	2.007	2.107	2.231	2.400	2.674	3.488
53	0.679	0.848	1.047	1.298	1.674	2.006	2.106	2.230	2.399	2.672	3.484
54	0.679	0.848	1.046	1.297	1.674	2.005	2.105	2.229	2.397	2.670	3.480
55	0.679	0.848	1.046	1.297	1.673	2.004	2.104	2.228	2.396	2.668	3.476
56	0.679	0.848	1.046	1.297	1.673	2.003	2.103	2.227	2.395	2.667	3.473
57	0.679	0.848	1.046	1.297	1.672	2.002	2.102	2.226	2.394	2.665	3.469
58	0.679	0.848	1.046	1.296	1.672	2.002	2.101	2.225	2.392	2.663	3.466
59	0.679	0.848	1.046	1.296	1.671	2.001	2.100	2.224	2.391	2.662	3.463
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.223	2.390	2.660	3.460

APPENDIX 3 χ^2 DISTRIBUTION

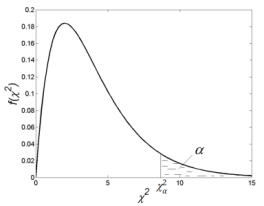


Table 3 $\chi^2_{\alpha,\nu}$ such that $\boldsymbol{p}\left(\chi^2 \geq \chi^2_{\alpha,\nu}\right) = \boldsymbol{\alpha}$

ν	α													
	0.999	0.99	0.98	0.97	0.96	0.95	0.94	0.93	0.92	0.91	0.90	0.80	0.70	0.60
1	0.000002	0.000157	0.000628	0.001414	0.002515	0.003932	0.005666	0.007717	0.010087	0.012778	0.015791	0.064185	0.148	0.275
2	0.002	0.020	0.040	0.061	0.082	0.103	0.124	0.145	0.167	0.189	0.211	0.446	0.713	1.022
3	0.024	0.115	0.185	0.245	0.300	0.352	0.401	0.449	0.495	0.540	0.584	1.005	1.424	1.869
4	0.091	0.297	0.429	0.535	0.627	0.711	0.788	0.862	0.931	0.999	1.064	1.649	2.195	2.753
5	0.210	0.554	0.752	0.903	1.031	1.145	1.250	1.347	1.439	1.526	1.610	2.343	3.000	3.656
6	0.381	0.872	1.134	1.330	1.492	1.635	1.765	1.885	1.997	2.103	2.204	3.070	3.828	4.570
7	0.599	1.239	1.564	1.802	1.997	2.167	2.320	2.461	2.592	2.716	2.833	3.822	4.671	5.493
8	0.857	1.647	2.032	2.310	2.537	2.733	2.908	3.068	3.217	3.357	3.490	4.594	5.527	6.423
9	1.152	2.088	2.532	2.848	3.105	3.325	3.521	3.700	3.866	4.021	4.168	5.380	6.393	7.357
10	1.479	2.558	3.059	3.412	3.697	3.940	4.157	4.353	4.535	4.705	4.865	6.179	7.267	8.295
11	1.834	3.053	3.609	3.997	4.309	4.575	4.810	5.024	5.221	5.405	5.578	6.989	8.148	9.237
12	2.214	3.571	4.178	4.601	4.939	5.226	5.480	5.710	5.921	6.118	6.304	7.807	9.034	10.182
13	2.617	4.107	4.765	5.221	5.584	5.892	6.163	6.409	6.634	6.844	7.041	8.634	9.926	11.129
14	3.041	4.660	5.368	5.856	6.243	6.571	6.859	7.120	7.359	7.581	7.790	9.467	10.821	12.078
15	3.483	5.229	5.985	6.503	6.914	7.261	7.566	7.841	8.093	8.327	8.547	10.307	11.721	13.030
16	3.942	5.812	6.614	7.163	7.596	7.962	8.283	8.572	8.836	9.082	9.312	11.152	12.624	13.983
17	4.416	6.408	7.255	7.832	8.288	8.672	9.008	9.311	9.588	9.845	10.085	12.002	13.531	14.937
18	4.905	7.015	7.906	8.512	8.989	9.390	9.742	10.058	10.347	10.614	10.865	12.857	14.440	15.893
19	5.407	7.633	8.567	9.200	9.698	10.117	10.483	10.812	11.112	11.391	11.651	13.716	15.352	16.850
20	5.921	8.260	9.237	9.897	10.415	10.851	11.231	11.573	11.884	12.173	12.443	14.578	16.266	17.809
21	6.447	8.897	9.915	10.601	11.140	11.591	11.986	12.339	12.662	12.961	13.240	15.445	17.182	18.768
22	6.983	9.542	10.600	11.313	11.870	12.338	12.746	13.112	13.445	13.753	14.041	16.314	18.101	19.729
23	7.529	10.196	11.293	12.030	12.607	13.091	13.512	13.889	14.233	14.551	14.848	17.187	19.021	20.690
24	8.085	10.856	11.992	12.754	13.350	13.848	14.283	14.672	15.026	15.353	15.659	18.062	19.943	21.652
25	8.649	11.524	12.697	13.484	14.098	14.611	15.059	15.459	15.823	16.159	16.473	18.940	20.867	22.616
26	9.222	12.198	13.409	14.219	14.851	15.379	15.839	16.250	16.624	16.970	17.292	19.820	21.792	23.579
27	9.803	12.878	14.125	14.959	15.609	16.151	16.624	17.045	17.429	17.783	18.114	20.703	22.719	24.544
28	10.391	13.565	14.847	15.704	16.371	16.928	17.412	17.844	18.238	18.601	18.939	21.588	23.647	25.509
29	10.986	14.256	15.574	16.454	17.138	17.708	18.204	18.647	19.050	19.421	19.768	22.475	24.577	26.475
30	11.588	14.953	16.306	17.208	17.908	18.493	19.000	19.453	19.865	20.245	20.599	23.364	25.508	27.442

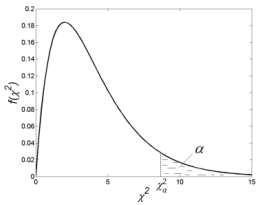


Table 3 continuation, $\chi^2_{\alpha,\nu}$ such that $\boldsymbol{p}\left(\chi^2 \geq \chi^2_{\alpha,\nu}\right) = \boldsymbol{\alpha}$

ν	α														
	0.50	0.40	0.30	0.20	0.10	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.001
1	0.455	0.708	1.074	1.642	2.706	2.874	3.065	3.283	3.537	3.841	4.218	4.709	5.412	6.635	10.827
2	1.386	1.833	2.408	3.219	4.605	4.816	5.051	5.319	5.627	5.991	6.438	7.013	7.824	9.210	13.815
3	2.366	2.946	3.665	4.642	6.251	6.491	6.759	7.060	7.407	7.815	8.311	8.947	9.837	11.345	16.266
4	3.357	4.045	4.878	5.989	7.779	8.043	8.337	8.666	9.044	9.488	10.026	10.712	11.668	13.277	18.466
5	4.351	5.132	6.064	7.289	9.236	9.521	9.837	10.191	10.596	11.070	11.644	12.375	13.388	15.086	20.515
6	5.348	6.211	7.231	8.558	10.645	10.948	11.283	11.660	12.090	12.592	13.198	13.968	15.033	16.812	22.457
7	6.346	7.283	8.383	9.803	12.017	12.337	12.691	13.088	13.540	14.067	14.703	15.509	16.622	18.475	24.321
8	7.344	8.351	9.524	11.030	13.362	13.697	14.068	14.484	14.956	15.507	16.171	17.011	18.168	20.090	26.124
9	8.343	9.414	10.656	12.242	14.684	15.034	15.421	15.854	16.346	16.919	17.608	18.480	19.679	21.666	27.877
10	9.342	10.473	11.781	13.442	15.987	16.352	16.753	17.203	17.713	18.307	19.021	19.922	21.161	23.209	29.588
11	10.341	11.530	12.899	14.631	17.275	17.653	18.069	18.533	19.061	19.675	20.412	21.342	22.618	24.725	31.264
12	11.340	12.584	14.011	15.812	18.549	18.939	19.369	19.849	20.393	21.026	21.785	22.742	24.054	26.217	32.909
13	12.340	13.636	15.119	16.985	19.812	20.214	20.657	21.151	21.711	22.362	23.142	24.125	25.471	27.688	34.527
14	13.339	14.685	16.222	18.151	21.064	21.478	21.933	22.441	23.017	23.685	24.485	25.493	26.873	29.141	36.124
15	14.339	15.733	17.322	19.311	22.307	22.732	23.199	23.720	24.311	24.996	25.816	26.848	28.259	30.578	37.698
16	15.338	16.780	18.418	20.465	23.542	23.977	24.456	24.990	25.595	26.296	27.136	28.191	29.633	32.000	39.252
17	16.338	17.824	19.511	21.615	24.769	25.215	25.705	26.251	26.870	27.587	28.445	29.523	30.995	33.409	40.791
18	17.338	18.868	20.601	22.760	25.989	26.445	26.947	27.505	28.137	28.869	29.745	30.845	32.346	34.805	42.312
19	18.338	19.910	21.689	23.900	27.204	27.669	28.181	28.751	29.396	30.144	31.037	32.158	33.687	36.191	43.819
20	19.337	20.951	22.775	25.038	28.412	28.887	29.410	29.991	30.649	31.410	32.321	33.462	35.020	37.566	45.314
21	20.337	21.992	23.858	26.171	29.615	30.100	30.632	31.225	31.895	32.671	33.597	34.759	36.343	38.932	46.796
22	21.337	23.031	24.939	27.301	30.813	31.307	31.849	32.453	33.135	33.924	34.867	36.049	37.659	40.289	48.268
23	22.337	24.069	26.018	28.429	32.007	32.510	33.062	33.675	34.370	35.172	36.131	37.332	38.968	41.638	49.728
24	23.337	25.106	27.096	29.553	33.196	33.708	34.269	34.893	35.599	36.415	37.389	38.609	40.270	42.980	51.179
25	24.337	26.143	28.172	30.675	34.382	34.902	35.472	36.106	36.824	37.652	38.642	39.880	41.566	44.314	52.619
26	25.336	27.179	29.246	31.795	35.563	36.091	36.671	37.315	38.044	38.885	39.889	41.146	42.856	45.642	54.051
27	26.336	28.214	30.319	32.912	36.741	37.278	37.866	38.520	39.259	40.113	41.132	42.407	44.140	46.963	55.475
28	27.336	29.249	31.391	34.027	37.916	38.460	39.058	39.721	40.471	41.337	42.370	43.662	45.419	48.278	56.892
29	28.336	30.283	32.461	35.139	39.087	39.640	40.246	40.919	41.679	42.557	43.604	44.913	46.693	49.588	58.301
30	29.336	31.316	33.530	36.250	40.256	40.816	41.430	42.113	42.883	43.773	44.834	46.160	47.962	50.892	59.702

APPENDIX 4 F-SNEDECORE DISTRIBUTION, α=0.01

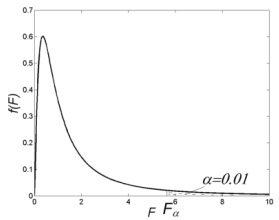


Table 4 $F_{\alpha,\nu}$ such that $p(F \geq F_{\alpha,\nu}) = \alpha$

	ν_1													
ν_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6083	6107	6126	6143
2	98.502	99.000	99.164	99.251	99.302	99.331	99.357	99.375	99.390	99.397	99.408	99.419	99.422	99.426
3	34.116	30.816	29.457	28.710	28.237	27.911	27.671	27.489	27.345	27.228	27.132	27.052	26.983	26.924
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546	14.452	14.374	14.306	14.249
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051	9.963	9.888	9.825	9.770
6	13.745	10.925	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874	7.790	7.718	7.657	7.605
7	12.246	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620	6.538	6.469	6.410	6.359
8	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814	5.734	5.667	5.609	5.559
9	10.562	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257	5.178	5.111	5.055	5.005
10	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849	4.772	4.706	4.650	4.601
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539	4.462	4.397	4.342	4.293
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296	4.220	4.155	4.100	4.052
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100	4.025	3.960	3.905	3.857
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939	3.864	3.800	3.745	3.698
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805	3.730	3.666	3.612	3.564
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691	3.616	3.553	3.498	3.451
17	8.400	6.112	5.185	4.669	4.336	4.101	3.927	3.791	3.682	3.593	3.518	3.455	3.401	3.353
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508	3.434	3.371	3.316	3.269
19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523	3.434	3.360	3.297	3.242	3.195
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368	3.294	3.231	3.177	3.130
30	7.562	5.390	4.510	4.018	3.699	3.473	3.305	3.173	3.067	2.979	2.906	2.843	2.789	2.742
40	7.314	5.178	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801	2.727	2.665	2.611	2.563
50	7.171	5.057	4.199	3.720	3.408	3.186	3.020	2.890	2.785	2.698	2.625	2.563	2.508	2.461
60	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718	2.632	2.559	2.496	2.442	2.394
70	7.011	4.922	4.074	3.600	3.291	3.071	2.906	2.777	2.672	2.585	2.512	2.450	2.395	2.348
80	6.963	4.881	4.036	3.563	3.255	3.036	2.871	2.742	2.637	2.551	2.478	2.415	2.361	2.313
90	6.925	4.849	4.007	3.535	3.228	3.009	2.845	2.715	2.611	2.524	2.451	2.389	2.334	2.286
100	6.895	4.824	3.984	3.513	3.206	2.988	2.823	2.694	2.590	2.503	2.430	2.368	2.313	2.265

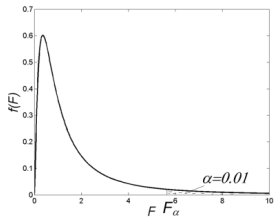


Table 4 continuation, $\boldsymbol{F}$ such that $\boldsymbol{p(F \geq F_{\alpha,\nu})} = \boldsymbol{\alpha}$

	ν_1													
ν_2	15	16	17	18	19	20	30	40	50	60	70	80	90	100
1	6157	6170	6181	6191	6201	6209	6260	6286	6302	6313	6321	6326	6331	6334
2	99.433	99.437	99.441	99.444	99.448	99.448	99.466	99.477	99.477	99.484	99.484	99.484	99.488	99.491
3	26.872	26.826	26.786	26.751	26.719	26.690	26.504	26.411	26.354	26.316	26.289	26.269	26.253	26.241
4	14.198	14.154	14.114	14.079	14.048	14.019	13.838	13.745	13.690	13.652	13.626	13.605	13.590	13.577
5	9.722	9.680	9.643	9.609	9.580	9.553	9.379	9.291	9.238	9.202	9.176	9.157	9.142	9.130
6	7.559	7.519	7.483	7.451	7.422	7.396	7.229	7.143	7.091	7.057	7.032	7.013	6.998	6.987
7	6.314	6.275	6.240	6.209	6.181	6.155	5.992	5.908	5.858	5.824	5.799	5.781	5.766	5.755
8	5.515	5.477	5.442	5.412	5.384	5.359	5.198	5.116	5.065	5.032	5.007	4.989	4.975	4.963
9	4.962	4.924	4.890	4.860	4.833	4.808	4.649	4.567	4.517	4.483	4.459	4.441	4.426	4.415
10	4.558	4.520	4.487	4.457	4.430	4.405	4.247	4.165	4.115	4.082	4.058	4.039	4.025	4.014
11	4.251	4.213	4.180	4.150	4.123	4.099	3.941	3.860	3.810	3.776	3.752	3.734	3.719	3.708
12	4.010	3.972	3.939	3.910	3.883	3.858	3.701	3.619	3.569	3.535	3.511	3.493	3.478	3.467
13	3.815	3.778	3.745	3.716	3.689	3.665	3.507	3.425	3.375	3.341	3.317	3.298	3.284	3.272
14	3.656	3.619	3.586	3.556	3.529	3.505	3.348	3.266	3.215	3.181	3.157	3.138	3.124	3.112
15	3.522	3.485	3.452	3.423	3.396	3.372	3.214	3.132	3.081	3.047	3.022	3.004	2.989	2.977
16	3.409	3.372	3.339	3.310	3.283	3.259	3.101	3.018	2.967	2.933	2.908	2.889	2.875	2.863
17	3.312	3.275	3.242	3.212	3.186	3.162	3.003	2.920	2.869	2.835	2.810	2.791	2.776	2.764
18	3.227	3.190	3.158	3.128	3.101	3.077	2.919	2.835	2.784	2.749	2.724	2.705	2.690	2.678
19	3.153	3.116	3.084	3.054	3.027	3.003	2.844	2.761	2.709	2.674	2.649	2.630	2.614	2.602
20	3.088	3.051	3.018	2.989	2.962	2.938	2.778	2.695	2.643	2.608	2.582	2.563	2.548	2.535
30	2.700	2.663	2.630	2.600	2.573	2.549	2.386	2.299	2.245	2.208	2.181	2.160	2.144	2.131
40	2.522	2.484	2.451	2.421	2.394	2.369	2.203	2.114	2.058	2.019	1.991	1.969	1.952	1.938
50	2.419	2.382	2.348	2.318	2.290	2.265	2.098	2.007	1.949	1.909	1.880	1.857	1.839	1.825
60	2.352	2.315	2.281	2.251	2.223	2.198	2.028	1.936	1.877	1.836	1.806	1.783	1.764	1.749
70	2.306	2.268	2.234	2.204	2.176	2.150	1.980	1.886	1.826	1.785	1.754	1.730	1.711	1.695
80	2.271	2.233	2.199	2.169	2.141	2.115	1.944	1.849	1.788	1.746	1.714	1.690	1.671	1.655
90	2.244	2.206	2.172	2.142	2.114	2.088	1.916	1.820	1.759	1.716	1.684	1.659	1.639	1.623
100	2.223	2.185	2.151	2.120	2.092	2.067	1.893	1.797	1.735	1.692	1.659	1.634	1.614	1.598

APPENDIX 5 F-SNEDECORE DISTRIBUTION. α=0.05

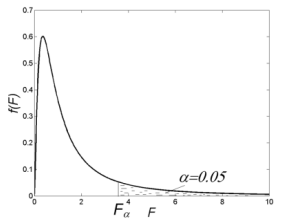


Table 5 F such that $p(F \geq F_{\alpha,\nu}) = \alpha$

	ν_1													
ν_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	161	199	216	225	230	234	237	239	241	242	243	244	245	245
2	18.513	19.000	19.164	19.247	19.296	19.329	19.353	19.371	19.385	19.396	19.405	19.412	19.419	19.424
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.785	8.763	8.745	8.729	8.715
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964	5.936	5.912	5.891	5.873
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735	4.704	4.678	4.655	4.636
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060	4.027	4.000	3.976	3.956
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637	3.603	3.575	3.550	3.529
8	5.318	4.459	4.066	3.838	3.688	3.581	3.500	3.438	3.388	3.347	3.313	3.284	3.259	3.237
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137	3.102	3.073	3.048	3.025
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978	2.943	2.913	2.887	2.865
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854	2.818	2.788	2.761	2.739
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753	2.717	2.687	2.660	2.637
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671	2.635	2.604	2.577	2.554
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602	2.565	2.534	2.507	2.484
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544	2.507	2.475	2.448	2.424
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494	2.456	2.425	2.397	2.373
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450	2.413	2.381	2.353	2.329
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412	2.374	2.342	2.314	2.290
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378	2.340	2.308	2.280	2.256
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348	2.310	2.278	2.250	2.225
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165	2.126	2.092	2.063	2.037
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077	2.038	2.003	1.974	1.948
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073	2.026	1.986	1.952	1.921	1.895
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993	1.952	1.917	1.887	1.860
70	3.978	3.128	2.736	2.503	2.346	2.231	2.143	2.074	2.017	1.969	1.928	1.893	1.863	1.836
80	3.960	3.111	2.719	2.486	2.329	2.214	2.126	2.056	1.999	1.951	1.910	1.875	1.845	1.817
90	3.947	3.098	2.706	2.473	2.316	2.201	2.113	2.043	1.986	1.938	1.897	1.861	1.830	1.803
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032	1.975	1.927	1.886	1.850	1.819	1.792

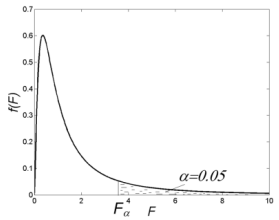


Table 5 continuation, $\boldsymbol{F}$ such that $\boldsymbol{p(F \geq F_{\alpha,\nu})} = \boldsymbol{\alpha}$

	ν_1													
ν_2	15	16	17	18	19	20	30	40	50	60	70	80	90	100
1	246	246	247	247	248	248	250	251	252	252	252	253	253	253
2	19.429	19.433	19.437	19.440	19.443	19.446	19.463	19.471	19.476	19.479	19.481	19.483	19.485	19.486
3	8.703	8.692	8.683	8.675	8.667	8.660	8.617	8.594	8.581	8.572	8.566	8.561	8.557	8.554
4	5.858	5.844	5.832	5.821	5.811	5.803	5.746	5.717	5.699	5.688	5.679	5.673	5.668	5.664
5	4.619	4.604	4.590	4.579	4.568	4.558	4.496	4.464	4.444	4.431	4.422	4.415	4.409	4.405
6	3.938	3.922	3.908	3.896	3.884	3.874	3.808	3.774	3.754	3.740	3.730	3.722	3.716	3.712
7	3.511	3.494	3.480	3.467	3.455	3.445	3.376	3.340	3.319	3.304	3.294	3.286	3.280	3.275
8	3.218	3.202	3.187	3.173	3.161	3.150	3.079	3.043	3.020	3.005	2.994	2.986	2.980	2.975
9	3.006	2.989	2.974	2.960	2.948	2.936	2.864	2.826	2.803	2.787	2.776	2.768	2.761	2.756
10	2.845	2.828	2.812	2.798	2.785	2.774	2.700	2.661	2.637	2.621	2.609	2.601	2.594	2.588
11	2.719	2.701	2.685	2.671	2.658	2.646	2.570	2.531	2.507	2.490	2.478	2.469	2.462	2.457
12	2.617	2.599	2.583	2.568	2.555	2.544	2.466	2.426	2.401	2.384	2.372	2.363	2.356	2.350
13	2.533	2.515	2.499	2.484	2.471	2.459	2.380	2.339	2.314	2.297	2.284	2.275	2.267	2.261
14	2.463	2.445	2.428	2.413	2.400	2.388	2.308	2.266	2.241	2.223	2.210	2.201	2.193	2.187
15	2.403	2.385	2.368	2.353	2.340	2.328	2.247	2.204	2.178	2.160	2.147	2.137	2.130	2.123
16	2.352	2.333	2.317	2.302	2.288	2.276	2.194	2.151	2.124	2.106	2.093	2.083	2.075	2.068
17	2.308	2.289	2.272	2.257	2.243	2.230	2.148	2.104	2.077	2.058	2.045	2.035	2.027	2.020
18	2.269	2.250	2.233	2.217	2.203	2.191	2.107	2.063	2.035	2.017	2.003	1.993	1.985	1.978
19	2.234	2.215	2.198	2.182	2.168	2.155	2.071	2.026	1.999	1.980	1.966	1.955	1.947	1.940
20	2.203	2.184	2.167	2.151	2.137	2.124	2.039	1.994	1.966	1.946	1.932	1.922	1.913	1.907
30	2.015	1.995	1.976	1.960	1.945	1.932	1.841	1.792	1.761	1.740	1.724	1.712	1.703	1.695
40	1.924	1.904	1.885	1.868	1.853	1.839	1.744	1.693	1.660	1.637	1.621	1.608	1.597	1.589
50	1.871	1.850	1.831	1.814	1.798	1.784	1.687	1.634	1.599	1.576	1.558	1.544	1.534	1.525
60	1.836	1.815	1.796	1.778	1.763	1.748	1.649	1.594	1.559	1.534	1.516	1.502	1.491	1.481
70	1.812	1.790	1.771	1.753	1.737	1.722	1.622	1.566	1.530	1.505	1.486	1.471	1.459	1.450
80	1.793	1.772	1.752	1.734	1.718	1.703	1.602	1.545	1.508	1.482	1.463	1.448	1.436	1.426
90	1.779	1.757	1.737	1.720	1.703	1.688	1.586	1.528	1.491	1.465	1.445	1.429	1.417	1.407
100	1.768	1.746	1.726	1.708	1.691	1.676	1.573	1.515	1.477	1.450	1.430	1.415	1.402	1.392

APPENDIX 6 K VALUES FOR CALCULATING TOLERANCE LIMITS

	confidence, q								
	90%			95%			99%		
	percentage, Q								
n	95%	99%	99.90%	95%	99%	99.90%	95%	99%	99.90%
2	18.800	24.167	30.227	37.674	48.430	60.573	188.491	242.300	303.054
3	6.919	8.974	11.309	9.916	12.861	16.208	22.401	29.055	36.616
4	4.943	6.440	8.149	6.370	8.299	10.502	11.150	14.527	18.383
5	4.152	5.423	6.879	5.079	6.634	8.415	7.855	10.260	13.015
6	3.723	4.870	6.188	4.414	5.775	7.337	6.345	8.301	10.548
7	3.452	4.521	5.750	4.007	5.248	6.676	5.488	7.187	9.142
8	3.264	4.278	5.446	3.732	4.891	6.226	4.936	6.468	8.234
9	3.125	4.098	5.220	3.532	4.631	5.899	4.550	5.966	7.600
10	3.018	3.959	5.046	3.379	4.433	5.649	4.265	5.594	7.129
15	2.713	3.562	4.545	2.954	3.878	4.949	3.507	4.605	5.876
20	2.564	3.368	4.300	2.752	3.615	4.614	3.168	4.161	5.312
25	2.474	3.251	4.151	2.631	3.457	4.413	2.972	3.904	4.985
30	2.413	3.170	4.049	2.549	3.350	4.278	2.841	3.733	4.768
35	2.368	3.112	3.974	2.490	3.272	4.179	2.748	3.611	4.611
40	2.334	3.066	3.917	2.445	3.213	4.104	2.677	3.518	4.493
45	2.306	3.030	3.871	2.408	3.165	4.042	2.621	3.444	4.399
50	2.284	3.001	3.833	2.379	3.126	3.993	2.576	3.385	4.323
55	2.265	2.976	3.801	2.354	3.094	3.951	2.538	3.335	4.260
60	2.333	2.248	2.955	3.774	3.066	3.916	2.506	3.293	4.206
65	2.235	2.937	3.751	2.315	3.042	3.886	2.478	3.257	4.160
70	2.222	2.920	3.730	2.299	3.021	3.859	2.454	3.225	4.120
75	2.211	2.906	3.712	2.285	3.002	3.853	2.433	3.197	4.084
80	2.202	2.894	3.696	2.272	2.986	3.814	2.414	3.173	4.053
85	2.193	2.882	3.682	2.261	2.971	3.795	2.397	3.150	4.024
90	2.185	2.872	3.669	2.251	2.958	3.778	2.382	3.130	3.999
95	2.178	2.863	3.657	2.241	2.945	3.763	2.368	3.112	3.976
100	2.172	2.854	3.646	2.233	2.934	3.748	2.355	3.096	3.954
110	2.160	2.839	3.626	2.218	2.915	3.723	2.333	3.066	3.917
120	2.150	2.826	3.610	2.205	2.898	3.702	2.314	3.041	3.885
130	2.141	2.814	3.595	2.194	2.883	3.683	2.298	3.019	3.857
140	2.134	2.804	3.582	2.184	2.870	3.666	2.283	3.000	3.833
150	2.127	2.795	3.571	2.175	2.859	3.652	2.270	2.983	3.811
160	2.121	2.787	3.561	2.167	2.848	3.638	2.259	2.968	3.792
170	2.116	2.780	3.552	2.160	2.839	3.527	2.248	2.955	3.774
180	2.111	2.774	3.543	2.154	2.831	3.616	2.239	2.942	3.759
190	2.106	2.768	3.536	2.148	2.823	3.606	2.230	2.931	3.744
200	2.102	2.762	3.529	2.143	2.816	3.597	2.222	2.921	3.731
250	2.085	2.740	3.501	2.121	2.788	3.561	2.191	2.880	3.678
300	2.073	2.725	3.481	2.106	2.767	3.535	2.169	2.850	3.641
400	2.057	2.703	3.453	2.084	2.739	3.499	2.138	2.809	3.589
500	2.046	2.689	3.434	2.070	2.721	3.475	2.117	2.783	3.555
600	2.038	2.678	3.421	2.060	2.707	3.458	2.102	2.763	3.530
700	2.032	2.670	3.411	2.052	2.697	3.445	2.091	2.748	3.511
800	2.027	2.663	3.402	2.046	2.688	3.434	2.082	2.736	3.495
900	2.023	2.658	3.396	2.040	2.682	3.426	2.075	2.726	3.483
1000	2.019	2.654	3.390	2.036	2.676	3.418	2.068	2.718	3.472
inf	1.960	2.576	3.291	1.960	2.576	3.291	1.960	2.576	3.291

APPENDIX 7 λ KOLMOGOROV DISTRIBUTION (LIMIT)

Table 7 $K(\lambda) = p(D\sqrt{n} \leq \lambda)$

λ	K(λ)	λ	K(λ)	λ	K(λ)	λ	K(λ)	λ	K(λ)
0.28	0.000001	0.75	0.372833	1.22	0.898104	1.69	0.993389	2.16	0.999822
0.29	0.000004	0.76	0.389640	1.23	0.902972	1.70	0.993828	2.17	0.999838
0.30	0.000009	0.77	0.406372	1.24	0.907648	1.71	0.994230	2.18	0.999852
0.31	0.000021	0.78	0.423002	1.25	0.912132	1.72	0.994612	2.19	0.999864
0.32	0.000046	0.79	0.439505	1.26	0.916432	1.73	0.994972	2.20	0.999874
0.33	0.000091	0.80	0.455857	1.27	0.920556	1.74	0.995309	2.21	0.999886
0.34	0.000171	0.81	0.472041	1.28	0.924505	1.75	0.995625	2.22	0.999896
0.35	0.000303	0.82	0.488030	1.29	0.928288	1.76	0.995922	2.23	0.999904
0.36	0.000511	0.83	0.503808	1.30	0.931908	1.77	0.996200	2.24	0.999912
0.37	0.000826	0.84	0.519366	1.31	0.935370	1.78	0.996460	2.25	0.999920
0.38	0.001285	0.85	0.534682	1.32	0.938682	1.79	0.996704	2.26	0.999926
0.39	0.001929	0.86	0.549744	1.33	0.941848	1.80	0.996912	2.27	0.999934
0.40	0.002808	0.87	0.564546	1.34	0.944872	1.81	0.997146	2.28	0.999940
0.41	0.003972	0.88	0.579070	1.35	0.947756	1.82	0.997346	2.29	0.999944
0.42	0.005476	0.89	0.593316	1.36	0.950512	1.83	0.997533	2.30	0.999949
0.43	0.007377	0.90	0.607270	1.37	0.953142	1.84	0.997707	2.31	0.999954
0.44	0.009730	0.91	0.620928	1.38	0.955650	1.85	0.997870	2.32	0.999958
0.45	0.012590	0.92	0.634286	1.39	0.958040	1.86	0.998023	2.33	0.999962
0.46	0.016005	0.93	0.647338	1.40	0.960318	1.87	0.998145	2.34	0.999965
0.47	0.020022	0.94	0.660082	1.41	0.962486	1.88	0.998297	2.35	0.999968
0.48	0.024682	0.95	0.672516	1.42	0.964552	1.89	0.998421	2.36	0.999970
0.49	0.030017	0.96	0.684636	1.43	0.966516	1.90	0.998536	2.37	0.999973
0.50	0.036055	0.97	0.696444	1.44	0.968382	1.91	0.998644	2.38	0.999976
0.51	0.042814	0.98	0.707940	1.45	0.970158	1.92	0.998744	2.39	0.999978
0.52	0.050306	0.99	0.719126	1.46	0.971846	1.93	0.998837	2.40	0.999980
0.53	0.058534	1.00	0.730000	1.47	0.973448	1.94	0.998924	2.41	0.999982
0.54	0.067497	1.01	0.740566	1.48	0.974970	1.95	0.999004	2.42	0.999984
0.55	0.077183	1.02	0.750826	1.49	0.976412	1.96	0.999079	2.43	0.999986
0.56	0.087577	1.03	0.760780	1.50	0.977782	1.97	0.999179	2.44	0.999987
0.57	0.098656	1.04	0.770434	1.51	0.979080	1.98	0.999213	2.45	0.999988
0.58	0.110395	1.05	0.779794	1.52	0.980310	1.99	0.999273	2.46	0.999989
0.59	0.122760	1.06	0.788860	1.53	0.981476	2.00	0.999329	2.47	0.999990
0.60	0.135718	1.07	0.797636	1.54	0.982578	2.01	0.999380	2.48	0.999991
0.61	0.149229	1.08	0.806128	1.55	0.983622	2.02	0.999428	2.49	0.999992
0.62	0.163225	1.09	0.814342	1.56	0.984610	2.03	0.999474	2.50	0.9999925
0.63	0.177753	1.10	0.822282	1.57	0.985544	2.04	0.999516	2.55	0.9999956
0.64	0.192677	1.11	0.829950	1.58	0.986426	2.05	0.999552	2.60	0.9999974
0.65	0.207987	1.12	0.837356	1.59	0.987260	2.06	0.999588	2.65	0.9999984
0.66	0.223637	1.13	0.844502	1.60	0.988048	2.07	0.999620	2.70	0.9999993
0.67	0.239582	1.14	0.851394	1.61	0.988791	2.08	0.999650	2.75	0.9999994
0.68	0.255780	1.15	0.858038	1.62	0.989492	2.09	0.999680	2.80	0.9999997
0.69	0.272189	1.16	0.864442	1.63	0.990154	2.10	0.999705	2.85	0.99999982
0.70	0.288765	1.17	0.870612	1.64	0.990777	2.11	0.999723	2.90	0.99999990
0.71	0.305471	1.18	0.876548	1.65	0.991364	2.12	0.999750	2.95	0.99999994
0.72	0.322265	1.19	0.882258	1.66	0.991917	2.13	0.999770	3.00	0.99999997
0.73	0.339113	1.20	0.887750	1.67	0.992438	2.14	0.999790		
0.74	0.355981	1.21	0.893303	1.68	0.992928	2.15	0.999806		

10 BIBLIOGRAPHY

1. M. R. Spiegel, J. J. Schiller, R. A. Srinivasan: Probability and Statistics, Schaum's Outlines, The McGraw-Hill Companies Inc., USA 2009.
2. L. J. Stephens: Engineering statistics demystified. The McGraw-Hill Companies Inc., USA 2007.
3. D.C. Montgomery, G.C. Runger: Applied Statisitcs and Probability for Engineers, John Wiley & Sons, Inc., USA 2007
4. D. T. Larose: Data Mining Methods and Models. John Wiley & Sons, Inc., USA 2006.
5. J. Koronacki, J. Mielniczuk: Statystyka dla studentów kierunków technicznych i przyrodniczych. WNT, Warszawa 2006.
6. A. Plucińska, E. Pluciński: Rachunek prawdopodobieństwa Statystyka matematyczna Procesy stochastyczne. WNT, Warszawa 2000.
7. J. Jóźwiak, J. Podgórski: Statystyka od podstaw. PWE, Warszawa 2000.
8. Engineering statistics. NIST. http://www.itl.nist.gov/div898/handbook/