Slezská univerzita v Opavě
Obchodně podnikatelská fakulta v Karviné

STATISTICAL METHODS FOR ECONOMISTS

Filip Tošenovský

Karviná 2014

Projekt OP VK č. CZ.1.07/2.2.00/28.0017
„Inovace studijních programů na Slezské univerzitě,
Obchodně podnikatelské fakultě v Karviné“

Field: Statistics

Annotation: This textbook presents to the reader important parts of statistical data analysis. The subject matter contained in the book focuses on statistical methods which constitute a standard part of scholarly materials used both at domestic and foreign universities. The methods include description of statistical characteristics, hypothesis testing, regression and correlation analysis, analysis of variance, and also other procedures abundantly used in industries for product quality control, such as design of experiments, Taguchi's methods based on loss functions and control charts.

Key words: Statistical characteristics, regression, correlation, hypothesis testing, analysis of variance, design of experiments, Taguchi's methods.

Author: **Ing. Filip Tošenovský, Ph.D.**

Reviewers: Prof. RNDr. Josef Tošenovský, CSc., Ing. Elena Mielcová, Ph.D.

ISBN **978-80-7510-033-7**

CONTENTS

INTRODUCTION

The presented textbook serves as a study material for the course Statistical Methods for Economists taught at the Karviná-based School of Business Administration of the Silesian University. The course Statistical Methods for Economists, a follow-up to the course Statistics, stresses the importance of application of statistical methods in economic disciplines, such as marketing, management, production planning and quality management.

The textbook is divided into twelve chapters, which corresponds to the usual twelve weeks of teaching the school term consists of. The chapters are more or less the same in terms of the extent of their contents and difficulty. The extent of each chapter corresponds to a two-hour lecture presented to full-time students at schools of economics. As part of a full-time study, the course lecture is accompanied by a seminar in which the explained subject matter is practised, using specific numerical examples and often a computer software, as well.

However, the Silesian University part-time students may also use the textbook. Part-time studying is a form of study which, in the case of Statistical Methods for Economists, requires students to work regularly and persistently, be able to concentrate on the subject and take an active approach to solving problems on their own. This is where the textbook should help substitute the fine full-time teaching, and serve as a study material. Other literary resources listed at the end of the textbook may also be of additional help in this respect.

To pass the course Statistical Methods for Economists successfully, it is assumed that students have passed the course Statistics in the first place. It is true that not all learnt in Statistics is necessary to master Statistical Methods for Economists because some of the subject matter presented earlier had a different purpose of being demonstrated. None the less, the ability of accurate and logical thinking experienced in the previous course will come in useful, and so will the ability to recognize mathematical symbols used and the knowledge of essentials of the probability theory and statistics.

Returning to the course Statistical Methods for Economists, let us describe its contents in a greater detail. A more accurate name for the subject could be Selected Statistical Methods for Economists, or even more accurately: Selected Statistical Methods of Marketing, Management and Quality Control. These are the three major areas of interest the university students often encounter in real life when applying statistics. Chapter 1 of the textbook revises elementary terms used in statistics, chapters 2-7 deal with the application of statistics in marketing and management, and chapters 8-12 are devoted to statistics in production planning and quality control. The subject matter and the related problems are studied using Excel, as long as the given problem allows Excel to find the solution. Students have already become familiarized with Excel in the course Statistics.

As was mentioned at the beginning of this introduction, the text is divided into twelve chapters. Each chapter requires about four to six hours of study. However, the reward waiting at the end of the study is worth it: it is the feeling that something significant has been overcome – an obstacle that separates the world of professionals from the world of non-professionals. With such knowledge, one can better analyse information we are all flooded with at present.

1 ESSENTIAL STATISTICAL TERMS, CHARACTERISTICS

All statistical methods work with certain terms. This way, both authors of the statistical theory and its users can communicate among themselves results of their analyses in a comprehensible way. To simplify the communication, specific terms are introduced in statistics. The advantage of this procedure lies in the fact that the terms are introduced only once, but their validity is permanent for all interested parties. Also, using the simplest terms, one may construct more sophisticated terms or methods. We shall revise as well as extend some of the terms introduced in the course Statistics, and define essential statistical characteristics, using the terms. The characteristics will in a convenient way summarize information contained in the data under scrutiny. **We note that from now on, whenever we refer to a closed interval, we shall denote the interval with parentheses „[ ]".**

The main objective of statistics is to analyse a certain data. Of course, there is a reason why data originated and is maintained. Its purpose is to help analyse the form and behaviour of a *statistical variable* the data is related to. An example of such a variable is the height of women in the Czech Republic, political preference of a citizen, gross domestic product of a country, an average of a ball bearing produced, etc. We shall be mainly interested in numerical variables that better suit the needs of mathematics. If this is our case, the analysed variable can take on different *values* (otherwise it wouldn't be a variable, of course). The set of all values a variable can take on is called *population*. Since population is related to a specific variable, it is a relative term. For instance, if we are interested in political preferences of the Czechs, the population consists of all Czech citizens, and will not be usually available unless census takes place and its results are available to the public. On the other hand, if we are interested in school results of a specific group of students who attend the course Statistical Methods for Economists, the group will represent the population which will be easily within reach. Statisticians, however, more often than not do not have populations at hand, and in such cases all they can do is perform a sampling from the population, which results in having a *data sample* of the population. There is more than one way how to obtain a data sample. There are also branches of statistics whose sole purpose is to analyse various forms of data sampling. In statistics, we usually require that the sampling be random. Random sampling means that every element of the population has the same probability of being selected, the same probability of being present in the final data sample. More precisely, random data sample of size n is a random vector $(X_1, X_2, \dots, X_n)$ where the random variables X_i's follow the same probability distribution (population), and are statistically independent. Such a sampling is required because it posseses certain „representative" properties the theory of statistics relies on.

If the data sample is available, we can analyse it with proper statistical methods, and based on this analysis we may formulate conclusions about the data structure of the population the data sample came from. Such conclusions constitute what is called *statistical inference*.

If the population can be obtained, the only ambition of statistics might be to describe the population. Methods that serve this purpose form *descriptive statistics*. Descriptive statistics provides us with characteristics that describe the population with a single number. These characteristics are called *population characteristics*. A characteristic summarizes information about the data. If the population consists of two thousand values, it is certainly better to use a single number – a characteristic to get a rough idea about the population, rather than name all its values. This aggregation, however, is not flawless: there must necessarily be a loss of the

original information about the population since a single number cannot obviously reflect the entire amount of the original information.

In case that only the data sample is available, not the entire population, the so-called *sample characteristics* are used to describe the data sample structure. It is a common habit to denote the population characteristics with Greek letters and the sample characteristics with Latin letters. In this way, an order is introduced to the notation, and all users of the theory know immediately whether they work with population characteristics or their sample counterparts.

We shall now introduce other terms and new characteristics for data that represent values of a single statistical variable. Later, data that contains information about two statistical variables will also be handled.

1.1 THE CASE OF A SINGLE VARIABLE

Let us have a population consisting of values $x_1, x_2, \dots, x_n$, where n is an integer, i.e. a finite number (we shall work with data of finite size only). Let X be a variable of interest. The numbers $x_1, x_2, \dots, x_n$ are the values which the variable can take on. If we apply a random sampling to this population, we can regard the variable X as a (discrete) random variable. Although the population contains the values $x_1, x_2, \dots, x_n$, not all these values must necessarily be different. Some of them may repeat. In such cases, X takes on only k different values $x_1^*, x_2^*, \dots, x_k^*$. The value x_1^* may appear in the population f_1 times. The number f_1 is called *absolute frequency* of the value x_1^*. Similarly, the value x_2^* appears in the population f_2 times, the value x_3^* ... f_3 times and so on…until the last value x_k^* appears in the population f_k times. Apart from absolute frequencies, we also work with other types of frequencies:

a) *relative frequency* of appearance of value x_l^* is given by the division f_l/n, where $n = \sum_{i=1}^{k} f_i = f_1 + f_2 + \cdots + f_k$ denotes the population size.

If we sort the values $x_1^*, x_2^*, \dots, x_k^*$ in the ascending order, we get a population where $x_{(1)}^* \leq x_{(2)}^* \leq \cdots \leq x_{(k)}^*$ holds. In this notation, $x_{(1)}^*$ is the minimum of the set $x_1^*, x_2^*, \dots, x_k^*$, $x_{(2)}^*$ is the second smallest number in the set, and so on. If the value $x_{(i)}^*$ has an absolute frequency of its occurence f_i^*, we may introduce the following new terms:

b) *absolute cumulative frequency of* $x_{(l)}^*$, which is given by the sum $\sum_{i=1}^{l} f_i^*$.

c) *Relative cumulative frequency of* $x_{(l)}^*$, given by the sum $\sum_{i=1}^{l} f_i^* / n$.

The foregoing types of frequencies can be used both for population and data sample.

1.1.1 MEASURES OF CENTRAL TENDENCY

Let us have a population consisting of values $x_1, x_2, \dots, x_n$. *Population arithmetic mean* μ is one of the most important measures of central tendency. It is defined as

1-1
$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Since the mean is related to a population, it is called population mean. If we performed a data sampling of size *m* from the population and obtained values $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_m$ by the sampling, we could estimate the usually unknown population mean by *sample mean*

1-2 $$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} \tilde{x}_i.$$

Excel: In Excel, both characteristics can be calculated using the function *mean*() which requires only one parameter: a reference to the area in the Excel spreadsheet that contains the data.

If we know the population $x_1, x_2, \dots, x_n$ contains only *k* different values: a value x_1^* exactly f_1 times, a value x_2^* … f_2 times, etc. ... and finally a value x_k^* … f_k times, we may rewrite 1-1 to

1-3 $$\mu = \frac{1}{n}\sum_{i=1}^{k} x_i^*.f_i = \frac{1}{\sum_{j=1}^{k} f_j}\sum_{i=1}^{k} x_i^*.f_i.$$

Similarly, equation 1-2 can be rewritten, using the absolute frequencies with which the values $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_m$ appear in the data sample. Also, we may regard 1-3 as a special case of what is called *weighted average*. Weighted average of values $x_1^*, x_2^*, \dots, x_k^*$ with weights $w_1, w_2, \dots, w_k$ is defined as

1-4 $$\mu_w = \frac{1}{\sum_{j=1}^{k} w_j}\sum_{i=1}^{k} x_i^*.w_i.$$

If the sum of weights is equal to one, it is clear that formulas 1-3 and 1-4 represent the same thing for $w_i = f_i / \sum_{j=1}^{k} f_j$, *i* = 1, 2,..., *k*.

Another measure of central tendency is *mode* $\hat{x}$, which is the value with the highest absolute frequency. This definition doesn't guarantee uniqueness of the mode, however. Thus, it may happen that the data contains more than one mode.

Yet another measure of central tendency is *median*, denoted $\tilde{x}$ or x_{50}. We also talk about a middle value. Median is generally not the same as average or mean. For a data sample consisting of values $x_1, x_2, \dots, x_n$, we may calculate the median in the following steps:

1) we sort the data in the ascending order to get a data $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$,
2) we calculate $z = n \cdot 0{,}5 + 0{,}5$,
3) if *z* is an integer (this is true when *n* is odd), then $\tilde{x} = x_{(z)}$. If *z* is not an integer (this happens when *n* is even), then $\tilde{x} = (x_{(z-0,5)} + x_{(z+0,5)})/2$.

Excel: Excel uses the function *median*() to determine median. The function requires only one parameter – a reference to the region in the Excel spreadsheet containing the data. Uniqueness of median is guaranteed by the definition in this case. We stress that to use the function correctly, each value of the data sample must be written out explicitly, i.e. it must not be a data region containing only different values of the sample together with their respective frequencies (see problem 1 how to proceed in this situation).

PROBLEM 1

Let a data sample contains number 7 with absolute frequency 234, number 9 with absolute frequency 672 and number 43 with absolute frequency 347. Calculate the relative frequency of each value, arithmetic mean of the sample, mode and median.

SOLUTION

Since mode is the value with the highest frequency, number 9 is the mode in this case. The relative frequency of 7 equals 234/(234+672+347), the relative frequency of 9 is 672/(234+672+347) and the relative frequency of 43 equals 347/(234+672+347). The arithmetic mean/average is

$$\frac{7 \cdot 234 + 9 \cdot 672 + 43 \cdot 347}{234 + 672 + 347} = 18{,}04.$$

The sample size is 1253 – an odd number. Therefore, the median is the 627th value in the sorted data sample, which is number 9.

1.1.2 MEASURES OF VARIABILITY

Measures of central tendency summarize in a certain sense information about where on the real line the values of the observed variable *X* typically are. However, the nature of these measures is such that they don't say anything about how far the values are from one another. For these purposes, measures of variability were introduced. They describe „a typical" mutual deviation of the individual values of *X*. In case of a population $x_1, x_2, \dots, x_n$, *population variance* σ^2, as one of the measures of variability, is defined as

1-5 $$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2.$$

If we only have a data sample $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_m$ of size *m* drawn from the population, we estimate the usually unknown population variance by *sample variance* s^2

1-6 $$s^2 = \frac{1}{(m-1)}\sum_{i=1}^{m}(\tilde{x}_i - \bar{x})^2,$$

where $\bar{x}$ is the arithmetic mean of the values $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_m$. We note that 1-6 is the typical formula for sample variance, but not the only one. Equation 1-5 tells us that the variance is a mean squared deviation of individual values of *X* from the population average μ.

Excel: To calculate the population variance, we use the Excel function *varp*() which demands only one parameter – the data region in the spreadsheet, for which the variance is to be determined. To calculate the sample variance, the function *var*() is used with the same argument.

Just like in the case of the arithmetic mean, we may rewrite equation 1-5 or 1-6 to its equivalent form that works with absolute frequencies: If we know the population $x_1, x_2, \dots, x_n$ contains only *k* different values – a value $x_1^* \dots f_1$ times, a value $x_2^* \dots f_2$ times,..., a value $x_k^* \dots f_k$ times, we may use the following formula instead of 1-5

1-7 $$\sigma^2 = \frac{1}{\sum_{j=1}^{k} f_j} \sum_{i=1}^{k} (x_i^* - \mu)^2 \cdot f_i.$$

The same logic/analogy applies to equation 1-6 if we use the absolute frequencies with which the data sample contains different values $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_m$.

Another measure of variability is *standard deviation*, defined as the square root of variance. If the underlying measure is a population variance, we talk about *population standard deviation* σ. If the underlying measure is a sample variance, we arrive at *sample standard deviation s* by taking the square root.

Range R, defined as $R = x_{max} - x_{min}$, where x_{max} is the highest value in the data and x_{min} is the lowest value in the data, also belongs among measures of variability.

Excel: The highest number in the data may be found using the Excel function *max*(), while the lowest number can be obtained with the function *min*(). Both functions require as the parameter a reference to the section of the spreadsheet containing the data.

We finalize the section on measures of variability with *coefficient of variation V*. If a population is available, we define its *population coefficient of variation* by equation

1-8 $$V = \frac{\sigma}{|\mu|}.$$

If only a sample is within reach, the corresponding *sample coefficient of variation* is calculated, using the sample standard deviation and sample mean:

1-9 $$V = \frac{s}{|\bar{x}|}.$$

The coefficient is suitable for situations when two data groups are to be compared in terms of their variability, but each of the groups relates to a variable with different physical units. Under these circuimstances, it is meaningless to use variance as a measure of variability because it would be calculated in different and squared physical units, making itself useless for any comparison.

The higher the coefficient of variation, the higher the variability of the given data group.

PROBLEM 2

Let a population contains number 7 with absolute frequency 234, number 9 with frequency 672 and number 43 with frequency 347. Calculate the population variance.

SOLUTION

We shall use formula 1-7. In the previous problem, which presented the same data group as a sample, we calculated the sample average 18,04. The same average appears here, but this time, it is a population average. According to 1-7, we have:

$$\sigma^2 = \frac{(7 - 18{,}04)^2 \cdot 234 + (9 - 18{,}04)^2 \cdot 672 + (43 - 18{,}04)^2 \cdot 347}{234 + 672 + 347} = 239{,}12.$$

1.1.3 MEASURES OF DATA CONCENTRATION

The last category of characteristics we are about to describe consists of measures that reflect in a sense to what extent or how the data under scrutiny are grouped together. Two major representatives of this category are kurtosis *Ku* and skewness *Sk*. If a population $x_1, x_2, \dots, x_n$ is available, *population kurtosis* is defined as

1-9 $$Ku = \frac{\sum_{i=1}^{n}(x_i-\mu)^4}{n\sigma^4}.$$

Formula 1-9 can also be written equivalently as

1-10 $$Ku = \frac{\sum_{i=1}^{k}(x_i^* - \mu)^4 \cdot f_i}{\sigma^4 \sum_{j=1}^{k} f_j}$$

provided the population $x_1, x_2, \dots, x_n$ contains only *k* different values: a value x_1^* … f_1 times, a value x_2^* … f_2 times, ... , a value x_k^* … f_k times. If $x_1, x_2, \dots, x_n$ represent only a data sample, we calculate *sample kurtosis*, using again equation 1-9 or 1-10, but replacing the population mean μ in the corresponding equation with the sample mean $\bar{x}$, and also replacing the fourth power of the population standard deviation σ^4 with the fourth power of the sample standard deviation s^4.

It is clear from 1-9 and 1-10 that kurtosis is nonnegative. Also, the intepretation of the characteristic is such that the higher the kurtosis, the higher the concentration of the data that lie closer to the mean, as compared to the data that are farther from the mean.

Formulas 1-9 and 1-10 sometimes appear in an altered form, with number 3 being substracted from 1-9 and 1-10. This modification compares the kurtosis of the analysed data with the kurtosis of normal distribution. The kurtosis of normal distribution is known to be 3, regardless of the parameters of the distribution. This means that if the modified kurtosis of the data is positive, the frequency distribution of the analysed data has a higher kurtosis than normal distribution. We note that such a modification of the kurtosis is not the only one.

Excel: Excel offers the function *kurt*() for the enumeration of kurtosis. The function has only one parameter, which is a reference to the area of the Excel spreadsheet containing the analysed data. However, the function calculates yet another modification of kurtosis, which **is not identical** to the most often used definitions 1-9 or 1-10. None the less, the Excel modification can still be used for comparison of two data groups in terms of their kurtosis, and its interpretation remains the same.

Population skewness is defined as

1-11 $$Sk = \frac{\sum_{i=1}^{n}(x_i-\mu)^3}{n\sigma^3},$$

or equivalently as

1-12 $$Sk = \frac{\sum_{i=1}^{k}(x_i^*-\mu)^3 \cdot f_i}{\sigma^3 \sum_{j=1}^{k} f_j},$$

if the population $x_1, x_2, \dots, x_n$ contains only *k* different values: a value $x_1^* \dots f_1$ times, a value $x_2^* \dots f_2$ times, ... , a value $x_k^* \dots f_k$ times. If we were to calculate *sample skewness*, the note made in the case of kurtosis applies here, as well: we use again formulas 1-11 or 1-12, with the population mean replaced by the sample mean, and the third power of the population standard deviation replaced by the third power of its sample counterpart s^3.

As is clear from the defining formulas, skewness can take on any real value. If skewness is zero, it means the frequency distribution of the data is symmetric. More intuitively, concentration of smaller values is the same as that of higher values. If skewness is positive, we say the frequency distribution of the data is skewed to the right, and concentration of smaller values is stronger than concentration of higher values. Finally, if skewness is negative, we say the data distribution is skewed to the left, and concentration of higher values is stronger than concentration of smaller values. In the case of a nonzero skewness, the frequency distribution of the data is said to be asymmetric.

Frequency distribution is often portrayed by a two-dimensional graph. The horizontal axis of the graph describes different values $x_1^*, x_2^*, \dots, x_k^*$ appearing in the data group, whereas the vertical axis of the graph measures the frequencies with which $x_1^*, x_2^*, \dots, x_k^*$ are contained in the data group.

PROBLEM 3

A population contains the following values: 111 with absolute frequency 500, 222 with absolute frequency 400, 333 with absolute frequency 600 and 444 with absolute frequency 300. Calculate the population skewness.

SOLUTION

The population size is 1800. The population average is 265,166, the population variance equals 13880,14. Thus, the third power of the population standard deviation is 1635275. According to 1-12, we get

$$Sk = \frac{(111 - 265{,}16)^3 \cdot 500 + \dots + (444 - 265{,}16)^3 \cdot 300}{1635275 \cdot 1800} = 0{,}01.$$

We may conclude the frequency distribution is almost perfectly symmetric in this case.

1.1.4 GENERAL MOMENTS

General moments are characteristics that look at data structures from a different angle. There are several reasons why we work with general moments. One reason is that frequency distributions and moments are related to each other uniquely under certain conditions: Data groups with the same moments have the same frequency distributions and vice versa. What we are interested in, however, relates to another reason why we work with the moments: it is the fact that some of the foregoing characteristics can be calculated in a more elegant way, using these moments.

For a population $x_1, x_2, \dots, x_n$, we define the *k-th general moment* M_k by equation

1-13 $$M_k = n^{-1} \cdot \sum_{i=1}^{n} x_i^k, \quad k = 1,2, \dots$$

Thus, the *k*-th moment is nothing but the average of the *k*-th power of the original data. If the data group contains only *m* different values x_i^*, i = 1, 2, …, m with frequencies f_i, we may also enumerate 1-13 with the formula $M_k = n^{-1} \cdot \sum_{i=1}^{m} (x_i^*)^k \cdot f_i \,, k = 1,2, \ldots$.

Now, the following relations hold true:

1-14
$$M_1 = \mu,$$
$$M_2 - M_1^2 = \sigma^2,$$
$$\sigma^{-3} \cdot (M_3 - 3M_1M_2 + 2M_1^3) = Sk,$$
$$\sigma^{-4} \cdot (M_4 - 4M_3M_1 + 6M_2M_1^3 - 3M_1^4) = Ku.$$

PROBLEM 4

A group of data *D* contains value 11 with absolute frequency 4235 and value 254 with absolute frequency 6543. Calculate the first two general moments.

SOLUTION

According to 1-13, we have

$$M_1 = (4235 + 6543)^{-1} \cdot (11 \cdot 4235 + 254 \cdot 6543) = 158{,}518,$$
$$M_2 = (4235 + 6543)^{-1} \cdot (11^2 \cdot 4235 + 254^2 \cdot 6543) = 39213{,}27.$$

1.2 THE CASE OF TWO VARIABLES

If we have a data group such that for each integer *i* = 1, 2, …, *m* and *j* = 1, 2, …, *n* it contains a pair of values (x_i, y_j) or even more pairs with these two values, we are working with a *data group of two statistical variables*. Absolute frequency of the pair (x_i, y_j) is called *joint frequency* of (x_i, y_j), and is denoted f_{ij}. The size of the data group is $r = \sum_{i,j} f_{ij}$.

The distribution of joint frequencies in the data group is described by a two-dimensional table called *contingency table* (see table 1). The heading of the table contains different categories of each variable, and the table itself contains joint frequencies with which the combinations of the categories occur in the data group.

A variable *y*, for instance, may represent family status, while a second variable, say *x*, can describe attained educational level. The joint frequency f_{11}, for example, will then describe the number of individuals who attained the level of education x_1, and at the same time have the family status y_1. A similar statement holds true for the other joint frequencies. The last column of the table is usually reserved for the sum of the joint frequencies that lie in the same row. The sum is denoted $f_{i.}$ if we work with the *i*-th row of the table. The last row of the table is reserved for the sum of the joint frequencies that lie in the same column. This sum is denoted $f_{.j}$ if we talk about the *j*-th column of the table. These summations are called *marginal frequencies*.

Table 1: Contingency table

y / x	y_1	y_2	...	y_n	
x_1	f_{11}	f_{12}	...	f_{1n}	$f_{1\cdot}$
x_2	f_{21}	f_{22}	...	f_{2n}	$f_{2\cdot}$
...	...	...	...	...	...
x_m	f_{m1}	f_{m2}	...	f_{mn}	$f_{m\cdot}$
	$f_{\cdot 1}$	$f_{\cdot 2}$		$f_{\cdot n}$	r

Source: author's

If we assume that the table represents an entire population, we can calculate basic characteristics for the two variables, using the symbols introduced for different types of frequencies, i.e. we can calculate the population means and population variances, using the following formulas:

1. Population means

$$\mu_X = \frac{1}{r}\sum_i x_i \sum_j f_{ij},$$

$$\mu_Y = \frac{1}{r}\sum_j y_j \sum_i f_{ij}.$$

2. Population variances

$$\sigma_X^2 = \frac{1}{r}\sum_i (x_i - \mu_X)^2 \sum_j f_{ij}$$

$$\sigma_Y^2 = \frac{1}{r}\sum_j (y_j - \mu_Y)^2 \sum_i f_{ij}\ .$$

On the other hand, if the table represented only a data sample, we would calculate the sample means and variances according to the formulas:

1. Sample means

$$\bar{x} = \frac{1}{r}\sum_i x_i \sum_j f_{ij},$$

$$\bar{y} = \frac{1}{r}\sum_j y_j \sum_i f_{ij}.$$

2. Sample variances

$$s_X^2 = \frac{1}{r-1}\sum_i (x_i - \bar{x})^2 \sum_j f_{ij},$$

$$s_Y^2 = \frac{1}{r-1}\sum_j (y_j - \overline{y})^2 \sum_i f_{ij}.$$

When working with two variables, the frequencies of which are given by the aforementioned contingency table, we also define another important characteristic called *covariance*. *Population covariance*, denoted $\mathrm{cov}(X,Y)$, of variables *X* and *Y* is defined by equation

1-15
$$\mathrm{cov}(X,Y) = \frac{1}{r}\sum_i \sum_j (x_i - \mu_X)(y_j - \mu_Y) f_{ij} = \frac{1}{r}\sum_i \sum_j x_i y_j f_{ij} - \mu_X \cdot \mu_Y.$$

If we work with a data sample of size *n*, $n \geq 2$, we also define *sample covariance*:

1-16
$$c_{XY} = \frac{1}{r-1}\sum_i \sum_j (x_i - \overline{x})(y_j - \overline{y}) f_{ij}.$$

PROBLEM 5

A variable *X* can take on values 3, 5, 4, 6, 7 and 9. For these values, values of another variable *Y* were measured: 1, 2, 7, 9, 11 and 13, respectively, i.e. 3 corresponds to 1, 5 corresponds to 2, etc. The absolute frequency of each value is one. Calculate the population covariance.

SOLUTION

We use 1-15, setting all the frequencies equal to one. The average of *X* is 5,66, the average of *Y* is 7,16. We get

$$\mathrm{cov}(X,Y) = \frac{1}{r}\sum_i \sum_j (x_i - \mu_X)(y_j - \mu_Y) f_{ij} = \frac{(3-5{,}66)\cdot(1-7{,}16) + \ldots + (9-5{,}66)\cdot(13-7{,}16)}{6} = 7{,}55.$$

Covariance is used to describe a mutual dependence of variables *X* and *Y*, the dependence taking the form of a line, i.e. we deal with a simple linear dependence. If the covariance is positive, we can say that a dependence in the form of a line exists between the two variables to a certain extent. The dependence is such that if one of the variables rises in value, the other rises to a certain extent, as well. On the contrary, if the covariance is negative, it signals existence of a comotion of the two variables but in opposite directions: if one of the variables rises in value, the other drops to an extent. In both cases, the movement of the other variable is to an extent proportionate to the change of the first variable. Zero covariance suggests there is no linear dependence between the variables. As we can see, it is the sign of the covariance that matters. The value of covariance is further transformed so that the transformed characteristic falls to closed interval [-1,1]. The transformation takes place to get a characteristic called *paired correlation coefficient*, which gives us a better interpretation as to how strong the linear dependence exists between the two variables. If we work with a population, we obtain *population paired correlation coefficient* through this transformation. If we work with a sample, the result is called *sample paired correlation coefficient*.

The population paired correlation coefficient is of the form

1-17
$$\rho = \frac{\mathrm{cov}(X,Y)}{\sigma_X \cdot \sigma_Y},$$

where σ_X is the population standard deviation of X and σ_Y is the population standard deviation of Y. The sample paired correlation coefficient is given by

1-18
$$r = \frac{c_{XY}}{s_X \cdot s_Y},$$

where s_X is the sample standard deviation of X and s_Y is the sample standard deviation of Y. Both the population and sample correlation coefficient can take on only values from interval [-1,1]. If the population paired correlation equals one, it means that there is an exact relation in the form of a rising line. If, on the contrary, the correlation equals minus one, there is an exact relation between the two variables in the form of a declining line. If the population correlation equals zero, the two variables are said to be uncorrelated. There are also other types of correlations, apart from the paired correlation. We shall discuss them in chapter 5.

CONTROL TEST 1

The following questions concern a data group which contains different values of a variable X together with their corresponding absolute frequencies.

Values of *X*	**Absolute frequencies**
23	2345
34	6213
33	456
35	8876
37	12134
31	5436
16	445

Source: author's

a. Calculate the arithmetic average, median and mode of X.
b. If the table represents a population, what do the variance, standard deviation, coefficient of variation and range look like? What would these characteristics look like if the table represented a sample?
c. Calculate the first two general moments of X.
d. The following table contains a data sample on two variables.

X	3	4	5	1	6	7	8
Y	5	3	4	6	7	8	9

Source: author's

Determine the sample covariance.

e. Estimate the paired correlation between X and Y, using the table above.

Complete the statements:

f. Median and average are measures of…….. .
g. Standard deviation and range are measures of …… .
h. Skewness and kurtosis are measures of…….. .
i. Coefficient of paired correlation takes on values from the interval…….. .

SOLUTIONS

a. Using 1-3, the average is 33,85. There are 35905 values available, which is an odd number. Therefore, the median equals the 17953th value in the data group sorted in the ascending order. The following table contains the sorted data:

Values	Frequencies
16	445
23	2345
31	5436
33	456
34	6213
35	8876
37	12134

Source: author's

The numbers 16 to 34 form a data subgroup of size 14895. The numbers 16 to 35 make up a data subgroup of size 23771. Thus, the median is obviously equal to 35. Mode is 37 in the problem.

b. The population variance is 16,56 according to 1-7. Its square root equals 4,07, which is the population standard deviation. The range is 37-16 = 21. The population coefficient of variation equals 4,07/33,85 = 0,12. The sample variance is (35905/35904).16,56 = 16,56. Given the relatively huge size of the data group, it is the same number as the population variance if the result is rounded to two decimal places. The calculation results from the relation between the two variances. The sample standard variation is 4,07, as well. The remaining sample characteristics will therefore be more or less the same (if we use the rounded numbers, otherwise, precisely speaking, they are not exactly the same).
c. The first general moment = average = 33,85. The second general moment = 1162,56.
d. The sample covariance = 3,166.
e. The sample correlation = 0,608.
f. Central tendency.
g. Variability.
h. Data concentration.
i. [-1, 1].

2 HYPOTHESIS TESTING IN MARKETING

The second chapter covers hypothesis testing, which is one of the most important techniques in statistics. The first part of the chapter revises some fundamental principles of hypothesis testing. A part of it was already covered in the course Statistics. Another section of the chapter describes statistical tests which could be considered to be elementary because this is how they are treated in many other scholarly texts. We shall also describe some other tests which suit particularly the needs of marketing. The technique of hypothesis testing is explained in a greater detail in accompanying examples.

2.1 TESTING STATISTICAL HYPOTHESES

Statistical hypotheses constitute only a part of all scientific hypotheses. They are related to random variables, and we divide the set of such hypotheses to a subset of parametric hypotheses and a subset of nonparametric hypotheses. Parametric hypotheses deal with parameters of the probability distribution of a random variable (or the population of an observed statistical variable). Nonparametric hypotheses are not related to parameters of such a distribution, they are related to some other properties of the distribution, such as its shape, for instance, because we may be interested in whether the behaviour of a random variable can be described properly by a binomial distribution or a normal distribution.

Every statistical test works with two hypotheses that stand against each other: a tested hypothesis (tested statement), called *null hypothesis* and denoted H_0, and an *alternative hypothesis*, denoted H_1. H_1 is usually the negation of H_0. What we usually have available for hypothesis testing is the result of a data sampling. Such a sampling can take the form of a marketing study or poll. Without sampling, which originates in a random way, it is not possible to perform a statistical test. Based on the result of the sampling, we are now to decide whether to accept or reject the null hypothesis. To make such a conclusion, we calculate what is called *test criterion T*, a function of the data gathered by the sampling. We also define a subset within the set of real numbers, the subset being called *critical region*. Different tests have different critical regions. If the test criterion T falls to the critical region, the null hypothesis is rejected. In the opposite case, the null hypothesis is accepted. The critical region is usually defined by a *critical value K* (or it can also be defined by a percentile). The critical value is either found in statistical tables or calculated with a suitable software (Excel, for instance).

Let us note that by accepting the null hypothesis, we are not proving the validity of the tested statement. Testing hypotheses does not necessarily lead to the right conclusion, which is natural since it is a process based on the limited amount of information stored in the data sample we work with. Uncertainty in the conclusion from a statistical test is related to *nivel of test* α, a parameter defined by whoever performs the statistical test (we will talk about this parameter in a moment). Let us point out again that the testing is based on the randomness of data sampling. In other words, it is based on the fact that the data to be used for the test were gathered independently of one another (independently in the statistical sense of the word). Whether this is true or not can also be tested [10].

For convenience, let us summarize the steps that lead to acceptance of the null hypothesis or its rejection.

The general procedure of hypothesis testing

1. Formulate the null hypothesis H_0 and the alternative hypothesis H_1,
2. Calculate the test criterion T,
3. Find the critical value K for a given nivel of test α (i.e. define the critical region C),
4. Compare K and T, i.e. determine whether $T \in C$, and based on this, accept or reject the hypothesis H_0.

The conclusion and credibility of hypothesis testing

If $T \in C$, H_0 is rejected. If $T \notin C$, H_0 is accepted. Since the decision whether to reject or accept the null hypothesis depends on the limited amount of information contained in the corresponding data sample, we can make a mistake that is of one of the following two kinds:

a. We reject the null hypothesis which actually holds true. By doing so, we make a mistake of the first kind. The probability that this mistake happens is denoted α, and is called nivel of test.

b. We accept the null hypothesis which in fact is not true. By doing so, we make a mistake of the second kind the probability of which is denoted β. The probability 1- β is called *power of the test*. It is the probability that the null hypothesis will be correctly rejected.

Nivel of test α is usually set at 0,05, 0,01 or less frequently at 0,1. If this is the case, we talk about a 5% nivel of test, a 1% nivel of test and a 10% nivel of test, respectively.

Apart from the nivel of test, the so-called *p-values* are also used in hypothesis testing. These values are often a part of statistical software outputs. A p-value tells us the probability of getting or exceeding the test criterion. If the p-value is smaller than or equal to the defined nivel of test, the null hypothesis is rejected. In the opposite case, the null hypothesis is accepted.

Basic statistical tests

We shall now present the standard and frequently used statistical tests. These are

(A) One-sample t – test.
(B) Two-sample t – test with equal variances.
(C) Two-sample t – test with unequal variances.
(D) Paired t-test.
(E) Two-sample F – test of variance equality.

Each of the tests is described now by the four-step general procedure of hypothesis testing. The tests can also be performed in Excel if the Excel add-in module *Data Analysis* is selected. The module contains many statistical methods, including the standard statistical tests.

(A) Testing the population mean (one-sample t – test)

Let $X = (X_1,..., X_n)$ be a random sample from a normal distribution $N(\mu, \sigma^2)$, where the variance σ^2 is unknown.

1. We test the null hypothesis H_0: $\mu = \mu_0$ against the alternative H_1: $\mu \neq \mu_0$, where μ_0 is a given value.

2. The test criterion takes the form

$$T = \frac{\bar{X} - \mu_0}{S} \cdot \sqrt{n},$$

where

$\bar{X}$ = sample mean calculated from the values $X_1,..., X_n$,
S = sample standard deviation calculated from the values $X_1,..., X_n$,
μ_0 = assumed population mean which is tested by the statistician
n = sample size.

3. The critical value K is related to a Student's distribution with n-1 degrees of freedom, and for a nivel of test α, it is denoted $t_{n-1}(\alpha)$. The critical value is defined as the number that satisfies the relationship $P(|X| \geq t_{n-1}(\alpha)) = \alpha$, where X is a random variable following the Student's distribution. The critical value can be either found in statistical tables, or calculated in Excel using the function TINV(α; n-1). The critical region of the test is $C = (-\infty, -K] \cup [K, +\infty)$.

4. If $|T| \geq t_{n-1}(\alpha)$, H_0 is rejected and H_1 is accepted; in all the other cases, H_0 is accepted.

(B) Testing a difference between two population means (two-sample t-test with equal variances)

Let there be two independent random samples of sizes n_1 and n_2, respectively, from normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively. The variances σ_1^2 and σ_2^2 are unknown, but $\sigma_1^2 = \sigma_2^2$ is assumed.

1. We test the null hypothesis H_0: $\mu_1 = \mu_2$ vs. the alternative H_1: $\mu_1 \neq \mu_2$.

2. The test criterion T takes the form:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}} \cdot \sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 - 2)}{n_1 + n_2}},$$

where $\overline{X}_1$ is the sample mean calculated from the data obtained from $N(\mu_1, \sigma_1^2)$, $\overline{X}_2$ is the sample mean calculated from the data obtained from $N(\mu_2, \sigma_2^2)$. We also calculate the sample variance S_1^2 of the first sample, and the sample variance S_2^2 of the second sample. The numbers n_1 and n_2 represent the size of the first and second sample, respectively.

3. The critical value K is related to a Student's distribution with n_1+n_2-2 degrees of freedom, and is denoted $t_{n_1+n_2-2}(\alpha)$ for a nivel of test α. The value can be either found in statistical tables, or calculated in Excel with the function TINV(α; n_1+n_2-2).

4. If $|T| \geq t_{n_1+n_2-2}(\alpha)$, the null H_0 is rejected and H_1 is accepted; in the opposite case, the null hypothesis is accepted.

(C) Testing a difference between two population means (two-sample t-test with unequal variances)

Let there be two independent samples of sizes n_1 and n_2, respectively, from probability distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively. The sample means $\overline{X}_1, \overline{X}_2$ are calculated from the two samples, as well as their sample variances S_1^2, S_2^2, respectively. The population variances σ_1^2 and σ_2^2 are unknown, however, inequality $\sigma_1^2 \neq \sigma_2^2$ is assumed this time.

1. We test the hypothesis H_0: $\mu_1 = \mu_2$ vs. the alternative H_1: $\mu_1 \neq \mu_2$.

2. The test criterion takes the form:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{V_1 + V_2}},$$

where

$$V_i = \frac{S_i^2}{n_i},\ i = 1, 2.$$

3. The critical value K is calculated using the formula

$$K = \frac{V_1 \cdot t_{n_1-1}(\alpha) + V_2 \cdot t_{n_2-1}(\alpha)}{V_1 + V_2},$$

where $t_{n_1-1}(\alpha)$ and $t_{n_2-1}(\alpha)$ are critical values of a Student's distribution with n_1-1 and n_2-1 degrees of freedom, respectively, both for a nivel of test α. The value K can be obtained, using the Excel function TINV(α; n_1-1) for $t_{n_1-1}(\alpha)$, and TINV(α; n_2-1) for $t_{n_2-1}(\alpha)$.

4. If $|T| \geq K$, H_0 is rejected and H_1 is accepted; in the opposite case, H_0 is accepted.

(D) Paired t-test

Let $X = X_1, X_2, ..., X_n$ be a random sample from $N(\mu_1, \sigma_1^2)$, and $Y = Y_1, Y_2, ..., Y_n$ be a random sample from $N(\mu_2, \sigma_2^2)$. The corresponding sample means are $\bar{X}$ and $\bar{Y}$, respectively.

1. We test the hypothesis H_0: $\mu_1 = \mu_2$ vs. the alternative H_1: $\mu_1 \neq \mu_2$.

2. The test criterion takes the form

$$T = \frac{\bar{D}}{S_D}\sqrt{n},$$

where

$$S_D = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2}\ ,\ D_i = X_i - Y_i,\ i = 1, 2..., n,\ \bar{D} = \bar{X} - \bar{Y}\ .$$

3. The critical value $K = t_{n-1}(\alpha)$, and is related to a Student's distribution with n-1 degrees of freedom, and a nive of test α. It can be obtained with the Excel function TINV(α; n-1).

4. If $|T| \geq K$, H_0 is rejected and H_1 accepted; in the opposite case, H_0 is accepted.

(E) Two-sample F – test of equality of variances

Let us have two independent random samples from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively, of sizes n_1 and n_2, respectively. Let S_1^2 and S_2^2 be their respective sample variances.

1. We test the hypothesis that the population variances are the same, i.e. the hypothesis H_0: $\sigma_1^2 = \sigma_2^2$, against the alternative H_1: $\sigma_1^2 \neq \sigma_2^2$.

2. The test criterion takes the form:

$$T = \frac{\max(S_1^2, S_2^2)}{\min(S_1^2, S_2^2)}.$$

3. The critical value $K = F_{n_1-1, n_2-1}(\alpha)$ can be found in statistical tables for a Fisher's distribution with n_1-1 and n_2-1 degrees of freedom, and a nivel of test α. Alternatively, the critical value can also be calculated with the Excel function FINV(α; n_1-1; n_2-1).

4. If $T \geq K$, H_0 is rejected and H_1 is accepted; in the opposite case, H_0 is accepted.

As was mentioned previously, statistical hypotheses represent only a part of all scientic hypotheses – that part which concerns random variables. These hypotheses/tests include parametric and nonparametric tests. Parametric tests deal with parameter(s) of a given probability distribution. Nonparametric tests are concerned not with parameters of a distribution but other statistical properties of the distribution. It must be noted, however, that nonparametric tests are used as a term in a more general sense: the term is used for those tests that do not have to comply with as many mathematical conditions for their use as other tests do. As we saw earlier, t-tests, for instance, required several such conditions including the prerequisite that the random sample come from a normal distribution. There are situations when such prerequisite cannot be met, and the question how to proceed then naturally arises. There are more robust statistical tests that demand only very general conditions to be met for their justified use. This is when we talk about nonparametric tests, although we may use the tests to check the particular form of parameters of a probability distribution, as well. In order not to get entangled in this terminology, we shall continue to perceive the term nonparametric test as we did to this point, i.e. we shall view such a procedure as a method that tests other distribution properties than those related directly to the parameters of the distribution.

2.2 MARKETING STUDY

To demonstrate the tests described above and introduce some new tests, let us draw on a marketing study named **Studie**. The study will be used now and then for the description of other statistical methods. The procedures that follow are accompanied by Excel calculations.

Studie

A company wants to bring out a new nonalcoholic beverage: a cola-based carbonated drink. There are three versions of the product that are to hit the market: Kafola, Kofikola and Kofolisima. A questionnaire-based poll was run, and answers from 47 respondents were gathered about the consumption of the new products. The results of the poll are contained in table 2 (the data represents a weekly consumption of the corresponding beverage in litres).

Table 2: Studie poll results

Respondent	Sex	Age	Kafola	Kofikola	Kofolisima
1	m	20	1,1	0,7	0,5
2	f	34	1	0,2	0,1
3	f	43	0,8	0,1	0,2
4	f	21	1,2	0,6	0,3
5	m	39	1,1	0,1	0,4
6	f	51	0,4	0	0,2
7	m	19	0,9	0,9	0,3
8	f	45	0,3	0,2	0,2
9	f	48	1,2	0,1	0,4
10	f	21	1,4	0,4	0,2
11	f	52	0,4	0	0,3
12	f	22	1,2	0,6	0,4
13	m	62	0,2	0	0,2
14	f	47	0,6	0,2	0,1
15	m	23	0,9	0,8	0,2
16	m	35	0,9	0,1	0,4

Respondent	Sex	Age	Kafola	Kofikola	Kofolisima
17	m	22	1	0,9	0,1
18	m	38	0,5	0,2	0,2
19	f	41	0,4	0,1	0,1
20	f	21	0,9	0,7	0,2
21	f	40	0,2	0	0,3
22	f	20	0,8	0,6	0,3
23	m	19	1,1	0,9	0
24	m	39	1	0,1	0
25	m	19	0,9	1,1	0,4
26	f	38	0,2	0,2	0,5
27	f	20	1,3	1,5	0,3
28	f	37	0,4	0,1	0,8
29	m	20	1,3	0,8	0,2
30	f	41	0,1	0,2	0,1
31	m	42	0,2	0,1	0,2
32	f	20	0,9	0,9	0,3
33	m	43	1,2	0,2	0,1
34	f	21	0,9	0,7	0,2
35	m	44	0,1	0,1	0,1
36	f	45	0	0,1	0,2
37	m	46	0,1	0,2	0,1
38	m	22	1	0,9	0,2
39	m	42	0,4	0,8	0,3
40	m	41	0,1	0,1	0,4
41	f	22	1,1	0,5	0,2
42	f	40	0,2	0,1	0,1
43	f	21	1,3	0,8	0
44	m	39	0,4	0,9	0,2
45	f	20	1,1	0,1	0,1
46	m	20	1	0,2	0,3
47	f	21	0,8	0,1	0,4

Source: author's

We note that the number of respondents is not particularly high in this case. Marketing studies usually address hundreds of respondents. Of course, we are primarily interested in the principles of working with the marketing data. Those principles are the same regardless of how many respondents take part in the poll.

We shall now perform the statistical tests described at the beginning of chapter two. The tests will try to answer various questions that could have been asked by the client who ordered the poll. Some of the tests will also be presented as a one-sided test. This is a kind of test the null hypothesis of which appears in the form of an inequality, not an equation. The alternative hypothesis remains the direct negation of the null hypothesis. We shall also use table 2 to describe and demonstrate other widely exploited statistical tests.

PROBLEM 1 (one-sample t-test)

Let us find out at five per cent nivel of test whether we can assume that the average weekly consumption of Kafola equals 0,7 litres. Thus, $H_0 : \mu = 0,7$ is tested against its alternative $H_1 : \mu \neq 0,7$; $\alpha = 0,05$. The test criterion satisfies

$$T = \frac{0,833 - 0,7}{0,166}\sqrt{47} = 2,239.$$

The sample average used in the criterion is calculated from the data in the column „Kafola", and it is equal to 0,833. The sample standard deviation is *s* = 0,166 (this may be calculated using its defining formula from chapter one, or by taking the square root of the Excel function *var*(), where the single argument is the reference to the column Kafola). Since it is possible to chain Excel functions, the function *sqrt*(*var*()) will give the standard deviation, as well.

The sample size is 47, therefore we work with a *t*-distribution (or Student's distribution) with 47-1 degrees of freedom. The critical value of the distribution is *K* = TINV(0,05, 46) = 2,012. Since the test criterion in absolute value exceeds the critical value, we reject the hypothesis that the average weekly consumption of Kafola equals 0,7 litres in the population.

The one-sample t-test can also be performed **in the form of a one-sided test**. In such a case, we formulate the null hypothesis as H_0: $\mu < \mu_0$ and its alternative as H_1: $\mu \geq \mu_0$. We use the test (A) described at the beginning of this chapter, however, in a slightly modified form.

1. We test H_0: $\mu < \mu_0$ against H_1: $\mu \geq \mu_0$

2. The test criterion takes the form

$$T = \frac{\bar{X} - \mu_0}{S} \cdot \sqrt{n},$$

where

$\bar{X}$ = sample average,
S = sample standard deviation,
μ_0 = assumption about the unknown population average; in our case, it is 0,7,
n = sample size; in our case, this is 47.

3. The critical region of the one-sided test is interval $C = [K, +\infty)$ given by the critical value *K* of the Student's distribution with *n*-1 degrees of freedom. The critical value in this case is such a value that the probability of exceeding it is equal to the preset nivel of test alpha. Given the definition of the Student's distribution critical values, it means that $K = t_{n-1}(2\alpha)$. This number can be calculated using the Excel function TINV(2α; *n*-1).

4. If $T \geq K$, H_0 is rejected and H_1 accepted. In the opposite case, H_0 is accepted.

As can be seen, the one-sided version of the t-test is very similar to its two-sided version. The difference is in the calculation of the critical value and the conclusion of the test. In our case, if we formulate the null hypothesis H_0: $\mu < 0{,}7$ against the alternative H_1: $\mu \geq 0{,}7$, the test criterion results in the same number, of course, but the critical value at five per cent nivel of test is equal to TINV(2·0,05; 46) = 1,68, and we again reject the tested hypothesis.

We could also try to answer the question for what nivel of test the null hypothesis would be accepted in the one-sided version of the test. In our case, we accept the null hypothesis if and only if $T < K$. To answer the question, it is convenient to use the Excel function TDIST(*K; n; tails*). The function returns *alfa* which satisfies the equality $P(|X| \geq K) = \alpha$, where X follows a t-distribution with n degrees of freedom if the argument „*tails*" is set at 2; alternatively, the function returns *alfa* satisfying the equation $P(X \geq K) = \alpha$, where X follows a t-distribution with n degrees of freedom if the argument „*tails*" is set at 1. In our case, if we substitute K for T, and use the function TDIST, we find that the probability of exceeding T is equal to TDIST(2,239; 46; 1) = 0,015. Thus, if $T < K$ is to hold, the critical value K must be such that the probability of exceeding it is smaller than 0,015. This probability, however, is called nivel of test. So the conclusion in the one-sided version of the test is such that any nivel of test smaller than 0,015 will lead to acceptance of the null hypothesis.

PROBLEM 2 (two-sample t-test with equal variances)

Let us demonstrate how to perform the two-sample t-test with equal variances. The equality of variances is assumed to hold true at the moment, although this should be tested, as well, using another statistical test (we will talk about the test later). Our objective now is to find out, using Studie, whether the average consumption of Kofikola is the same as that of Kofolisima. The nivel of test is five per cent. The test criterion satisfies

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}} \cdot \sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 - 2)}{n_1 + n_2}},$$

from which it is obvious that we need to calculate the sample averages in both data samples (the column „Kofikola" represents one sample in this case, the column „Kofolisima" is the other sample) and also the sample variances and samples sizes. These characteristics are contained in table 3. Again, the calculation of the characteristics can be carried out using the corresponding defining formulas or the Excel functions.

Table 3: entry characteristics for the two-sample t-test

	Kofikola	**Kofolisima**
average	0,40851064	0,24042553
variance	0,14123034	0,02289547
sample size	47	47

Source: author's

Thus, using the data from table 3, the test criterion T = 2,844. The critical value K = TINV(0,05; 47+47-2) = 1,986. Therefore, the hypothesis of equal population means is rejected, the nivel of test being five per cent.

Excel: The test can be performed, using the Excel analytical tools located at **Data/Data Analysis** or **Tools/Data Analysis**. If Excel is installed for the first time, it may happen that the Data Analysis module is hidden in the program. In such cases, the module must be installed through Tools/Add-ins (older Excel versions) or through File/Options/Add-ins (newer Excel versions). The module contains nineteen statistical methods of which five are related to statistical tests. The main advantage of the module is that the corresponding formulas don't have to be constructed and calculated. Everything is done automatically by the module itself. Each test is integrated into a single dialogue window, and its results are presented in a unified table.

If we run the module Data Analysis, a dialogue window with various statistical methods pops up. We select the two-sample t-test with equal variances, and confirm the option. In the subsequent window that Excel offers, we place the computer mouse to the sub-window Sample 1 and designate the area in the spreadsheet containing the data of interest – in our case, the data for the Kofikola consumption, for instance. Similarly, the Sample 2 sub-window will contain a reference to the area in the Excel spreadsheet containing the data on Kofolisima consumption. We keep the default alpha at 0,05 as well as the default output location, offered by the dialogue window. After confirming these options, Excel generates the following table 4 with all necessary results.

Table 4: Excel output for the two-sample t-test with equal variances

	sample 1	*sample 2*
Mean	0,408510638	0,240425532
Variance	0,141230342	0,022895467
Sample size	47	47
:	:	
:	:	
:	:	
t Stat	**2,84439379**	
P(T<=t) (1)	0,002741733	
t krit (1)	1,661585397	
P(T<=t) (2)	0,005483465	
t krit (2)	**1,986086317**	

The table contains characteristics necessary for carrying out the test, and also the test criterion **t Stat** and the critical vaue of the two-sided test **t krit (2)**. Both values confirm that our previous calculations were correct. The conclusion therefore is the same.

PROBLEM 3 (F-test)

If the two-sample t-test with equal variances is to be credible, we must confirm whether the assumption of equal population variances is correct. We shall now test the assumption at one per cent nivel of test. The test criterion related to this test is of the form

$$T = \frac{\max(S_1^2, S_2^2)}{\min(S_1^2, S_2^2)}.$$

It's a division of two sample variances. In our case, the sample variance for the weekly consumption of Kofikola is $S_1^2 = 0{,}141$, and the sample variance for Kofolisima $S_2^2 = 0{,}0228$. Therefore, T = 6,168. The critical value of the test K = FINV(0,01;47-1;47-1) = 2. This means the null hypothesis is rejected with a very small nivel of test.

Excel: The same test can be realized using the Data Analysis module in Excel if we select the F-test of equal variances option in the module. In order for this procedure to give the same result, it is necessary that the data sample with the higher sample variance be used as Sample 1 in the dialogue window that follows the confirmation of the option of the F-test in the module. In our example, the higher variance relates to the data sample on Kofikola: $S_1^2 = 0{,}141$. Therefore, Sample 2 option in the dialogue window will contain the reference to the data on Kofolisima. The nivel of test alpha is 0,05 by default. We shall reset the level at 0,01 for our purposes. Confirming the options, Excel returns results in the form of table 5.

Table 5: Excel output on the F-test of equal variances

F-test of equal variances

	Sample 1	*Sample 2*
Mean	0,408510638	0,240425532
Variance	0,141230342	0,022895467
Observations	47	47
Difference	46	46
F	**6,168484848**	
P(F<=f) (1)	3,7416E-09	
F krit (1)	**2,006833595**	

Here, F stands for the test criterion and F krit (1) stands for the critical value of the test: the real number such that the probability of exceeding it by the test criterion is 0,01. In this case, the critical value equals 2,007.

As we can see, the procedure used in the previous problem, where we worked with the two-sample t-test with equal variances, was not appropriate, as it was based on the assumption of equal variances. This assumption has just been rejected. The appropriate procedure is to use the two-sample t-test with unequal variances, as demonstrated in the following problem.

PROBLEM 4 (two-sample t-test with unequal variances)

We said that in this case the test criterion takes the form

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{V_1 + V_2}},$$

where $V_i = S_i^2 / n_i$, i = 1, 2, S_i^2 = sample variance of the i-th sample and n_i = size of the i-th sample. Applying these formulas to our case of nonalcoholic beverages, we get

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{V_1 + V_2}} = \frac{0,408 - 0,24}{\sqrt{0,14/47 + 0,0228/47}} = 2,85.$$

The critical value of the test is

$$K = \frac{V_1 \cdot t_{n_1-1}(\alpha) + V_2 \cdot t_{n_2-1}(\alpha)}{V_1 + V_2} = \frac{0,003 \cdot 2,013 + 0,00048 \cdot 2,013}{0,003 + 0,00048} = 2,013.$$

The conclusion of the test is such that we reject the null hypothesis of equal means.

Excel: using the Data Analysis module in Excel, we can select the two-sample t-test with unequal variances option. Upon confirming the selection, we fill in the required information in the corresponding dialogue window: we insert references to the areas of the Excel spreadsheet containing the data on Sample 1 and Sample 2, just like in the case of the two-sample t-test with equal variances. Also, we set the nivel of test at alpha (we leave it at five per cent here). The output of Excel calculations is contained in table 6.

Table 6: Excel output of the two-sample t-test with unequal variances

Two-sample t-test with unequal variances

	Sample 1	*Sample 2*
Means	0,408510638	0,240425532
Variances	0,141230342	0,022895467
Observations	47	47
:	:	
:	:	
t Stat	**2,84439379**	
P(T<=t) (1)	0,00302417	
t krit (1)	1,670219484	
P(T<=t) (2)	0,006048339	
t krit (2)	**1,999623585**	

Let us comment on the results of table 6: t stat represents the test criterion, t krit (2) is the critical value. It is necessary to note that as in the case of several fundamental statistical characteristics (skewness and kurtosis, in particular), Excel carries out some calculations differently, compared to how such calculations are performed in rigorous statistical texts. The critical value of the test is calculated in more than one way, to be more precise. Different procedures aim to approximate differently the degrees of freedom of the Student's distribution related to this test. Therefore, it is quite likely that the critical value for this test provided by Excel will differ from the one defined at the beginning of this chapter. Nonetheless, despite the difference in the critical value, the conclusion to our problem remains the same. Also, as demonstrated above, the difference in the critical values is certainly not severe.

PROBLEM 5 (paired test)

We emphasize again that while applying the two-sample t-tests with equal or unequal variances, we assume, among other things, that the two samples are independent of each other. This is a very important prerequisite. If this is not the case, it is better to apply the paired test. If the condition of independence is met, and the analyst applies the paired test instead of the two-sample t-test, nothing serious happens. However, such a procedure is not optimal because the paired test requires two samples of the same size, as opposed to what is required by the two-sample t-test. Thus, if the analyst wants to use the paired test, it might be the case that he or she will have to throw away some of the data provided the samples to be worked with are of different sizes. On the contrary, if the situation requires the paired test to be applied because of the conditions necessary for this test, and the analyst uses the two-sample t-test instead, it will be a serious mistake, and the conclusions based on such a technique will be completely incredulous.

We shall use the paired test now to find out whether the average weekly consumption of Kofikola is the same as the average weekly consumption of Kofolisima. First, let us substract the values on consumption, which are in the same row of table 2, and are in columns Kofikola or Kofolisima. Doing so, we get the substractions $D_i = X_i - Y_i$, where X_i is the consumption of Kofikola from the *i*-th row of table 2 and from column Kofikola, and Y_i is the consumption of Kofolisima from the *i*-th row of the same table and from column Kofolisima. The average substraction equals $\bar{D} = \bar{X} - \bar{Y} = 0,408 - 0,24 = 0,168$. Secondly, let us calculate the sample standard deviation of the substractions

$$S_D = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2} = 0,4076.$$

Here, $n = 47$ = sample size = number of rows in table 2. The test criterion of the paired test is

$$T = \frac{\bar{D}}{S_D}\sqrt{n} = (0,168 / 0,4076)\cdot\sqrt{47} = 2,826.$$

The critical value $K = t_{n-1}(\alpha) = TINV(0,05;47-1) = 2,013.$ Since $|T| \geq K$, H_0 is rejected.

Excel: Selecting the Data/Data Analysis/Two-sample paired t-test option in Excel and using subsequently the Sample 1 and Sample 2 options for references to the Excel Spreadsheet area with data on Kofikola and Kofolisima, we get the following output (Table 7).

Table 7: Excel output for the paired test

Two-sample paired test of means

	Sample 1	*Sample 2*
Averages	0,408510638	0,240425532
Variances	0,141230342	0,022895467
Observations	47	47
Pears. correlation	-0,017650836	
Hyp. Difference of means	0	
t Stat	**2,827157048**	
P(T<=t) (1)	0,003464955	
t krit (1)	1,678660414	
P(T<=t) (2)	0,006929909	
t krit (2)	**2,012895599**	

T stat corresponds to the test criterion of the paired test, t krit (2) is the critical value of the test. As is confirmed by the table, Excel calculations correspond to the calculations we made by hand.

In the next section of the chapter, we shall present some other statistical tests. We will describe their purpose, and use the study material Studie to show how to proceed in their case.

2.3 MEDIAN TEST

It is one of the tests that doesn't require too many conditions for its use. As the name of the test suggests, median test tries to confirm or reject the hypothesis about the median of a probability distribution. If such a distribution possesses the property that its median is equal to its mean, the test can be regarded as an alternative to the one-sample t-test. The only condition that must be met for the median test to be justified is the requirement that the respective data sampling be made in a population with continuous probability distribution. Thus, normality is not required here, as in the case of one-sample t-test.

Let us denote the unknown median as $\tilde{\mu}$, and the size of the data sample used to perform the test as *n.* We assume that *n* is large enough since the precision of the test we are about to describe improves with *n* increasing.

1. We test H_0: $\tilde{\mu} = \tilde{\mu}_0$ versus H_1: $\tilde{\mu} \neq \tilde{\mu}_0$. Here, $\tilde{\mu}_0$ is a specific value defined by the statistican.

2. The test criterion is $T = \frac{|2m - n|}{\sqrt{n}}$, where *m* is the number of observations in the data sample, which are smaller than $\tilde{\mu}_0$.

3. The critical value $K = z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the critical value of the standard normal distribution N(0,1) for a nivel of test α, i.e. it is the real number $z_{1-\alpha/2}$ such that the

probability of exceeding it equals $1-\alpha/2$. The critical value can be found either in statistical tables or using the Excel function NORMSINV(1-α/2).

4. If $T \geq K$, H_0 is rejected. In the opposite case, H_0 is accepted.

PROBLEM 6 (Median test)

Let us test the hypothesis that the average age of the cola-based drink consumer is 33 years (this is supposed to be the median age). The nivel of test is five per cent. The data available for the test are contained in Studie. Looking at the data, we see that in 20 out of all the 47 cases, the age of the consumers is below 33. Therefore,

$$T = \frac{|2m-n|}{\sqrt{n}} = \frac{|2\cdot 20-47|}{\sqrt{47}} = 1,02.$$

The critical value $K = z_{1-\alpha/2}$ = NORMSINV(1-0,05/2) = 1,96. This leads to the conclusion that the null hypothesis is accepted.

2.4 CHI-SQUARED TESTS

The last category of statistical tests we are going to deal with is related to chi-squared tests. We shall demonstrate two of these tests, as they are frequently used in social surveys. The first test focuses on the type of the probability distribution the data sample used for the test came from, the second test verifies the hypothesis of statistical independence of two random variables. Since a chi-squared distribution is exploited in the two tests, it is clear where the name of the tests originated.

2.4.1 TESTING A DISCRETE PROBABILITY DISTRIBUTION

As is well known from mathematical statistics, the most frequently used probability distributions are of discrete or continuous type. The chi-squared test can be used for any of these two situations. For simplicity, we shall work with discrete distributions only. What follows is a theoretical description of the test and an example that demonstrates its purpose.

Let X be an observed variable (not necessarily a numerical variable). A type of beverages consumed may serve as an example of such a variable. Let us assume that there are k different categories of the variable: $X_1, X_2, ..., X_k$ (k different types of beverages, for instance). Let p_i represent the relative frequency of occurrence of the i-th category X_i in the population. If we sample data **randomly** from this population, and write down the absolute frequencies with which different categories $X_1, X_2, ..., X_k$ occured in the sample, we can view the absolute frequencies as realizations of the random variables $X_1, X_2, ..., X_k$. The expression

$$T = \sum_{i=1}^{k} \frac{\left(X_i - np_i\right)^2}{np_i}$$

is then a random variable itself, following approximately a chi-squared distribution with *k*-1 degrees of freedom. The higher the *n*, the more precise the approximation is. The absolute frequency of the *i*-th category X_i in the sample is called **empirical frequency**, the term np_i is called **theoretical or expected frequency.**

This setting can be used to test a hypothesis about the p_i's, *i* = 1, 2, …, *k*, concerning the variables $X_1, X_2, ..., X_k$. This means from the statistical point of view that parameters of a probability distribution (multinomial in this case) are tested. To do the test, a null hypothesis about the parameters p_i is formulated, a sampling is carried out from the corresponding population, and the aforementioned test criterion *T* is calculated. If $T \geq \chi^2_{k-1}(\alpha)$, where $\chi^2_{k-1}(\alpha)$ is the critical value for a chi-squared distribution with *k*-1 degrees of freedom and a five per cent nivel of test, the null hypothesis is rejected. In the opposite case, when $T < \chi^2_{k-1}(\alpha)$, the null hypothesis is accepted. The critical value can be found either in statistical tables or by using the Excel function CHIINV(alpha, k-1).

PROBLEM 7 (Chi-square test)

To demonstrate the test, let us use the data on cola-based drinks. The population is represented by the Faculty of Business Administration of the Silesian University. We assume that only cola-based drinks are sold at the faculty. We are interested in whether it is true that all the three cola-based drinks are consumed in the same amounts. Statistically speaking, this means that we test whether the variable *X*, a cola-based drink, follows a uniform distribution. We set the nivel of test at five per cent.

Table 8 depicts a random sampling result – absolute frequencies for each of the three drinks.

Table 8: Frequencies of consumed drinks

	Number of bottles
Kofola	87
Kofikola	93
Kofolisima	101

Source: author's

We have *n* = 87+93+101 = 281, *k* = 3. The null hypothesis is $H_0 : p_1 = p_2 = p_3 = 1/3$. The test criterion takes the form

$$T = \sum_{i=1}^{k} \frac{\left(X_i - np_i\right)^2}{np_i} = \frac{(87 - 281/3)^2}{281/3} + \frac{(93 - 281/3)^2}{281/3} + \frac{(101 - 281/3)^2}{281/3} = 1,053.$$

The critical value *K* = CHIINV(0,05,3-1) = 5,99. Therefore, we accept the null hypothesis about the equal consumption of all the three drinks.

2.4.2 CHI-SQUARED TEST OF INDEPENDENCE

There is also another problem related to chi-square testing, and a *contingency table* is constructed to solve the problem. Two variables are assumed: a variable *A* (sex status: male or female, for instance) and a second variable *B* (remuneration at work, for example). The variable *A*, a *classification variable*, exists in two forms A_1 and A_2. Similarly, the variable *B* exists in *s* possible forms $B_1, B_2, ..., B_s$, $s \geq 2$. A contingency table (see table 9) is constructed.

Table 9: Contingency table for the chi-squared test of independence

Categories of *A* / *B*	B_1	B_2	B_3	...	B_s	Sum
A_1	n_{11}	n_{12}	n_{13}	…	n_{1s}	$n_{1.}$
A_2	n_{21}	n_{22}	n_{23}	…	n_{2s}	$n_{2.}$
Sum	$n_{.1}$	$n_{.2}$	$n_{.3}$	…	$n_{.s}$	n

Source: author's

The symbol n_{ij} stands for the number of cases when the variable *A* took on the value (category) A_i, and at the same time the variable *B* reached the level (category) B_j. The symbol $n_{i\bullet} = \sum_{j=1}^{s} n_{ij}$ expresses the number of cases when *A* fell in the *i*-th category, regardless of what category the variable *B* fell in, and the symbol $n_{\bullet j} = n_{1j} + n_{2j}$ represents similarly the number of cases when *B* fell in the *j*-th category. These frequencies are called *marginal frequencies*.

We want to know whether the two variables *A* and *B* are statistically independent.

The test procedure

1. We are testing, for a nivel of test alpha, the hypothesis H_0: *A* and *B* **are independent variables** vs. the alternative hypothesis H_1: *A* a nd*B* **are not independent**.

2. The test criterion is of the form

$$T = \sum_{j=1}^{s} \sum_{i=1}^{2} \frac{\left(n_{ij} - n_{ij}^{T}\right)^2}{n_{ij}^{T}},$$

where $n_{ij}^{T} = (n_{i\bullet} \cdot n_{\bullet j}) / n$ are theoretical (expected) frequencies. The values n_{ij} represent empirical frequencies acquired by random sampling.

3. The critical value is $K = \chi_{s-1}^{2}(\alpha)$.

4. Conclusion of the test: if $T \geq K$, H_0 is rejected. In the opposite case, H_0 is accepted.

PROBLEM 8

Let *A* be a variable = sex status of respondents and another variable *B* = form of remuneration awarded to respondents in competitive sport events. Table 10 shows the numbers of respondents who were randomly selected, and fall into different categories classified by the two variables.

Table 10: data on types of remuneration and sex status of respondents

Sex/Remuneration	**Financial reward**	**Soft drink**	**Subtotal**
Men	78	42	120
Women	46	34	80
Subtotal	124	76	200

Source: author's

For example, the number 78 tells us that this is the number of respondents who were men and who said that they had received a financial reward for their sports achievement. The subtotals are calculated by the poll worker who is to process the data.
Tables 11 and 12 contain the calculations of the theoretical frequencies (left table) and the terms appearing in the sum of the test criterion (right table). The symbol E_{ij} stands for the empirical frequency and the symbol O_{ij} describes the corresponding theoretical frequency. The empirical frequencies are calculated from table 10, taking the appropriate row and column marginal frequencies (subtotals¨), multiplying them and dividing by the number of all respondents (which is 200). Therefore, we get, for instance, $74,4 = 120 \cdot 124 / 200$ for the first theoretical frequency, and a similar procedure applies to other theoretical frequencies, as well.

Table 11: theoretical frequencies

O_{ij}	**Fin. reward**	**Soft drinks**
Men	74,4	45,6
Women	49,6	30,4

Table 12: terms for the test criterion

$(E_{ij}-O_{ij})^2/O_{ij}$	
0,174193548	0,284210526
0,261290323	0,426315789

Source: author's

The final table 13 contains the test criterion *T*, degrees of freedom df = s-1 of the test, and the critical value *K* for 5% nivel of test.

Table 13: The test criterion *T* and critical value *K*

T	1,14
alfa	0,05
df	s-1 = 2-1
K	3,84

Source: author's

T = 1,14 and the critical value K = CHIINV(0,05; 2-1) = 3,84. Since the test criterion is smaller than the critical value, we accept the hypothesis that there is no relation between the form of reward and the sex status of the rewarded. As a final note, the chi-squared test can be realized in Excel, using the function CHITEST(actual; expected) which has two parameters: the parameter „actual" is a reference to the Excel spreadsheet area containing the empirical frequencies, whereas the parameter „expected" is a reference to the spreadsheet area with the expected/theoretical frequencies. The function returns a *p-value*, which means that the conclusion of the test is constructed as follows: if the *p-value* is smaller that the nivel of test alpha (or equal), the null hypothesis is rejected; if the *p-value* is greater than alpha, the null hypothesis is accepted.

CONTROL TEST 2

a. The data appearing in the table below represents the result of a random sampling related to a variable *Y*. Using the one-sample t-test, find out whether the population mean of *Y* is 17,8, the nivel of test being five per cent. As part of the calculations, state the value of the test criterion and the critical value. Will the conclusion of the test change if the nivel of test at ten per cent is used instead? Provide the critical value for the latter case.

Y	16	15	17	18	19	14	13

Source: author's

b. Let the data on two variables *Y* and *X* is available:

X	6	25	17	18	29	4	15

Y	16	15	17	18	19	14	34

Source: author's

Find out, using the F-test, whether both samples came from populations with the same variance, the nivel of test being set again at five per cent. State the test criterion and critical value as part of your calculations. Check your results against those provided by the Data Analysis module of Excel.

c. Perform the two-sample t-test with equal variances and confirm or reject the hypothesis that the variables X and Y have the same population mean.

d. A poll concerning cell phone trademarks that are popular among customers led to the results shown in the following table. Set the nivel of test at ten per cent, and test the validity of the hypothesis that 25% of all customers use the Mobil1 cell phones, 33% of

all the customers incline to the Mobil2 cell phones and the remaining 42% of the customers prefer the Mobil3 trademark. Again, state the test criterion and critical value.

	Number of users
Mobil1	2340
Mobil2	3124
Mobil3	3000

Source: author's

e. Using the test of independence, verify whether severity of car crash depends on sex status of car driver. Do so with one per cent nivel of test. As part of your calculation, state the test criterion, the critical value and the test conclusion. Available are following data:

	Man	**Woman**
Minor accidents	134	127
Accidents of intermediate severity	254	301
Severe accidents	14	4

Source: author's

SOLUTIONS

a. Test criterion = -2,2. Critical value = 2,44. The test criterion in absolute value is smaller than the critical value, implying that the null hypothesis is accepted. If the nivel of test was ten per cent, the critical value would be 1,94. In the latter case, the null hypothesis would be rejected.
b. Test criterion = 1,78. Critical value = 4,28. The null hypothesis on variance equality is accepted.
c. Test criterion = 0,63. Critical value = 3,05. The null hypothesis on equality of means is accepted.
d. Test criterion = 43,25. Critical value = 4,6. The hypothesis on uniform distribution is rejected.
e. Test criterion = 8,65. Critical value = 9,21. We accept the hypothesis that severity of car crash and car driver sex status are two independent variables.

3 REGRESSION ANALYSIS

Regression analysis deals with dependence of a quantitative variable on one or more quantitative variables. In the case of one variable depending on another variable, we talk about simple regression, as opposed to the case when there are more explanatory variables. In the latter case, we talk about multiple regression. In this chapter, the reader should deepen their knowledge on regression presented in the course Statistics [5], in particular, as regards the multiple regression. Elementary regression terms and conditions are presented at the beginning of this chapter. Further, a formula for the calculation of regression coefficients is derived, as well as a statistical test verifying significance of the coefficients. At the end of the chapter, statistical significance of the entire regression model is tested.

3.1 THE CONCEPT OF REGRESSION ANALYSIS

Regression analysis aims to find a mathematical relation – an equation which in a certain sense describes changes of a random variable Y dependent on changes of random variables $X_1, X_2, \ldots, X_k$. We shall assume the standard case presented in literature, i.e. the case when only some values of the variables $X_1, X_2,\ldots, X_k$ are known or available. These realizations of the random variables are denoted x_{ij} = the i-th value of the j-th variable X_j. As far as he values are concerned, they are usually a part of a controlled experiment in which the analyst defines/selects the values of $X_1, X_2,\ldots, X_k$, and then finds or measures the values of Y that corresponds to the values of $X_1, X_2,\ldots, X_k$. The value of Y, measured or found for the i-th value of $X_1, X_2,\ldots, X_k$, is denoted Y_i. To give an example, let Y = GDP which is influenced by factors $X_1, X_2,\ldots, X_k$. Different constellations of the factors will give a different GDP value, the behaviour of GDP being a random variable, as well, since it is almost certain that we will not be able to define k factors that describe it completely. Thus, GDP and so the variable Y , as well, will be in general a random variable, and its probability distribution will change as the level of the factors $X_1, X_2,\ldots, X_k$ changes. Therefore, we use the lower index i in the symbol Y_i. In this controlled experiment, which tries to find a <u>concrete</u> form of the relation between Y and the variables $X_2, \ldots, X_k$ on a *specified subset* of the set of all possible values of $X_1, X_2, \ldots, X_k$, we assume that the relation takes the form $Y = f(X_1, X_2,..., X_k) + \varepsilon$. Here, Y depends on the **regression function** f, which contains unknown parameters, and on the random term ε which completes the full description of the random behaviour of Y. The systematic part of the model $f(X_1, X_2,..., X_k)$ is not able to provide the full description of the behaviour of Y. As was already outlined at the beginning of the chapter, the problem of finding an appropriate relation between the variables will be resolved for the case when the variable Y, the so-called *dependent variable*, depends on k *independent variables* or a vector $X = (X_1,..., X_k)$.

The systematic part $f(X_1, X_2,..., X_k)$ can take on different forms:

$$f(X_1, X_2,..., X_k) = \beta_1 + \beta_2 X_1,$$
$$f(X_1, X_2,..., X_k) = \beta_0 + \beta_1 X_1 + \beta_2 X_2^2, \text{ etc.}$$

The parameters $\beta_0, \beta_1, \beta_2$ are unknown!

If the systematic part satisfies $f(X_1, X_2, ..., X_k) = \beta_1 f_1(X) + \beta_2 f_2(X) + ... + \beta_k f_k(X)$, we talk about *linear regression* (linear in parameters), or about *linear regression model*. We usually consider the model:

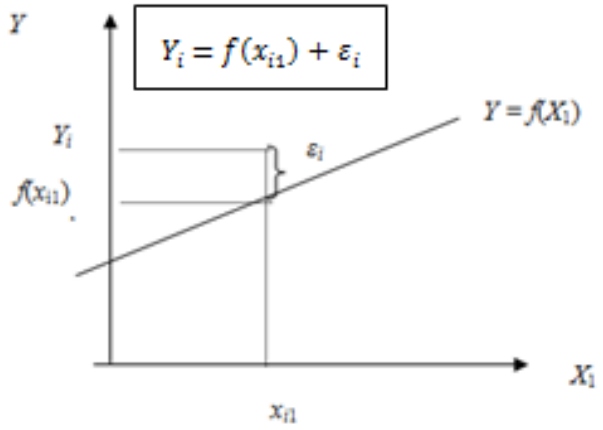
3-1 $$f(X_1, X_2, ..., X_k) = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k .$$

We shall start with its simplest form in which *f* is a linear function of one independent variable:

3-2 $$f(X_1) = \beta_1 + \beta_2 X_1 .$$

Thus, we consider the relation $Y = \beta_1 + \beta_2 X_1 + \varepsilon = f(X_1) + \varepsilon$. Our situation is depicted in figure 1.

Figure 1: Regression dependence in the form of a line



The graph shows that the behaviour of *Y* is determined by the systematic part of the model, i.e. by the function $f(X_1)$ which reflects the effect of a single variable on the variable *Y*. However, it does not suffice to use the function $f(X_1)$ to describe the behaviour of *Y*, and it is necessary to add the influence of other factors – those which are represented by the term ε. And what was just said is also true for any specific value of X_1, of course: for a value x_{i1}, for example. At the point x_{i1}, equation $Y_i = f(x_{i1}) + \varepsilon_i$ holds true, where $f(x_{i1})$ is a specific value. Nevertheless, even though $f(x_{i1})$ is a specific value, we don't know this value because the expression $f(x_{i1})$ depends on unknown parameters. We may even know the exact mathematical form of the expression (for instance, we may know that it is a line), and we still won't be able to evaluate it. **The objective of regression is to estimate the unknown parameters**.

To make the estimation, we need to have some data. For the estimation of a regression line, for instance, data $(x_{11}, y_1), (x_{21}, y_2), ..., (x_{n1}, y_n)$ are usually available. In other words, *n* points from a plane are available. The first coordinates $x_{11}, x_{21}, ..., x_{n1}$ of these points are the specified values of the independent variable X_1. These values are defined by the analyst, and the corresponding y's are obtained later. In our case, we obtained a single value of *Y* for each value of X_1.

The estimate of the function $f(X_1, X_2, ..., X_k) = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$, calculated for the i-th value of the variables $X_1, X_2, ..., X_k$, is denoted as $\hat{Y}_i = b_1 x_{i1} + b_2 x_{i2} + ... + b_k x_{ik}$.
If Y depends on two variables X_1, X_2, the points, obtained experimentally, will be of the form:

$$(x_{11}, x_{12}, y_1)$$
$$(x_{21}, x_{22}, y_2)$$
$$...$$
$$(x_{n1}, x_{n2}, y_n).$$

These points lie in a three-dimensional space, and are interspersed with a function of the form $\hat{Y} = b_1 x_1 + b_2 x_2$. The function approximates the relation between Y and the variables X_1, X_2.

More generally, if Y depends on k variables $X_2,\ldots,X_k$, we assume that the following points from a k+1-dimensional space are available:

$$(x_{11}, x_{12},\ldots, x_{1k}, y_1)$$
$$(x_{21}, x_{22},\ldots, x_{2k}, y_2)$$
$$...$$
$$(x_{n1}, x_{n2},\ldots, x_{nk}, y_n)$$

These points are interspersed with a hyperplane of the form $\hat{Y} = b_1 x_1 + b_2 x_2 + ... + b_k x_k$. This function approximates the relation between Y and the variables $X_1, X_2,\ldots, X_k$.

To justify the procedures to be explained, which result in an estimation of the unknown regression coefficients, it is imperative that the following conditions are satisfied. The conditions are related to the random part ε of the regression model:

1. Expected value of ε_i is zero, i.e. $E(\varepsilon_i) = 0$ for each i.
2. Variance of ε_i is constant, independent of i, i.e. $Var(\varepsilon_i) = \sigma^2$ for each i.
3. Variables ε_i and ε_j are not correlated, i.e. $Cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$.
4. Variables ε_i's are normally distributed, i.e. $\varepsilon_i \sim N(0, \sigma^2)$ for each i.

As is usually the case, expected value is denoted as E, variance is denoted as *Var* and covariance uses the symbol *Cov*. If the reader forgot the symbols, we recommend revision of the foundations of statistics contained in the course Statistics.

3.2 ESTIMATION OF REGRESSION COEFFICIENTS

We assume a regression function of the form $f(X_1, X_2, ..., X_k) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$.
We determine the i-th value of each explanatory variable and obtain a vector $(x_{i1}, x_{i2},\ldots, x_{ik})$. We do so for $i = 1, 2,\ldots, n$, so we end up with n vectors (or points). We then find or measure a particular value of Y for each vector $(x_{i1}, x_{i2},\ldots, x_{ik})$, $i = 1, 2,\ldots, n$. Since there are n vectors, we shall obtain n values of Y. This is all we have to make the estimation. Therefore, aside from the conditions 1-4, this is what greatly affects the quality of the resulting estimates.

Of course, we talk about estimates because we only work with a data *sample*. The vector of unknown regression parameters $\vec{\beta} = (\beta_0, ..., \beta_k)$ corresponds to the set of all possible points of

the form (x_{i1}, x_{i2},…, x_{ik}, y_i). Estimating the unknown parameters, we end up with an estimator $\vec{b} = (b_0,...,b_k)$.

The vector $\vec{b} = (b_0,...,b_k)$ is obtained by

3-3
$$\vec{b}^T = \left(X^T.X\right)^{-1} X^T.\vec{Y},$$

where *X*, the matrix of regressors, satisfies

3-4
$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & ... & x_{1k} \\ 1 & x_{21} & x_{22} & ... & x_{2k} \\ ... & ... & ... & ... & ... \\ 1 & x_{n1} & x_{n2} & ... & x_{nk} \end{bmatrix},$$

and

3-5
$$\vec{Y} = (\mathrm{Y}_1, \mathrm{Y}_2, ..., \mathrm{Y}_n)^T.$$

Symbol Z^T denotes the transposition of matrix *Z*, Z^{-1} means the inverse of *Z*. To evaluate 3-3, the following data, as mentioned previously, must be available

$$(x_{11}, x_{12},\ldots, x_{1k}, y_1)$$
$$...$$
$$(x_{n1}, x_{n2},\ldots, x_{nk}, y_n).$$

PROBLEM 1

Estimate dependence of electricity consumption *Y* on power-supply distance X_1 and amount of electricity supplied X_2. The regression function is assumed to be of the form $f\left(X_1, X_2\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. Available are the following data:

Table 14: data for problem 1

X_1	X_2	Y
1,2	3,6	3,2
1,3	3,7	3,3
1,3	3,8	3,4
1,4	3,8	3,5
1,4	3,9	3,6
1,5	3,9	3,6
1,5	4	3,7
1,6	4	3,8
1,6	4,1	3,9
1,7	4,2	4

Source: author's

SOLUTION

Table 14 represents points which are used to construct the matrices *X* and *Y*:

$$X = \begin{bmatrix} 1 & 1,2 & 3,6 \\ 1 & 1,3 & 3,7 \\ 1 & 1,3 & 3,8 \\ 1 & 1,4 & 3,8 \\ 1 & 1,4 & 3,9 \\ 1 & 1,5 & 3,9 \\ 1 & 1,5 & 4 \\ 1 & 1,6 & 4 \\ 1 & 1,6 & 4,1 \\ 1 & 1,7 & 4,2 \end{bmatrix} \qquad \vec{Y} = \begin{bmatrix} 3,2 \\ 3,3 \\ 3,4 \\ 3,5 \\ 3,6 \\ 3,6 \\ 3,7 \\ 3,8 \\ 3,9 \\ 4 \end{bmatrix}$$

We shall now calculate the vector $\vec{b} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix}$ in several steps, using 3-3:

$$X^T \cdot X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1,2 & 1,3 & 1,3 & 1,4 & 1,4 & 1,5 & 1,5 & 1,6 & 1,6 & 1,7 \\ 3,6 & 3,7 & 3,8 & 3,8 & 3,9 & 3,9 & 4 & 4 & 4,1 & 4,2 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1,2 & 3,6 \\ 1 & 1,3 & 3,7 \\ 1 & 1,3 & 3,8 \\ 1 & 1,4 & 3,8 \\ 1 & 1,4 & 3,9 \\ 1 & 1,5 & 3,9 \\ 1 & 1,5 & 4 \\ 1 & 1,6 & 4 \\ 1 & 1,6 & 4,1 \\ 1 & 1,7 & 4,2 \end{bmatrix} = \begin{bmatrix} 10 & 14,5 & 39 \\ 14,5 & 21,25 & 56,8 \\ 39 & 56,8 & 152,4 \end{bmatrix}.$$

$$\left(X^T \cdot X\right)^{-1} = \begin{bmatrix} 245,2 & 108 & -103 \\ 108 & 60 & -50 \\ -103 & -50 & 45 \end{bmatrix}.$$

$$X^T \cdot Y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1,2 & 1,3 & 1,3 & 1,4 & 1,4 & 1,5 & 1,5 & 1,6 & 1,6 & 1,7 \\ 3,6 & 3,7 & 3,8 & 3,8 & 3,9 & 3,9 & 4 & 4 & 4,1 & 4,2 \end{bmatrix} \cdot \begin{bmatrix} 3,2 \\ 3,3 \\ 3,4 \\ 3,5 \\ 3,6 \\ 3,6 \\ 3,7 \\ 3,8 \\ 3,9 \\ 4 \end{bmatrix} = \begin{bmatrix} 36 \\ 52,56 \\ 140,82 \end{bmatrix}.$$

Using 3-3, we now obtain:

$$\vec{b}^T = \left(X^T \cdot X \right)^{-1} \cdot X^T \cdot \vec{Y} = \begin{bmatrix} 245,2 & 108 & -103 \\ 108 & 60 & -50 \\ -103 & -50 & 45 \end{bmatrix} \cdot \begin{bmatrix} 36 \\ 52,56 \\ 140,82 \end{bmatrix} = \begin{bmatrix} -0,78 \\ 0,60 \\ 0,90 \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}.$$

Therefore, the resulting regression function is $\vec{Y} = -0,78 + 0,60x_1 + 0,90x_2$.

Theoretical values

Theoretical values $\widehat{Y}_1, \widehat{Y}_2, ..., \widehat{Y}_n$ are calculated from the relation $\widehat{Y} = -0,78 + 0,60x_1 + 0,90x_2$ by inserting specific values x_1 and x_2 to the equation:

$$\widehat{Y}_1 = -0,78 + 0,60 \cdot 1,2 + 0,90 \cdot 3,6 = 3,18,$$
$$\widehat{Y}_2 = -0,78 + 0,60 \cdot 1,3 + 0,90 \cdot 3,7 = 3,33,$$
$$\dots$$
$$\widehat{Y}_{10} = -0,78 + 0,60 \cdot 1,7 + 0,90 \cdot 4,2 = 4,02.$$

In matrix form, the resulting values can be written as

$$\widehat{Y} = X \cdot \vec{b}^T = \begin{bmatrix} 1 & 1,2 & 3,6 \\ 1 & 1,3 & 3,7 \\ 1 & 1,3 & 3,8 \\ 1 & 1,4 & 3,8 \\ 1 & 1,4 & 3,9 \\ 1 & 1,5 & 3,9 \\ 1 & 1,5 & 4 \\ 1 & 1,6 & 4 \\ 1 & 1,6 & 4,1 \\ 1 & 1,7 & 4,2 \end{bmatrix} \cdot \begin{bmatrix} -0,78 \\ 0,6 \\ 0,9 \end{bmatrix} = \begin{bmatrix} 3,18 \\ 3,33 \\ 3,42 \\ 3,48 \\ 3,57 \\ 3,63 \\ 3,72 \\ 3,78 \\ 3,87 \\ 4,02 \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ \hat{y}_{10} \end{bmatrix}$$

Vector of residuals

The difference between the theoretical value and empirical value is called residual. In vector form, the difference $\vec{e} = \vec{Y} - \hat{Y}$ represents the vector of residuals. In our example, we have:

$$\vec{e} = \vec{Y} - \hat{Y} = \begin{bmatrix} 3,2 \\ 3,3 \\ 3,4 \\ 3,5 \\ 3,6 \\ 3,6 \\ 3,7 \\ 3,9 \\ 3,9 \\ 4 \end{bmatrix} - \begin{bmatrix} 3,18 \\ 3,33 \\ 3,42 \\ 3,48 \\ 3,57 \\ 3,63 \\ 3,72 \\ 3,78 \\ 3,87 \\ 4,02 \end{bmatrix} = \begin{bmatrix} 0,02 \\ -0,03 \\ -0,02 \\ 0,02 \\ 0,03 \\ -0,03 \\ 0,02 \\ 0,02 \\ 0,03 \\ -0,02 \end{bmatrix} = \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ e_{10} \end{bmatrix}$$

The differences are calculated by substracting the corresponding vector coordinates.

Variance of the estimates of regression coefficients

Since finding regression coefficients results in estimates of the unknown population coefficients, it is convenient to introduce variances of the estimates. The variances reflect the precision of the estimates, and can be found on the main diagonal of the matrix

3-6
$$Var(\vec{b}) = s^2 \cdot \left(X^T \cdot X \right)^{-1},$$

where $s^2 = \dfrac{\sum_{i=1}^{n} e_i^2}{n-k}$ is an estimate of the variance of ε. Also,

e_i = i-th residual,
n = sample size (number of points we work with),
k = number of parameters in the regression model.

For our example, we have:

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-k} = \frac{0,006}{10-3} = 0,0008571.$$

Thus,

$$Var(\vec{b}) = s^2 \cdot \left(X^T \cdot X \right)^{-1} = 0,0008571 \cdot \begin{bmatrix} 245,2 & 108 & -103 \\ 108 & 60 & -50 \\ -103 & -50 & 45 \end{bmatrix} = \begin{bmatrix} 0,2102 & 0,0926 & -0,0883 \\ 0,0926 & 0,0514 & -0,0429 \\ 0,0883 & -0,0429 & 0,0386 \end{bmatrix}.$$

The main diagonal of the rightmost matrix contains the coefficient variances:

$s^2(b_0) = 0{,}2102$, and the standard deviation estimate $s(b_0) = 0{,}4584$.
$s^2(b_1) = 0{,}0514$, and the standard deviation estimate $s(b_1) = 0{,}2267$.
$s^2(b_2) = 0{,}0386$, and the standard deviation estimate $s(b_2) = 0{,}1965$.

When the regression model and the variances of the coefficients are estimated, the result is usually expressed in the form of an equation with the estimated standard deviations, i.e. the square roots of the estimated variances, written under the estimated coefficients that appear in the equation:

$$\hat{Y} = \underset{(0{,}4584)}{-0{,}78} + \underset{(0{,}2267)}{0{,}60}\, x_1 + \underset{(0{,}1965)}{0{,}90}\, x_2$$

It may happen that the order of magnitude of the differences among the coefficients can be quite substantial. For instance, it can be the case that $b_1 = 200$ and $b_2 = 0{,}02$. It is then reasonable to ask the question whether there is any point in adding the small coefficient to the model. To find the answer to this question, one can use the following statistical test.

3.3 TESTING SIGNIFICANCE OF REGRESSION COEFFICIENTS

The structure of the test is as follows:

1. The tested hypothesis is:

$$H_0: \beta_i = 0,$$
$$H_1: \beta_i \neq 0.$$

2. The test criterion is

$$T = \frac{b_i}{s(b_i)},$$

where b_i is the estimate of β_i, $s(b_i)$ is the estimated standard deviation of b_i.

3. The critical value $K = t_{n-k}(\alpha)$ for a nivel of test α.

4. If $|T| > K$, we reject the hypothesis H_0 and accept the alternative hypothesis H_1 according to which β_i is considered to be nonzero or statistically significant. In the opposite case, the null hypothesis is accepted, and the tested parameter is thought to be equal to zero or statistically insignificant.

In our case, we have

$$T_1 = \frac{b_1}{s(b_1)} = \frac{0{,}60}{\sqrt{0{,}0514}} = 2{,}65,$$

$$T_2 = \frac{b_2}{s(b_2)} = 4{,}58,$$

where $t_{n-k}(\alpha) = t_{10-3}\ (0{,}05) = 2{,}365$. Since $T_1 > 2{,}365$ and also $T_2 > 2{,}365$, both population coefficients are statistically significant, and should be considered in the model.

As was said before, one needs to have a sample of points ($x_{j1}, x_{j2},\ldots, x_{jk}, y_j$) to find the vector of coefficients $\vec{b} = (b_1,...,b_k)$, the estimate of the coefficients $\vec{\beta} = (\beta_1,...,\beta_k)$. In other words, the vector of observations Y and the matrix of regressors X must be known.

In practice, tha analyst has to decide how to select the values x_{ij} and how many of them should be included in the model. Answers to these questions have a major impact on the final estimates of the coefficients because the subsequent calculations are only a routine procedure. There is a statistical discipline which, among other things, aims to answer these questions. The discipline is called *design of experiments*. We shall analyse some fundamental designs of experiments in later chapters of this textbook.

3.4 CONFIDENCE INTERVALS FOR REGRESSION COEFFICIENTS

Confidence intervals for the parameters $\beta_1,\ldots,\beta_k$ are intervals in which the parameters lie with probability 1-α. Each of these intervals is of the form

3-7 $$[b_i - t_{n-p}(\alpha).s(b_i), b_i + t_{n-p}(\alpha).s(b_i)],$$

where

b_i = estimate of β_i,
$s(b_i)$ = estimated standard deviation of b_i,
$t_{n-p}(\alpha)$ = critical value of a Student's distribution,
n = sample size (number of points),
p = number of parameters in the model,
α = nivel of test,

The unknown parameter β_i lies in the interval 3-7 with probability 1-α. It is necessary that the random term ε of the regression model is normally distributed for the interval to hold true.

3.5 TESTING MODEL SIGNIFICANCE

Significance of the regression model can be verified by the following test, which again requires that the random term of the model is normally distributed. The structure of the test is as follows:

1. The hypothesis is:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0,$$

or, using vectors, $H_0 : \vec{\beta} = \vec{0}.$

$$H_1 : \vec{\beta} \neq \vec{0}.$$

2. The test criterion: $$T = \frac{S_{\hat{Y}}/(k)}{S_e/(n-k-1)},$$

where

$$S_{\hat{Y}} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2, \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

$$S_e = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2.$$

3. The critical value is $K = F_{k,n-k-1}(\alpha)$, where $F_{k,n-k-1}(\alpha)$ is the critical value of a Fisher's distribution with $df1 = k$ a $df_2 = n\text{-}k\text{-}1$ degrees of freedom. In Excel, one can obtain the value with the function FINV(α; df_1; df_2).

4. If $T \geq K$, H_0 is rejected. In the opposite case, H_0 is accepted.

In our example, we have:

$$T = \frac{0{,}594/(2)}{0{,}006/(10-3)} = 346{,}5, \quad K = F_{2,7}(0{,}05) = 4{,}73.$$

Since T exceeds the critical value K, H_0 is rejected and the model is viewed as satisfactory, i.e. we reject the hypothesis that all regression coefficients except the constant term β_0 are zeros.

SUMMARY

This chapter dealt with regression analysis which aims to find a relation between a quantitative variable Y, or an explained/dependent variable, and other quantitative variables called explanatory/independent variables. In particular, we were concerned with linear regression models (linear in parameters). At the end of the chapter, different statistical tests were presented, regarding statistical significance of the regression coefficients and the model as a whole, as well as confidence intervals for the estimates of the coefficients.

The text was accompanied by examples. The following terms were explained: linear regression, estimates of regression coefficients, theoretical value, residual, variance and standard deviation of regression coefficients, test of regression coefficients, test of the model, confidence intervals for regression coefficients.

Some more examples follow.

PROBLEM 2

a) Estimate the regression coefficients of the model $\hat{Y} = b_0 + b_1 x_1 + b_2 x_2$,
b) Calculate the theoretical values of Y,
c) Calculate the residuals of the model,
d) Calculate the variance of the coefficient estimates,
e) Test significance of the coefficients.
f) For the following entry values $X_0 = \begin{bmatrix} 1 & 11 & 3 \\ 1 & 12 & 5 \end{bmatrix}$, predict the dependent variable Y_0.

Perform the tasks a)-f), using the following data:

y	x_1	x_2
10	1	0
25	3	-1
32	4	0
43	5	1
58	7	-1
62	8	0
67	10	-1
71	10	2

Source: author's

SOLUTION

a. The estimates of the coefficients:

$$X^T Y = \begin{bmatrix} 368 \\ 2710 \\ 35 \end{bmatrix},$$

$$\vec{b}^T = \frac{1}{4664}\begin{bmatrix} 2887 & -384 & 240 \\ -384 & 64 & -40 \\ -240 & -40 & 608 \end{bmatrix}.\begin{bmatrix} 368 \\ 2710 \\ 35 \end{bmatrix} = \begin{bmatrix} 6{,}47 \\ 6{,}59 \\ 0{,}26 \end{bmatrix}.$$

b. The theoretical values:

$$\hat{Y} = (13{,}06\,,\, 25{,}98\,,\, 32{,}83\,,\, 39{,}68\,,\, 52{,}34\,,\, 59{,}19\,,\, 72{,}11\,,\, 72{,}89)\,.$$

c. The residuals:

$$\vec{e} = (-3{,}06\,,\, -0{,}98\,,\, -0{,}83\,,\, 3{,}32\,,\, 5{,}66\,,\, 2{,}81\,,\, -5{,}11\,,\, -1{,}89)\,.$$

d. The variances of the coefficient estimates:

Since $\sum_i e_i^2 = 91{,}65$,

$$Var(\vec{b}) = \frac{91{,}65}{8-3}.\frac{1}{4664}\begin{bmatrix} 2887 & -384 & 240 \\ -384 & 64 & -40 \\ 240 & -40 & 608 \end{bmatrix} = \begin{bmatrix} 11{,}35 & \ldots & \ldots \\ \ldots & 0{,}25 & \ldots \\ \ldots & \ldots & 2{,}39 \end{bmatrix}.$$

e. The test of coefficients:

$$T_0 = \frac{b_0}{s(b_0)} = \frac{6{,}47}{3{,}37} = 1{,}92\ ,\ T_1 = \frac{6{,}59}{0{,}5} = 13{,}18\ ,\ T_2 = 0{,}17\ .$$

$$K = t_{n-p}(\alpha) = t_{8-3}(0{,}05) = 2{,}571.$$

Statistical significance pertains only to β_1 because $T_1 > K$.

f. Two predictions of Y_0 (at two different points of X):

Since $X_0 = \begin{bmatrix} 1 & 11 & 3 \\ 1 & 12 & 5 \end{bmatrix}$,

$$Y_0 = \begin{bmatrix} 1 & 11 & 3 \\ 1 & 12 & 5 \end{bmatrix} \cdot \begin{bmatrix} 6{,}47 \\ 6{,}59 \\ 0{,}26 \end{bmatrix} = \begin{bmatrix} 79{,}74 \\ 86{,}85 \end{bmatrix}.$$

PROBLEM 3

Find out whether production depends on corporate investments. A potential dependence is reflected by the parameter β_1 in the regression function with two unknown parameters. We know, based on twelve data, that the estimate of β_1 is $b_1 = 2{,}1622$. We also know the standard deviation of the estimate is $s(b_1) = 0{,}615516$. Verify or reject the existence of the dependence by testing the hypothesis: $\beta_1 = 0$. The nivel of test is five per cent.

SOLUTION

Since

$$T = \frac{b_1}{s(b_1)},$$

we get

$$T = \frac{2{,}1622}{0{,}615516} = 3{,}513.$$

The appropriate critical value, related to a Student's distribution with 12 – 2 = 10 degrees of freedom, is $t_{10}(0{,}05) = 2{,}228$. Since 3,513 > 2,228, we reject the null hypothesis about the zero value of the coefficient. Thus, the coefficient is statistically significant, and the dependence may be assumed to exist.

PROBLEM 4 (HOTEL SERVICES)

Find the linear regression model which describes a dependence of total monthly revenues Y (in tens of thousands of crowns) of a hotel on revenues X_1 generated by the catering services

of the hotel (in tens of thousands of crowns) and on X_2 which is a product of the number of beds at the hotel and the number of days in a given month. The data is in the following table.

Data

Y	x_1	x_2
12,0	2,0	150
8,0	1,2	94
76,4	14,8	811
17,0	8,3	254
21,3	8,4	399
10,0	3,0	95
12,5	4,8	149
97,3	15,6	312
88,0	16,1	952
25,0	11,5	247
38,6	14,2	400
47,3	14,0	312

Source: author's

SOLUTION

The vectors for the dependent and independent variables have the following form:

$$Y = \begin{bmatrix} 12,0 \\ 8,0 \\ 76,4 \\ 17,0 \\ 21,3 \\ 10,0 \\ 12,5 \\ 97,3 \\ 88,0 \\ 25,0 \\ 38,6 \\ 47,3 \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & 2,0 & 150 \\ 1 & 1,2 & 94 \\ 1 & 14,8 & 811 \\ 1 & 8,3 & 254 \\ 1 & 8,4 & 399 \\ 1 & 3,0 & 95 \\ 1 & 4,8 & 149 \\ 1 & 15,6 & 312 \\ 1 & 16,1 & 952 \\ 1 & 11,5 & 247 \\ 1 & 14,2 & 400 \\ 1 & 14,0 & 312 \end{bmatrix}.$$

Therefore

$$X^T X = \begin{bmatrix} 12,0 & 113,90 & 4175,0 \\ 113,9 & 1428,43 & 51958,5 \\ 4175,0 & 51958,50 & 2266001 \end{bmatrix}, \quad X^T Y = \begin{bmatrix} 453,4 \\ 6006,8 \\ 230647,8 \end{bmatrix},$$

$$\left(X^T X\right)^{-1} = \begin{bmatrix} 0,343 & -0,02629 & -0,0003 \\ -0,026 & 0,006234 & -0,000094 \\ -0,00003 & -0,000094 & 0,00000266 \end{bmatrix},$$

$$\left(X^T X\right)^{-1} \cdot \left(X^T Y\right) = \vec{b}^T = \begin{bmatrix} -9,126450 \\ 3,729273 \\ 0,033091 \end{bmatrix}.$$

The regression function is

$$\hat{Y} = -9,126450 + 3,729273 x_1 + 0,033091 x_2.$$

PROBLEM 5

Test statistical significance of the coefficients from the previous example: β_1, β_2. The nivel of test is five per cent.

SOLUTION

First of all, the standard deviations of the coefficients are calculated. Their values are:

$$s(b_1) = 1,371, \quad s(b_2) = 0,0283.$$

The corresponding test criterion is $T = \dfrac{b_j}{s(b_j)}$, and so

$$T_1 = \frac{3,729273}{1,371} = 2,7201, \qquad T_2 = \frac{0,033091}{0,0283} = 1,1693.$$

The critical value of the test is found for a Student's distribution with 12 – 3 = 9 degrees of freedom and the five per cent nivel of test: $t_9(0,05) = 2,26$. Comparing the test criterions with the critical value, we see that the null hypothesis concerning the zero value of the coefficient β_1 is rejected. The case of the other parameter is different, however. The parameter β_2 seems to be insignificant and the second variable should be removed from the model.

CONTROL TEST 3

Yes/No answers:

3.1 Regression analysis examines dependence among quantitative variables?

3.2 Deviation of an empirical value of *Y* from its theoretical value, modelled by a regression function, is called residual?

3.3 Regression analysis deals only with linear functions?

3.4 The test of significance of regression coefficients requires the critical value of a normal distribution?

3.5 The null hypothesis in the test of model significance is: $H_0 : \beta_1 \neq \beta_2 \neq ... \neq \beta_k \neq 0$?

3.6 The classical regression model assumes that random terms in the model have _____ expected value and _____ variance.

3.7 The test examining the zero value of an individual regression coefficient is called _____

3.8 If the model $\hat{Y} = b_0 + b_1 x_1 + ... + b_k x_k$ contains the term b_0, the first column of the matrix of regressors *X* consists of value(s) _____

3.9 Regression analysis exploits dependence among_____ variables.

3.10 Variances of estimated regression coefficients can be found on _____ _____ of the matrix $Var(\vec{b}) = s^2 (X^T X)^{-1}$.

3.11 A personnel department gathered the following data on age (*X*) of 20 randomly selected employees and the amount of time (*Y*) they spent out of work due to health reasons.

x	***y***	***x***	***y***
20	4	58	20
35	14	46	13
35	15	43	16
34	10	33	10
32	10	29	10
28	9	36	11
25	12	48	14
46	15	55	15
38	15	36	14
50	16	19	6

Source: author's

Estimate regression coefficients of the model $\hat{Y} = b_0 + b_1 x$.

3.12 A statistical office examined dependence of yearly savings on yearly income. Both variables are related to families with two children. The result of the survey is in the following table.

Income (thousands of crowns)	104	125	146	167	111	135	189	196	205	210	170	230
Savings (thousands of crowns)	6	5,6	9,2	14	8	9,1	20,5	29	23,2	38,5	25	40

Source: author's

Find the linear regression model that explains a dependence of savings on income. Using the model, estimate the savings of a family whose yearly income is 205 thousand crowns.

3.13 Eight families were randomly selected from national accounting records. Their gross yearly income (= explanatory variable *x*, measured in crowns) and their yearly expenses on industrial products (= explained variable *Y*, measured in crowns) were analysed. The results are in the table.

x	211399	306502	250251	264138	274060	297046	328645	249987
Y	42276	72341	49852	53827	54914	60409	71729	47997

Source: author's

a. Estimate the linear regression function which describes a dependence of expenses on income.
b. Calculate the theoretical expenses of a family with income exceeding 300 thousand crowns.

3.14

a. Using the data of problem 3.13, calculate the standard deviation of the estimates $b_i,\ i = 0, 1$.
b. Using the data of problem 3.13, calculate the test criterion used to test the statistical insignificance of b_1.

3.15 The following data is available on the production of France (= variable *Y* in millions of euros), its amount of fixed capital (= variable X_1 in millions of euros) and employment (= variable X_2 in thousands of people). Estimate regression coefficients of the model $f(X_1, X_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where $Y = f(X_1, X_2) + \varepsilon$.

Economic sector	Y_i	x_{i1}	x_{i2}
Agriculture	288443	18781	1055
Food and drinks	393828	13990	551
Power	330300	33813	223
Semi-products	602182	32022	1101
Production equipment	426720	19520	965
Household goods	34008	1258	49
Transportation means	185887	10462	358
Consumer products	427766	16392	1030
Construction	436926	19828	1472
Trade	495319	36354	2691
Transportation	417147	58196	1268
Market services	1002132	116083	4617
Insurance services	61827	2053	158
Financial services	709297	6908	441
Non-market services	840622	136923	6148

Source: author's

SOLUTIONS

3.1 yes
3.2 yes
3.3 no
3.4 no
3.5 no
3.6 zero, constant
3.7 *t*-test
3.8 ones
3.9 quantitative
3.10 the main diagonal
3.11 $\widehat{Y} = 1,394 + 0,296x$,
3.12 $\widehat{Y} = -26,399 + 0,274x$; 29 711 crowns
3.13

a. $\widehat{Y} = -19599,4 + 0,2796x,$

b. for $x_i = 300000\,crowns,$ we get $\widehat{Y}_i = 64298$ crowns. At least 64298 is the answer.

3.14

a. $s(b_1) = 0,03375,$

b. The test criterion T = 8,284, the critical value $t_6(0,05) = 2,447 \Rightarrow$ we accept the hypothesis H_1 on dependence of expenses on income.

3.15 $\widehat{Y} = 263684,7 + 2,2331x_1 + 66,7912x_2.$

4 CORRELATION ANALYSIS

In the previous chapter, we were solving the problem of finding a functional relation which would describe dependence of one variable *Y* on other explanatory variables represented by a vector *X*. Mathematically speaking, the relations were linear in parameters. In this chapter, we will be preoccupied with the problem of measuring intensity of dependence among variables. There is more than one way how to do it. Perhaps the simplest way of measuring dependence has to do with what is called *correlation analysis*.

Correlation analysis is closely related to regression [2], as it benefits from the theory of linear regression models. The objective of correlation is different, however. It does not seek a reasonable form of relations among variables because it a priori assumes that relations are linear in parameters and even in variables, as well. Instead, it focuses on the construction of measures of dependence among the variables.

This chapter is accompanied by examples to give a better understanding of the subject. After studying correlation, the reader is advised to calculate the problems at the end of the chapter.

4.1 CORRELATION COEFFICIENT

In the simplest form, we study dependence between two random variables *Y* and *X*. If this is the case, *paired correlation coefficient* ρ_{xy} is used to measure the level of linear dependence between the two variables. The coefficient is defined as

4-1
$$\rho_{xy} = \frac{Cov(X,Y)}{\sigma(X).\sigma(Y)} \text{ for } \sigma(X) > 0, \sigma(Y) > 0,$$

$$= 0 \text{ otherwise.}$$

Here, $Cov(X,Y) = E(XY) - E(X) \cdot E(Y)$ is the covariance of the random variables *X* and *Y*, the characteristic having been defined in chapter one. Also, $\sigma(X)$ and $\sigma(Y)$ are the standard deviations of *X* and *Y*, respectively. The symbol *E* stands for the expected value of a random variable. Expected value was explained in the course Statistics. The paired correlation coefficient is an element of the closed interval $[-1,1]$, i.e. $\rho_{xy} \in [-1,1]$. If $\rho_{xy} = 0$, we say that the variables *X* and *Y* are uncorrelated. If $\rho_{xy} = 1$ or $\rho_{xy} = -1$, an exact functional relation exists between *X* and *Y*, the function being a line. If $\rho_{xy} = 1$, the line is increasing. If $\rho_{xy} = -1$, the line is decreasing.

If $\rho_{xy} = 0$, we can only conclude that the variables are uncorrelated. We cannot say that they are (statistically) independent. Although independent variables are uncorrelated, the opposite statement is not generally true.

PROBLEM 1

Let us calculate the correlation coefficient ρ_{xy} for the data in table 15:

Table 15: Entry data for correlation analysis

X	-2	-1	0	1	2
Y	4	1	0	1	4

Source: author's

All the pairs occur with the same probability p. The following table 16 adds some preparatory calculations to make things easier.

Table 16: Preliminary calculations

X_i	Y_i	$X_i.Y_i$
-2	4	-8
-1	1	-1
0	0	0
1	1	1
2	4	8
$\Sigma x_i = 0$	$\Sigma y_i = 10$	$\Sigma x_i.y_i = 0$

We have $Cov(X,Y) = E(XY) - E(X) \cdot E(Y) = p\sum_i x_i y_i - p^2 \sum_i x_i \sum_i y_i = 0$, and thus $\rho_{xy} = 0$.

At the same time, it can be seen that the two variables are not independent – on the contrary, they are even perfectly dependent, one of the variables being the second power of the other.

Formula 4-1 defines the population/theoretical correlation coefficient, which in most cases cannot be calculated since the population characteristics $Cov(X,Y)$, $\sigma(X)$ and $\sigma(Y)$ will most likely be unknown. The presented example is devised artifically, of course. For these reasons, the population coefficient is in practice usually estimated by its sample version r_{xy}, which is calculated from a data sample. The sample correlation coefficient is defined by

4-2
$$r_{xy} = \frac{n.\sum x_i.y_i - \sum x_i.\sum y_i}{\sqrt{\left[n.\sum x_i^2 - (\sum x_i)^2\right]\left[n.\sum y_i^2 - (\sum y_i)^2\right]}}.$$

To decide reasonably whether there is any amount of linear dependence between Y and X, the correlation is tested for significance. The null hypothesis of the test states that $\rho_{xy} = 0$, and the alternative hypothesis says the opposite – there exists a nonzero correlation. To perform the test, the sample correlation coefficient is used.

Testing zero value of paired population correlation

1. The null hypothesis is H_0: $\rho_{xy} = 0$ vs. the alternative H_1: $\rho_{xy} \neq 0$.

2. The test criterion is of the form:

$$T = \frac{r_{xy}.\sqrt{n-2}}{\sqrt{1-r_{xy}^2}},$$

where n = number of sample pairs (x_i, y_i).

3. The critical value of the test for a nivel of test alpha is $K = t_{n-2}(\alpha)$. Thus, it concerns a Student's distribution with n-2 degrees of freedom.

4. If $|T| < K$, H_0 is accepted, i.e. Y is not linearly dependent on X. In the opposite case, H_1 is accepted, which means that Y is (at least to a certain extent)) linearly dependent on X.

Let us note that there are conditions that must be satisfied for the test to be valid: the main condition is that the pairs were drawn from a two-dimensional normal distribution.

PROBLEM 2

Let us have the following sample pairs made up of values x_i and y_i (the first two columns of table 17):

Table 17: Result of a random sampling, preparatory calculations

x_i	y_i	$x_i\, y_i$	x_i^2	y_i^2
-2	-5	10	4	25
-1	-3	3	1	9
0	0	0	0	0
1	1	1	1	1
2	4	8	4	16
Σ = 0	Σ = -3	Σ = 22	Σ = 10	Σ = 51

Source: author's

Using 4-2, we obtain

$$r_{xy} = \frac{5.22 - 0.(-3)}{\sqrt{(5.10-0).(5.51-(-3)^2)}} = 0,9918.$$

This value is a clear signal that there probably is a linear relationship between Y and X. However, the sample size is very small, and so we prefer to test the (in)significance of the correlation, using a one per cent nivel of test.

$$T = \frac{0,9918.\sqrt{5-2}}{\sqrt{1-0,9918^2}} = \frac{1,718}{\sqrt{0,016}} = 13,443.$$

Since $\alpha = 0{,}01$, the critical value is $K = t_{5\text{-}2}(0{,}01) = \text{TINV}(0{,}01;\ 3) = 5{,}84$. As $T > K$, there is a significant (nonzero) linear dependence of Y on X. This is also confirmed by the p-value of the test, which is TDIST(13,443; 3; 2) = 0,00089. The value is substantially smaller than the nivel of test, suggesting that the correlation is significant for nivels of test 0,00089 and higher, i.e. for all reasonable nivels of test.

4.2 CORRELATION INDEX

If the regression function, based on which the correlation of two variables is assessed, isn't linear, it is possible to measure a dependence of two variables with *correlation index*:

4-3
$$I_{xy} = \sqrt{\frac{S_{\hat{Y}}}{S_Y}},$$

where

$$S_{\hat{Y}} = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2,$$

$$S_Y = \sum_{i=1}^{n} (Y_i - \overline{Y})^2.$$

The symbols were already used in regression analysis. The calculation of I_{xy} is more laborous than that of r_{xy} because the regression function has to be found first, so that the theoretical values of the dependent variable $\widehat{Y}_i$ are available. The values Y_i are measured, and their average is $\overline{Y}$. The theoretical values $\widehat{Y}_i$ correspond to the values of the explanatory variable X appearing in the regression, as was described in the chapter on regression analysis:

$$\widehat{Y}_i = f(x_i).$$

The index satisfies $0 \le I_{xy} \le 1$. Discussions about the resulting value of the index are similar to those about r_{xy}, testing its significance is not performed, however. I_{xy} can also be used for the case of a regression line. It is then identical to the absolute value of the paired correlation coefficient r_{xy}.

4.3 SPEARMAN'S RANK CORRELATION COEFFICIENT

If ranks of two variables X, Y are known, not their original values, *Spearman's rank correlation coefficient* r_s is used instead to measure dependence of the variables:

4-4
$$r_S = 1 - \frac{6 \cdot \sum_i d_i^2}{n(n^2 - 1)}.$$

Here, d_i is the difference of the i-th ranks of X and Y, and n is the number of numerical pairs of the two variables that are available through a data sampling.

PROBLEM 3

Products were sorted by their quality, the sorting having been implemented by two committees: specialists were on one committee, laymen selected from the general public on another committee. Determine whether the resulting assessments of the product quality depends on what committee is considered for this purpose. Here, the dependence is understood in the sense of a correlation. Entry data are in table 18, together with the differences in the rankings.

Table 18: product rankings

Product	Ranks by laymen	Ranks by specialists	d_i	d_i^2
1	7	8	-1	1
2	9	9	0	0
3	8	7	1	1
4	10	10	0	0
5	6	6	0	0
6	5	4	1	1
7	3	5	-2	4
8	4	3	1	1
9	2	2	0	0
10	1	1	0	0

Source: author's

$$r_S = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)} = 1 - \frac{6.8}{10.99} = 0,95.$$

The coefficient can be used to test statistical independence (not only correlation !) of the two variables:

1. The tested hypothesis is H_0: *X*, *Y* are independent vs. H_1: *X*, *Y* are not independent.

2. The test criterion is:

$$T = \sqrt{(n-1)} \cdot r_S .$$

3. The critical value *K* is that of the standard normal distribution *N*(0,1). For a nivel of test alpha, the value is calculated with the Excel function NORMSINV(1-alpha/2).

4. If $|T| \geq K$, we reject H_0. In the opposite case, we accept H_0.

If H_0 is accepted, we know the variables are independent, and thus uncorrelated as well. If the null hypothesis is rejected, we know the variables are not independent, but we don't know if they are uncorrelated or not. The test is approximately valid provided that $n \geq 30$ and the random vector (*X*,*Y*) follows a two-dimensional continuous probability distribution.

4.4 MULTIVARIATE DEPENDENCE- THE CASE OF TWO VARIABLES

If we want to examine the linear dependence of a variable *Y* on variables $X_1, X_2, ..., X_p$, *p*>1, we use either:

a. coefficients of partial correlation,
b. coefficients of multivariate correlation.

Ad a. *Partial correlation coefficient* $r_{yx_1 \bullet x_2,...,x_p}$ measures the intensity of the linear dependence of *Y* on X_1 provided a certain effect of variables $X_2, ..., X_p$ is removed. These are the variables listed behind the symbol „•". Partial correlation tries to solve the problem that the effect of X_1 might be distorted by the contemporaneous effects of variables $X_2, ..., X_p$. We shall restrict our analysis to the case of $p = 2$.

The coefficient of partial correlation appears again in two forms – as a population coefficient and a sample coefficient. In the latter case, and for $p = 2$, the sample version is calculated as

$$r_{yx_1 \bullet x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1 x_2}}{\sqrt{\left(1 - r_{yx_2}^2\right)\left(1 - r_{x_1 x_2}^2\right)}},$$

or

$$r_{yx_2 \bullet x_1} = \frac{r_{yx_2} - r_{yx_1} r_{x_1 x_2}}{\sqrt{\left(1 - r_{yx_1}^2\right)\left(1 - r_{x_1 x_2}^2\right)}}.$$

In both cases, the coefficients take on a value from interval [-1,1]. As can be seen from 4-5 and 4-6, to calculate the partial correlation, it is necessary to evaluate various combinations of paired correlations. Since we talk about sample partial correlation, it is possible to test the significance of its population version.

Testing statistical significance of partial correlation (p = 2):

1. $H_0: \ \rho_{yx_1 \bullet x_2} = 0$, $H_1: \ \rho_{yx_1 \bullet x_2} \neq 0$.

2. The test criterion is:

$$T = \frac{r_{yx_1 \bullet x_2} \sqrt{n-3}}{\sqrt{1 - r_{yx_1 \bullet x_2}^2}}.$$

3. The critical value for a nivel of test alpha is $K = t_{n-3}(\alpha)$ = TINV(α,*n-3*).

4. IF $|T| \geq t_{n-3}(\alpha)$, the coefficient of partial correlation is significant, i.e. nonzero.

The test is valid provided the random vector (Y, X_1, X_2) follows a three-dimensional normal distribution. It is also assumed that $n > 3$.

Ad b. *Coefficient of multiple correlation* measures dependence of a variable Y on all explanatory variables $X_1, X_2, ..., X_p$. If two explanatory variables are considered, the sample version of the coefficient satisfies

$$r_{y\bullet x_1x_2} = \sqrt{\frac{r_{yx_1}^2 - 2r_{yx_1}r_{yx_2}r_{x_1x_2} + r_{yx_2}^2}{1 - r_{x_1x_2}^2}}, \; 0 \le r_{y\bullet x_1x_2} \le 1.$$

The significance of the coefficient can also be tested:

Testing statistical significance of multiple correlation:

1. $H_0: \; \rho_{y\bullet x_1x_2} = 0$ vs. $H_1: \; \rho_{y\bullet x_1x_2} \neq 0$.

2. The test criterion:

$$T = \frac{r^2_{y\bullet x_1x_2} \cdot (n-3)}{2 \cdot \left(1 - r^2_{y\bullet x_1x_2}\right)} \; .$$

3. The critical value of the test is related to a Fisher's distribution this time. The distribution has 2 and n-3 degrees of freedom: the critical value is written as $F_{2,n-3}(\alpha)$ for a nivel of test alpha. In Excel, it can be obtained with the function FINV(α,2,n-3).

4. If $T \ge F_{2,n-3}(\alpha)$, the coefficient of multiple correlation is statistically significant (the null hypothesis is rejected). In the opposite case, it is statistically insignificant (the null hypothesis is accepted).

The test is valid on condition that the random vector (Y, X_1, X_2) follows a three-dimensional normal distribution. Is is also assumed that $n > 3$.

SUMMARY

In this chapter, we became familiarized with another important statistical term: correlation analysis. We learnt how to compute the correlation coefficient, correlation index and the Spearman's rank correlation coefficient. The end of the chapter discussed partial correlation and multiple correlation. The theory and examples worked with the case when two explanatory variables are considered because the calculations become more cumbersome if more explanatory variables are considered.

The text is now followed by more examples.

PROBLEM 4

Check if there is any linear dependence among the following variables. Calculate the partial correlation coefficients and test the smaller one on significance. The nivel of test is 5%. Also, calculate the coefficient of multiple correlation and test its significance for 5% nivel of test. Use table 19 for the calculations. The dependent variable is *Y*, the other variables are explanatory.

Table 19: Entry data for problem 4

Y	12	8	76,4	17	21,3	10
X_1	2	1,2	14,8	8,3	8,4	3
X_2	150	94	811	254	399	95

Y	12,5	97,3	88	25	38,6	47,3
X_1	4,8	15,6	16,1	11,5	14,2	14
X_2	149	312	952	247	400	312

Source: author's

SOLUTION

The paired correlations are as follows:

$$r_{yx_1} = 0,85, \quad r_{yx_2} = 0,75, \quad r_{x_1x_2} = 0,73.$$

Inserting them in equations

$$r_{yx_1 \bullet x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_2}^2\right)\left(1 - r_{x_1x_2}^2\right)}},$$

$$r_{yx_2 \bullet x_1} = \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_1}^2\right)\left(1 - r_{x_1x_2}^2\right)}},$$

we get the partial correlations:

$$r_{yx_1 \bullet x_2} = 0,67\,, \; r_{yx_2 \bullet x_1} = 0,36\,.$$

Testing of $r_{yx_2 \bullet x_1} = 0,36$:

1. $H_0: \; \rho_{yx_2 \bullet x_1} = 0$ vs. $H_1: \; \rho_{yx_2 \bullet x_1} \neq 0$.

2. The test criterion: $T = \dfrac{r_{yx_2 \bullet x_1}\sqrt{12-3}}{\sqrt{1 - r_{yx_2 \bullet x_1}^2}} = 1,16.$

3. The critical value: $t_{12-3}\left(0,05\right) = 2,262$.

4. Since $|T| < t_9(0,05)$, we accept the null hypothesis anad conclude the partial correlation is insignificant.

In the end, we evaluate the multiple correlation coefficient and test its significance. Using equation 4-7, we get:

$$r_{y\bullet x_1x_2} = \sqrt{\frac{r_{yx_1}^2 - 2r_{yx_1}r_{yx_2}r_{x_1x_2} + r_{yx_2}^2}{1 - r_{x_1x_2}^2}} = 0,87.$$

As we can see, the value is greater than all the paired correlation coefficients. Testing the significance, we have:

1. $H_0: \ \rho_{y\bullet x_1x_2} = 0$ (no linear dependence) vs. $H_1: \ \rho_{y\bullet x_1x_2} \neq 0$.

2. The test criterion: $T = \dfrac{r^2_{y\bullet x_1x_2}(12-3)}{2\cdot\left(1-r^2_{y\bullet x_1x_2}\right)} = 14,54.$

3. The critical value = FINV(0,05,2,9) = 4,26.

4. Since the test criterion falls in the critical region, we reject the null hypothesis, and we conclude that there is a combined linear effect of the *X*'s on *Y*.

PROBLEM 5

The sample coefficient of paired correlation $r_{xy} = 0,23$ has been calculated, based on a data sample of size n = 25. Verify for 1% nivel of test whether there is a linear indepdence between the variables X and Y in the population.

SOLUTION

The test criterion satisfies

$$T = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} = \frac{0,23\sqrt{25-2}}{\sqrt{1-0,23^2}} = 1,133.$$

The critical value of the test, found either in statistical tables or in Excel, is equal to

$$t_{23}(0,01) = 2,8.$$

Since 1,133 < 2,8, we cannot reject the null hypothesis, i.e. existence of linear dependence has not been proved.

PROBLEM 6

Canard Company has been monitoring a potential dependence of its operational costs per unit of production *Y* on total production *X* (in thousands of pieces).

Table 20: Costs *Y* and production *X* of Canard Company

x_i	60	71	92	144	192	306
y_i	5157	2620	1986	1582	1100	954

x_i	437	481	747	989	1383
y_i	729	456	200	196	110

Source: author's

Calculate the correlation index provided that a hyperbolic dependence of the form

$$Y = \frac{a}{X} + b + \varepsilon$$

is assumed. Here, *a* and *b* are unknown regression coefficients.

SOLUTION

The least squares method is used to estimate the coefficients *a* and *b*. The estimates $\hat{a}, \hat{b}$ are then inserted in the regression equation, and the theoretical values $\hat{y}_i = (\hat{a} / x_i) + \hat{b}$ are found for different values of x_i. Also, the average value $\bar{y}$, calculated from all the empirical values of the dependent variable *Y*, must be evaluated. Equation 4-3 then gives

$$I_{yx} = \sqrt{\frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}} = \sqrt{\frac{19813814}{22155242}} = 0{,}945.$$

The index suggests a high level of dependence of the costs on the production, which is not surprising, of course. However, what we are mainly interested in is that it is the hyperbolic equation that seems to describe the relation well, as suggested by the index.

CONTROL TEST 4

Yes/No answers:

4.1 Correlation coefficient measures a dependence of *Y* on *X*?
4.2 Correlation coefficient can take on any value from interval $[0,1]$?
4.3 The null hypothesis of the test that verifies significance of the correlation coefficient assumes that two variables are uncorrelated?
4.4 It is much simpler to calculate the index of correlation than the paired correlation coefficient?
4.5 Spearman's correlation coefficient can take on any value from interval $[-1,1]$?

Complete the statement:

4.6 Correlation analysis seeks a measure of _____

4.7 If $r_{xy} = 1$, then the line which describes the relation is _____

4.8 The correlation index can take on any value from interval _____

4.9 If values of variables *X, Y* represent ranks, then _____ correlation coefficient is used to describe a linear dependence between the two variables.

4.10 If *Y* is linearly dependent on $X = (X_1, X_2, ..., X_m)$, we use the coefficient of _____ _____to measure the dependence.

4.11 Calculate the paired correlation between coal extraction (in thousands of tons) and costs per ton of the extracted coal (in crowns). The available data are in the table.

Mine	x_i	y_i
1	350	37
2	351	38
3	329	38
4	329	38,5
5	327	37,5
6	322	39,1
7	321	39,6
8	316	42,1
9	298	42,9
10	286	43,5
$\sum$	3229	396,2

Source: author's

4.12 National accounts were used for a random selection of eight families. The accounts show the gross yearly income *X* of the families (in crowns) and their yearly expenses *Y* on consumer products. The data are in the table below. Calculate the correlation index for the case the linear dependence takes the form of a line, and calculate the paired correlation coefficient, as well.

x_i	211399	306502	250251	264138
y_i	42276	72341	49852	53827

x_i	274060	297046	328645	249987
y_i	59914	60409	71729	47997

Source: author's

4.13 Ten films were presented to the jury at a film festival, and viewers also participated in the rating of the movies. The final ranking is in the table.

Film	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
Rankings by jury	5	7	9	1	2	8	3	4	6	10
Rankings by viewers	1	6	4	3	8	7	2	5	10	9

Source: author's

Estimate the correlation between the rankings, using Spearman's correlation coefficient. Test significance of the coefficient for 5% nivel of test.

4.14 Calculate the coefficient of multiple correlation for the data in the next table. The data describes a dependence of output volume Y on fixed capital X_1 and employment X_2.

Economic sector	y_i	x_{i1}	x_{i2}
Agriculture	288443	18781	1055
Food and Drinks	393828	13990	551
Power sector	330300	33813	223
Semi-products	602182	32022	1101
Production equipment	426720	19520	965
Household goods	34008	1258	49
Transit means	185887	10462	358
Consumer goods	427766	16392	1030
Construction sector	436926	19828	1472

Source: author's

SOLUTIONS

4.1 yes
4.2 no
4.3 yes
4.4 no
4.5 yes
4.6 measures dependence
4.7 rising
4.8 [0,1]
4.9 Spearman's
4.10 Multiple correlation
4.11 $r_{yx} = -0,8967$
4.12 $r_{yx} = 0,9196 = I_{yx}$
4.13 Spearman's correlation $r_s = 0,38$, $T = 3,44$, $K = 1,64$. The coefficient is significant.
4.14 $r^2_{y\bullet x_1 x_2} = 0,6069.$

5 METHODS FOR SALES PREDICTIONS

Time series theory represents today a very important part of econometrics. The theory enables us to describe systems that change their behaviour in time. It is necessary to say that the dynamics of these systems deepens as the world globalization advances. National economy is a typical example of where the time series analysis is exploited. We might be interested in monthly movements of the aggregate price levels, published by the national statistical office, currency exchange rate closing quotes, etc. Time series, however, do not originate only in economy, but in other spheres of human activity, as well. For instance, birthrates and death rates are observed in demography, maximal nad minimal temperatures are recorded in meteorology, or blood pressure readings of a patient, as part of preventive check-ups, are documented by the doctor. The time series analysis aims to understand the mechanism which generated the series in question. Understanding the mechanism allows one, to an extent, to control the functioning of the system which generated the series, and thus to set a desired future course of the system by defining appropriately its input parameters. One may also use the insight into the mechanism for prediction of the future behaviour of the system. The system which created a time series is described by a mathematical model.

The time series theory is very extensive. It may as well be the most extensive branch of statistics, as some scholars note. In this chapter, we shall deal with the so-called classical time series theory. If we recall the general form of a regression model, the equation consisted of a systematic part and a random part. Whereas the systematic part reflected the systematic effect of the most important factors on the modelled variable, the random part represented the effect of all other and less important factors the separate influence of which is hard, if not impossible, to capture. Different approaches exist, regarding how to build a model that will approximate the mechanism of generating the time series. The classical approach focuses on the systematic part of the regression model.

The classical analysis assumes that the systematic part of the model can be decomposed into several elements of a specific type, which will shed more light on the origin of the series. It is understood that it is easier to detect these elements when they are separated rather than when their effect is aggregated. Another reason that stands behind the decomposition is the effort to discover potential seasonality, since it is a common practice to deprive the series of seasonality for different reasons.

In our case, to keep things simpler, we shall also assume that the mathematical model describing the time series containts only one explanatory variable *t* which represents a point in time.

5.1 TIME SERIES

A time series $\{y_t\}$ (TS) is a sequence of values that represent a realization of a sequence of random variables. We shall be mainly interested in **economic time series**, and in time development of sales, in particular.

It is usually assumed for a time series that:

- the main factor of change is the time *t*,
- equidistant time intervals, i.e. there is always the same time distance between any two neighbouring points in time *t* and *t'*, for which we get values y_t and $y_{t'}$ of the series.

A mathematical model is used to describe the time series. The main objective of building the model is to use it for prediction of the future values of the series. Two types of predictions are distinguished: point prediction and interval prediction.

5.2 TIME SERIES MODEL DECOMPOSITION

It is assumed that the model of a time series can be decomposed into four elements which describe different aspects of the time development of the analysed variable:

- trend component T_t,
- seasonal component S_t,
- cyclical component C_t,
- random component ε_t.

The trend component describes the fundamental character of the time development of the series. It tells us the essential character of its movement (whether it rises or drops, whether its level will eventually taper off or accelerate upwards, etc.). The trend expresses the systematic and long-term effect of factors that keep affecting the series in the same way. The trend is either rising or declining. If it's neither rising nor declining, it is a series with no trend.

The seasonal and cyclical components, which combined form the **periodic component,** capture regular oscilations of the series. The former relates to oscilations that take place within one year. These oscilations repeat every year, at the same moment. The oscilations can be attributed to natural phenomena (different seasons of the year – spring, summer, autumn, winter), or social habits (construction activities are more pronounced in the summer than in winter, for instance). The important feature of seasonality is that the regular oscilations of the same type take place every twelve months at the latest. **The cyclical component** represents the effect of factors that give rise to longer-term oscilations. We also talk about oscilations around the trend, and the time delay between two oscilations is more than twelve months. It is usually difficult to describe mathematically the cyclical component. Therefore, it is sometimes not included in the time series model at all. The difficulty is given by the fact that the cyclical oscilations are not as regular, and they often vary in their intensity.

The trend, seasonal and cyclical components form the **deterministic component** of the model. It is usually assumed that their combined effect is a result of their addition, i.e. the following model is assumed

5-1 $$Y_t = T_t + S_t + C_t + \varepsilon_t, \quad t = 0, \pm 1, \pm 2, \ldots .$$

In this case, we talk about the **additive model** of a time series. Two special cases of 5-1 are worth mentioning, as they often appear in economic applications: 1) the case without the periodic component, i.e. the case $S_t = C_t = 0$, so that we have

5-2 $$Y_t = T_t + \varepsilon_t, \quad t = 0, \pm 1, \pm 2, \ldots .$$

In the other case, $C_t = 0$ is assumed, which turns 5-1 into

5-3 $$y_t = T_t + S_t + \varepsilon_t, \quad t = 0, \pm 1, \pm 2, \ldots .$$

This is a time series model with seasonality.

Apart from the additive structure of 5-1, there is also a multiplicative version of the model:

5-4 $$y_t = T_t \cdot S_t \cdot C_t \cdot \varepsilon_t, \quad t = 0, \pm 1, \pm 2, \ldots .$$

The main objective of the time series analysis is to quantify the individual components of the time series model.

The stochastic process that generated the time series can also be analysed with other mathematical means, including the so-called adaptive approaches, such as moving averages and exponential smoothing. We shall talk about moving averages later.

The following modelling techniques focus on the practical aspect of working with time series, as is the case of other statistical methods in this course. The reader who is more interested in the time series analysis may take specialized courses that cover the subject in a greater detail.

5.2.1 TREND

As was mentioned already, time *t* is now assumed to be the only factor determining the dynamics of the analysed variable. The assumption, although largely simplifying the reality, makes it much more straightforward to model the time series under consideration and separate its individual components. One of these components – the trend – is the most important part the series.

Let us assume the model can be written as in 5-2. The trend T_t of this model is very often described by a linear function, polynomial of degree two, exponential function, modified exponential function, logistic curve or Gompertz's curve. The functions differ in their complexity, which further affects the way their parameters are estimated. If we work with a linear function or polynomial of degree two, both these function are linear in parameters, and thus their parameters can be estimated with the least squares method explained in the chapter on regression. The case of the other functions is different, since the mathematical description of the curves is not linear in parameters any more. Therefore, a different method must be used to find estimates of the unknown model parameters.

Linear trend (or polynomial of degree two)

Assuming a linear trend, the model 5-2 can be written as

5-5 $$Y_t = \beta_0 + \beta_1 t + \varepsilon_t, \quad t = 0, \pm 1, \pm 2, \ldots .$$

Estimates of the parameters β_0, β_1 are obtained with the least squares method.

PROBLEM 1

The following table 21 contains a time series data.

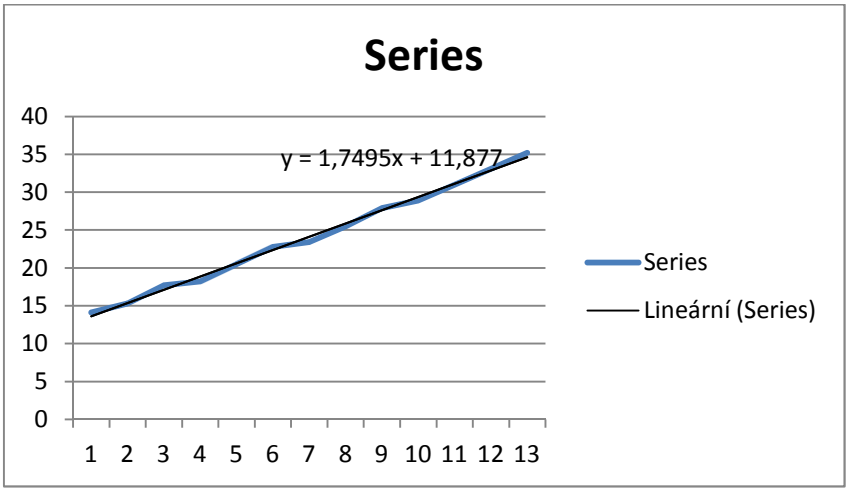
Table 21: A time series for problem 1

y_t	14,1	15,3	17,7	18,2	20,5	22,8	23,4	25,5	27,9	28,9	31	33,1	35,2
t	1	2	3	4	5	6	7	8	9	10	11	12	13

Source: author's

Excel: you can have the series depicted in Excel, together with the trend calculated by the least squares method. The output is in figure 2.

Figure 2: The time series and its trend in the form of a line



To get the figure in Excel (version 2010), the following steps must be taken: highlight the area of Excel containing the time series data *Yt*, *t* = 1, 2, …, *n*, and select

Insert → Graph → XY dot → with straight connecting lines.

This procedure will draw the graph of the original series. If it is necessary to convert the scale on the x-axis of the graph into values 1,2, …, *n*, click on the graph and press the right button of the computer mouse, choose „Select data" and „Adjust the x-axis", as proposed in the dialogue window. If you click on the graph again, you will see Graph tools at the top of Excel and Trend line in the corresponding window. Here, you may select the linear trend line and also its equation from the options.

One can proceed similarly in the case of polynomials of degree two, the equation of which is

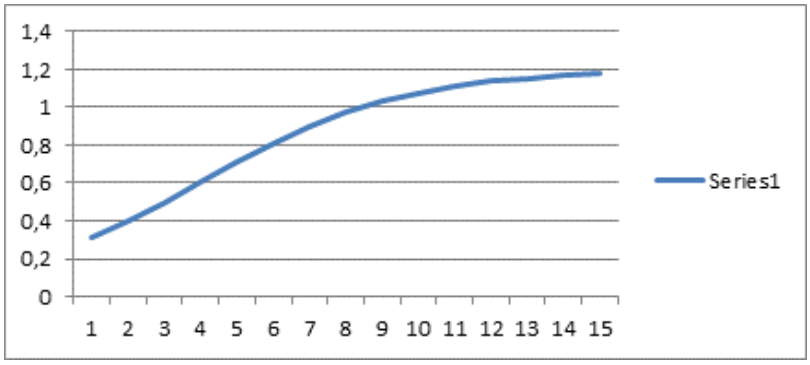
$$T_t = \beta_0 + \beta_1 t + \beta_2 t^2, \; t = 0, \pm 1, \pm 2, ...,$$

or in the case of polynomials of higher degrees.

Logistic trend

Logistic curves belong to the set of **S-curves**. This type of curves is often used in situations where a certain business cycle occurs. These cycles experience phases including the phase of saturation, which is what logistic curves capture well because they have a horizontal asymptote. For instance, when a new product is brought to the market, it is expected that it will take some time before customers register the product and try it out. Thus, at the early stages of the product marketing, the product sales volumes will go up slowly. Later, however, when the product comes into use, the sales growth curve will probably be steeper as more customers start to move from older versions of the product to the new version. When the sales volumes reach their climax, the product will dominate the market, as it becomes a fad. Later, competitors will catch up, and the sales volumes of the product will start to weaken. All these stages can be depicted by the logistic curve (figure 3).

Figure 3: The S-shape of a logistic trend



Logistic trend is given by equation

$$T_t = \frac{\kappa}{1+\beta_0\beta_1^t}, \quad \kappa > 0, \beta_1 > 0, \text{ t} = 1,2,\ldots .$$

The function has an S-curve shape for $\beta_1 < 1$, $\beta_0 > 1$. The unknown parameters of the trend can be estimated using a **method of selected points**: Let the length of the time series be T, where T is an odd number, and let us select chronologically the first, the middle (the p-th, say) and the last observation of the series. Then the parameter estimates satisfy

5-6

$$b_1 = \sqrt[p-1]{\frac{(1/y_p)-(1/y_T)}{(1/y_1)-(1/y_p)}},$$

$$k = \frac{y_1(1-b_1^{p-1})}{(y_1/y_p)-b_1^{p-1}},$$

$$b_0 = \frac{(1/y_1)-(1/y_p)}{b_1-b_1^p}\cdot k.$$

If T is even, it is necessary to select other values of the series: for example, the first, the p-th and the r-th such that r-p=p-1, and then formula 5-6 can be used again.

PROBLEM 2

Table 22 contains values of a time series. Describe the series with a logistic trend, using the method of selected points.

Table 22: Values of a time series to be captured by a logistic trend

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$\mathbf{y_t}$	0,4	0,6	0,6	0,7	0,7	0,8	0,9	0,9	1	1,1	1,2	1,2	1,8	1,4	1,6	1,7

t	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
$\mathbf{y_t}$	2,3	2,5	2,2	2,6	2,3	2,6	2,7	2,8	3,2	3	3,1	3,1	3,3	3,4	3,6

Source: author's

The method of selected points gives

$$b_1 = \sqrt[p-1]{\frac{(1/y_p)-(1/y_T)}{(1/y_1)-(1/y_p)}} = \sqrt[15]{\frac{(1/1{,}73)-(1/3{,}56)}{(1/0{,}35)-(1/1{,}73)}} = 0{,}872996836,$$

$$k = \frac{y_1(1-b_1^{p-1})}{(y_1/y_p)-b_1^{p-1}} = 4{,}2309685,$$

$$b_0 = \frac{(1/y_1)-(1/y_p)}{b_1-b_1^p}\cdot k = 12{,}70162866.$$

Thus, the final model of the series is

$$\hat{y}_t = \frac{4{,}23}{1+12{,}7\cdot 0{,}87^t}.$$

Table 23 contains the theoretical values of the series generated by the final model. This series and the original series are shown in figure 4 for visual comparison.

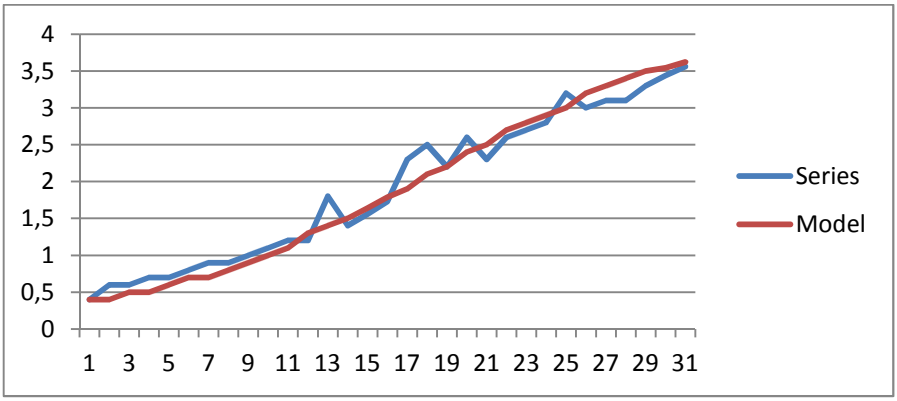
Table 23: The original and modelled (theoretical) time series

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$\mathbf{y_t}$	0,4	0,6	0,6	0,7	0,7	0,8	0,9	0,9	1	1,1	1,2	1,2	1,8	1,4	1,56	1,73
$\widehat{\boldsymbol{y}}_{\boldsymbol{t}}$	0,4	0,4	0,5	0,5	0,6	0,7	0,7	0,8	0,9	1	1,1	1,3	1,4	1,5	1,64	1,79

t	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
$\mathbf{y_t}$	2,3	2,5	2,2	2,6	2,3	2,6	2,7	2,8	3,2	3	3,1	3,1	3,3	3,44	3,56
$\widehat{\boldsymbol{y}}_{\boldsymbol{t}}$	1,9	2,1	2,2	2,4	2,5	2,7	2,8	2,9	3	3,2	3,3	3,4	3,5	3,54	3,62

Source: author's

Figure 4: The original series and its logistic model



There are more trends to choose from, of course, such as exponential trend, modified exponential trend, Gompertz's curve, etc. As we have said at the beginning of the chapter, the readers interested particularly in this subject, may take a specialized course devoted to time series. We note that a suitable trend can be selected, using differences of the original data: $\Delta^1 y_t = y_t - y_{t-1}$, $\Delta^2 y_t = \Delta^1 y_t - \Delta^1 y_{t-1}$. The selection rule based on the differences is as follows:

Table 24: Trend selection

Criterion	**Trend**
$\Delta^1 y_t \approx$ ***constant***	**Linear**
$\Delta^1 y_t \approx$ ***linear,*** $\Delta^2 y_t \approx$ ***constant***	**Quadratic**
$y_t - y_{t-1} \approx$ ***Gauss curve***	**Logistic**

5.2.2 SEASONAL COMPONENT – THE CASE OF CONSTANT SEASONALITY

Effect of seasonal factors can be described not only by a suitable moving average (to be discussed later), but also by a mathematical curve to which the principles of regression are applied. This concept is based on expanding the regression function, which contains the trend component already, by adding the seasonal component to it, the component depending on unknown parameters to be estimated, as well. As a result, more parameters of the regression function in question are to be estimated – some of these parameters relate to the trend component, while others belong to the seasonal component. The seasonal component is represented by auxiliary variables x_i, i = 2, 3, ..., s, where s is the number of seasons the model works with. Each variable x_i is a dichotomous variable, which means that it can only take on values 0 and 1. The variable x_i takes on value 1 for the i – th season and 0 otherwise (for all the other seasons). There are s-1 auxiliary variables in the model because otherwise, if the number was equal to s, perfect multicollinearity would exist, which is a problem that inhibits one from estimating the model parameters (the inverse of the matrix $X^T X$ ceases to exist). Therefore, the effect of one of the seasons is incorporated into the absolute term of the model.

The expanded model is of the form

$$y_t = T_t + \alpha_2 x_2 + \alpha_3 x_3 + \cdots + \alpha_s x_s + \varepsilon_t.$$

If a polynomial trend is assumed, we may write

$$y_t = \beta_0 + \beta_1 t + \ldots + \beta_k t^k + \alpha_2 x_2 + \cdots + \alpha_s x_s + \varepsilon_t,$$

where the absolute term β_0 contains the effect of the first season. **This representation of seasonality also assumes that the sum of all seasonal fluctuations equals zero**, i.e. the fluctuations cancel out. Also, independence of the seasonality on the trend is assumed as well as the additive decomposition of the entire regression model. Should there be a dependence between the trend and seasonality, the so-called proportional seasonality might be more convenient for this situation. Let us demonstrate now how to work with the model.

PROBLEM 3

Table 25 contains values of a time series that spans a four-year time period. These are quaterly data and, as suggested by figure 5, seasonality probably occurs in each quater of the year. We also adopt the view, based on the graph, that the trend component of our model could be linear. Therefore, we decided to use the model

$$y_t = \beta_0 + \beta_1 t + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4 + \varepsilon_t.$$

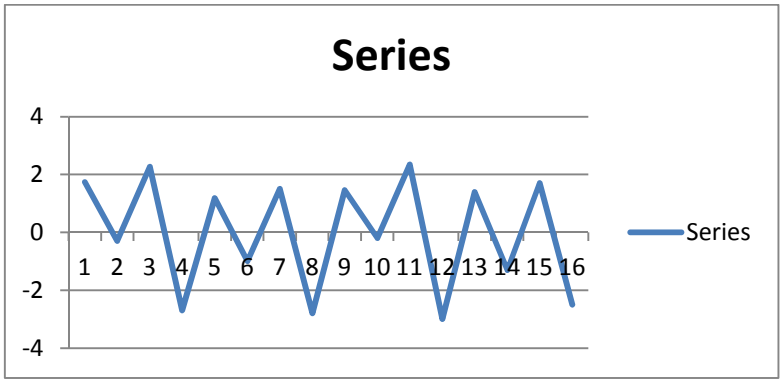
Table 25: Quaterly time series data

Season(i)	Sz(1)	Sz(2)	Sz(3)	Sz(4)	Sz(1)	Sz(2)	Sz(3)	Sz(4)
t	1	2	3	4	5	6	7	8
y_t	1,74	-0,3	2,27	-2,7	1,19	-1	1,51	-2,8

Season(i)	Sz(1)	Sz(2)	Sz(3)	Sz(4)	Sz(1)	Sz(2)	Sz(3)	Sz(4)
t	9	10	11	12	13	14	15	16
y_t	1,46	-0,2	2,35	-3	1,39	-1,3	1,7	-2,5

Source: author's

Figure 5: Time series data from table 25



Let us estimate the unknown coefficients of the model, using the least squares method (see the chapter on regression analysis). We get

$$\beta_0=1{,}605,\ \beta_1=-0{,}023,\ \alpha_2=-2{,}1,\ \alpha_3=0{,}56,\ \alpha_4=-4{,}13.$$

Thus,

$$\hat{Y}_t = 1{,}605 - 0{,}023t - 2{,}1x_2 + 0{,}56x_3 - 4{,}13x_4.$$

However, this is not where the calculation ends. We must realize that the value 1,605 includes the effect of the first-quater seasonality, and we would like to know that effect. Also, the estimates α_i do not represent the effect of the corresponding i-th quarter yet. For example, the value 0,56 reflects an increase in the third quarter, taking into account the effect of the first quarter which is automatically absorbed in the value 1,605. To isolate the seasonal effects Sz_i, $i = 1, 2, 3, 4$, for all the quarters, we calculate

$$Sz_1 = -\frac{\alpha_2 + \alpha_3 + \alpha_4}{4},\ Sz_2 = \alpha_2 + Sz_1,\ Sz_3 = \alpha_3 + Sz_1,\ Sz_4 = \alpha_4 + Sz_1.$$

In our case, $Sz_1 = 1{,}4175$, so the other seasonal effects are $Sz_2 = -0{,}68$, $Sz_3 = 1{,}977$, $Sz_4 = -2{,}71$. Using one more auxiliary variable x_1 for the first season, we now have

$$\hat{y}_t=(1{,}605-1{,}4175)-0{,}023t+1{,}4175x_1-0{,}68x_2+1{,}977x_3-2{,}71x_4.$$

We can evaluate the theoretical values given by the resulting model. This is done in table 26. These values are visually compared to the original values of the time series (figure 6). The graph proves that the model has been selected well enough.

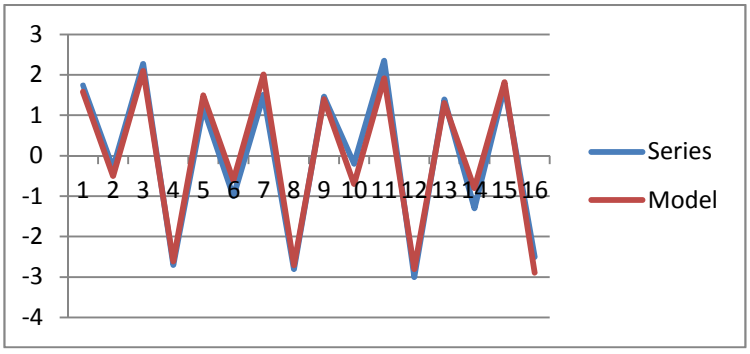
Table 26:Theoretical values given by the regression model

t	1	2	3	4	5	6	7	8
y-theoretical	1,582	-0,5	2,096	-2,617	1,49	-0,6	2,004	-2,709

t	9	10	11	12	13	14	15	16
y-theoretical	1,398	-0,7	1,912	-2,801	1,306	-0,8	1,82	-2,893

Source: author's

Figure 6: Comparison of the theoretical and empirical time series data



5.2.3 PROPERTIES OF THE RANDOM COMPONENT OF A REGRESSION MODEL

We considered in our models, based on their decomposition into the trend and seasonal component, a random component as well. It is this component that gives the final answer to how the time series in question behaves. We also used the principles of regression to estimate the unknown parameters that appeared in our model. Therefore, care should be taken to make sure that such estimated parameters have good statistical properties. To achieve this goal, it is necessary that the random part of the model satisfies the conditions of the classical regression model. These conditions are listed in the chapter on regression, and we repeat them here for convenience:

1. Expected value of ε_i is zero, or $E(\varepsilon_t) = 0$ for each t.
2. Variance of ε_t is constant, i.e. independent of t: $Var(\varepsilon_t) = \sigma^2$ for each t.
3. Variables ε_i, ε_j are uncorrelated, i.e. $Cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$.
4. Variable ε_i follows a normal distribution $N(0, \sigma^2)$ for each t.

If the condition 2 holds, we talk about **homoscedasticity** (in the opposite case, when the condition does not hold, we talk about **heteroscedasticity**).

The conditions above should be verified with appropriate statistical methods. We shall focus now on the third condition which typically isn't met in the case of time series. Before we do so, we note the following: condition 1 is not usually verified, and is assumed to be true. If condition 4 holds together with all the other listed conditions, the least squares estimates are the best of all unbiased estimates in a certain sense. If „only" the first three conditions hold, which is not to be taken for granted, of course, the least squares estimates will „only" be the best estimates of all the so-called linear unbiased estimates. There are some terms we just used that require an explanation. We shall not provide an explanation at this point because the main reason for making the note is to stress that we can still get reasonably good estimates of the unknown regression coefficients, using the least squares method, even if condition 4 does not hold. As far as condition 2 is concerned, one may use the Goldfeld-Quandt's test to verify the validity of the condition, or some other more general test, such as the White's test. However, heteroscedasticity is a problem, typical of cross-section data, not of time series data, and so we shall not be preoccupied with it here. For these reasons, we shall focus exclusively on the problem of autocorrelation (condition 3). An in-depth coverage of the other problems may be found in different school subjects (in Econometrics, for instance).

The analysis of condition 3 is based on analysing the residuals of the model, i.e. it is based on the values $e_t = Y_t - \hat{T}_t - \hat{S}_t, t = 0, \pm 1, \pm 2, \ldots$. Here, the original regression model, describing the time series, is of the form $Y_t = T_t + S_t + \varepsilon_t, t = 0, \pm 1, \pm 2, \ldots$. In the expression $e_t = Y_t - \hat{T}_t - \hat{S}_t$, Y_t represents an individual value of the series, $\hat{T}_t$ is an estimate of the trend, and $\hat{S}_t$ is an estimate of the seasonal component of the model.

To verify that the random terms of the model are uncorrelated, the Durbin-Watson's test is frequently used.

5.2.4 DURBIN-WATSON'S TEST

The test verifies the null hypothesis: the random terms of the model are uncorelated versus the alternative hypothesis: the random terms are correlated, the correlation taking the

following first-order autoregressive form AR(1): $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$, where u_t satisfies the conditions of the classical regression. The AR(1) model also contains an unknown parameter ρ, which is the population paired correlation coefficient measuring the correlation between ε_t and ε_{t-1}. The test analyses validity of the null hypothesis that there is no autocorrelation in the model versus the alternative hypothesis that there is an autocorrelation of the form AR(1). The test is realized in several steps. First, the least squares estimates of the unknown regression parameters of the model are found, and the estimates are used to calculate the residuals of the model e_t. Second, the residuals are used to construct the following test criterion

5-7
$$T = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2},$$

where *T* is the length of the times series. Special statistical tables are then used to compare the criterion with the critical value of the test. The tables are at the end of this textbook. To use the tables properly, for the given number of observations T, nivel of test α and the number of model parameters k, which does not include the absolute term of the model, a lower level d_L and an upper level d_H are found in the tables. Also, the sample paired correlation *r* between the neighbouring residuals has to be calculated according to

5-8
$$r = \frac{\sum_{t=2}^{T} e_t e_{t-1}}{\sum_{T=1}^{T} e_t^2}.$$

If r is positive, the conclusion to the test is such that if the test criterion 5-7 is greater than d_H, the null hypothesis is accepted, whereas if the criterion is smaller than d_L, the null hypothesis is rejected. If r is negative, the statistic $T^* = 4 - T$ is used instead of *T*, and the just-described conclusion is worded based on T^*. If *T* or T^* falls between the values d_L and d_H, the test is inconclusive, however, it is recommended that an autocorrelation be assumed preventively because this is usually the case in time series models.

PROBLEM 4

Table 27 contains fictitious data on households' monthly expenses on food (in millions of crowns). The data spans the time period from January 2000 (t = 1) to March 2001 (t = 15).

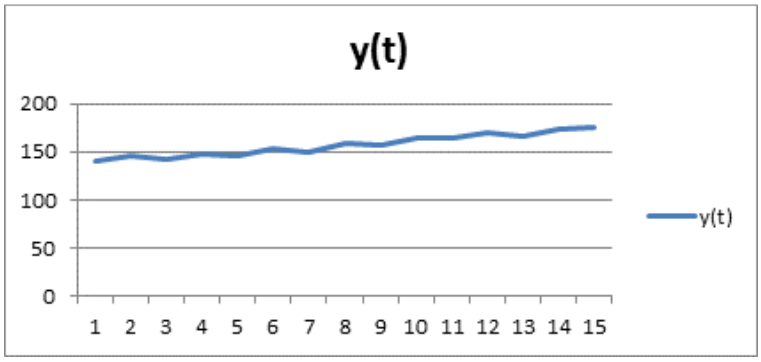
Table 27: Monthly expenditures Y_t

Y_t	141	145	142	147	146	154	150	158	157	165	164	170	167	174	175
t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Source: author's

Given the character of the series (figure 7), we shall assume that it follows a linear trend $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$ with a single explanatory variable $x_t = t$. The seasonal component is not assumed. We start with the estimation of the parameters, and the resulting model will be tested for autocorrelation, using the Durbin – Watson's test.

Figure 7: The time series Y_t



The least squares method gives

$$\boldsymbol{b} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \begin{pmatrix} 0{,}295 & -0{,}028 \\ -0{,}028 & 0{,}0036 \end{pmatrix} \cdot \begin{pmatrix} 2355 \\ 19552 \end{pmatrix} = \begin{pmatrix} 136{,}66 \\ 2{,}543 \end{pmatrix}.$$

Inserting the estimates in the model, we get the theoretical values $\hat{Y}_t = b_0 + b_1 t$ and the residuals $e_t = Y_t - \hat{Y}_t$, $e_{t-1} = Y_{t-1} - \hat{Y}_{t-1}$. Calculations necessary for the test are in table 28.

Table 28: Residuals and preparatory calculations

e_t	e_{t-1}	$(e_t - e_{t-1})^2$	e_t^2	$e_t \cdot e_{t-1}$
1,8			3,24	
3,25	1,8	2,1025	10,5625	5,85
-2,28	3,25	30,5809	5,1984	-7,41
0,17	-2,28	6,0025	0,0289	-0,3876
-3,37	0,17	12,5316	11,3569	-0,5729
2,08	-3,37	29,7025	4,3264	-7,0096
-4	2,08	42,6409	19,8025	-9,256
1	-4	29,7025	1	-4,45
-2,54	1	12,5316	6,4516	-2,54
2,91	-2,54	29,7025	8,4681	-7,3914
-0,62	2,91	12,4609	0,3844	-1,8042
2,82	-0,62	11,8336	7,9524	-1,7484
-2,71	2,82	30,5809	7,3441	-7,6422
1,74	-2,71	19,8025	3,0276	-4,7154
0,2	1,74	2,3716	0,04	0,348
	Sum	272,547	85,9438	-48,7297

For the Durbin – Watson's test, we get the criterion $DW = 272{,}547/85{,}9 = 3{,}1$. The estimated correlation between the residuals is according to 5-8: $r = -48{,}73/85{,}9 = -0{,}56$. In the tables at the end of the textbook, we find $d_L = 1{,}077$ and $d_H = 1{,}361$ for $T = 15$, the number of regressors without the absolute term = 1 and the nivel of test $\alpha = 0{,}05$. Since the sample paired correlation between the residuals is negative, we shall use the alternative test

criterion $T^* = 4 - DW = 0{,}9$. This number is very low, which means that we may assume the presence of an AR(1) autocorrelation in the model.

5.3 MOVING AVERAGES

The trends we have devoted ourselves to assumed that the objective was to find a single mathematical curve that would intersperse and describe the entire times series. Such an effort can be justified, for instance, in situations when the analysed time series is fairly short, so that it makes sense to explain the way it was generated with a single function. Another example is the situation when the user of the final model is interested in a longer-term trend of the series rather than its short-term fluctuations. However, we are often encountered with circuimstances under which using a single function to describe the series behaviour is a too ambitious objective. To give an example, let us consider a company that is interested in the short-term future of the series. If this is the case, the company will probably also be interested in the time period of the series that has just preceded the present, for it is this recent past that will most likely affect the future behaviour of the series to a great extent, as opposed to the more distant past. In such a case, it makes more sense to work with models that reflect potential shifts in the behaviour of the series. The recognition of the shifts means, in other words, that we distinguish between the behaviour of the series from the recent past and its behaviour from the distant past. This approach does not assume that the mechanism which generated the series remains the same across time. Approaches which assume shifts in the behaviour of the series are called **adaptive approaches**, and include, among other techniques, moving averages.

Moving averages are based on the principle that only a part of the series is modelled with a selected function, and for later analyses, only one of the values of the function – the middle value - is used as a representative. The function used to model the part of the series depends on unknown parameters which can be estimated with the least squares technique. It is assumed that the parameters may change in time.

In this chapter, we shall again work with an additive regression function $Y_t = T_t + \varepsilon_t$, where ε_t satisfies all the required conditions, and we also assume that the *analytical form* of the trend remains the same across time. However, as we move from one part of the series to another, the model used to describe the new part will generally have different parameters. Therefore, it is said that **models with changeable parameters** are applied. There are different types of moving averages. If a linear function is used to model various sections of the time series, we talk about a simple moving average. If a second-order polynomial is used for these purposes, we talk about a weighted moving average. Let us take an example to see the theoretical and practical aspects of working with moving averages.

5.3.1 SIMPLE MOVING AVERAGES

Let $Y_1, Y_2, \ldots, Y_n$ be a time series. To use the moving average technique, the length m of the series to be modelled must be determined first. Also, the order of the polynomial that will be used as a model must be selected. If a linear function is used, it will be a polynomial of order one. The length m is usually chosen to be an odd number which can be written as $m = 2p + 1$, where p is a positive integer. Each part o the series to be modelled has its center point. These are the values Y_t, where $t = p + 1, p + 2, \ldots, n - p$. This means that the subsequent part of the series to be modelled is the previous part of the series shifted forward by one observation. This is how the modelled section of the series is moved forward, hence the name of the technique. The center of the first section is given by Y_{p+1}, the center of the

second part equals Y_{p+2}, etc. … until the center of the final part of the series is Y_{n-p}. The k-th part of the series includes observations $Y_{p+k-p}, Y_{p+k-(p-1)}, \dots, Y_{p+k}, Y_{p+k+1}, \dots, Y_{p+k+p}$, which can be written as $\{Y_{p+k+j}\}_{j=-p}^{p}$, or $\{Y_{t+j}\}_{j=-p}^{p}$ for $t = p + k$. To model the part of the series with center point at time t, using the least squares, means to minimize the criterion

5-9
$$\sum_{j=-p}^{p} \left[Y_{t+j} - b_0(t) - b_1(t)j\right]^2.$$

Calculating the corresponding partial derivatives and putting them equal to zero to find the minimum, we arrive at equations

5-10
$$2\sum_{j=-p}^{p} \left[Y_{t+j} - b_0(t) - b_1(t)j\right](-1) = 0,$$
$$2\sum_{j=-p}^{p} \left[Y_{t+j} - b_0(t) - b_1(t)j\right](-j) = 0.$$

We said the principle of moving averages lies in working with a single representative of the modelled part of the series, which is to be the theoretical value of the center-point observation of that part. At this center point, $j = 0$, which means that $\hat{Y}_t = b_0(t)$ must hold. Thus, it suffices to calculate only the absolute term of the model, using 5-10. Doing so, we get

5-11
$$b_0(t) = \frac{1}{m}\sum_{j=-p}^{p} Y_{t+j}.$$

As we can see, the center-point representative of a given section of the series is the simple average of all the observations that make up the section. And this is where the name of the technique – simple moving average - came from.

PROBLEM 5

Table 29 contains a time series which we will now model with a simple moving average of length five.

Table 29: A time series for problem 5

t	1	2	3	4	5	6	7	8	9	10
Y_t	34	40	37	42	45	47	44	51	52	58

t	11	12	13	14	15	16	17	18	19	20
Y_t	55	64	59	66	68	62	72	75	72	77

Source: author's

Since $m = 5$, we have $p = 2$. The first theoretical value is $\hat{Y}_3 = \frac{1}{5}\sum_{j=-2}^{2} Y_{3+j} = 39{,}6$, the second theoretical value is $\hat{Y}_5 = \frac{1}{5}\sum_{j=-2}^{2} Y_{4+j} = 42{,}2$, etc.…until the last theoretical value is $\hat{Y}_{18} = \frac{1}{5}\sum_{j=-2}^{2} Y_{18+j} = 71{,}6$. Extending table 29 with the theoretical values, we get table 30.

Table 30: The original time series and the moving averages of length five

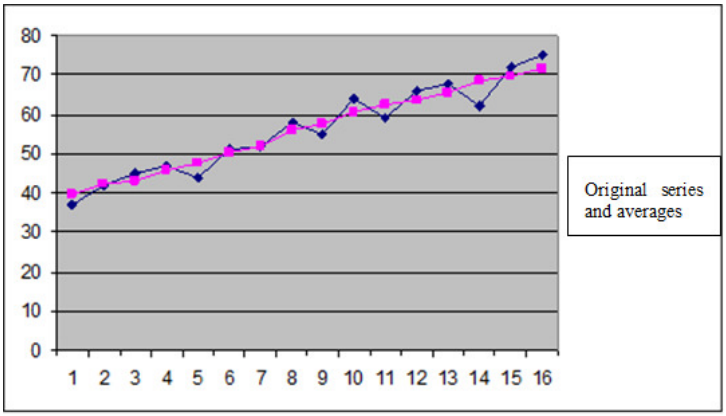
t	1	2	3	4	5	6	7	8	9	10
Y_t	34	40	37	42	45	47	44	51	52	58
average			40	42	43	46	48	50	52	56

t	11	12	13	14	15	16	17	18	19	20
Y_t	55	64	59	66	68	62	72	75	72	77
average	58	60	62	64	65	69	70	72		

Source: author's

Figure 8 compares the empirical and theoretical values of the series. The original series was shortened by leaving out the first two and the last two observations.

Figure 8: The time series and its moving averages



5.4 MAKING PREDICTIONS WITH TIME SERIES MODELS

One of the most important reasons why any time series model is constructed is to use the model for predictions of the future behaviour of the time series. Sometimes we also talk about making extrapolations. Predictions are based on the principle that the past mechanism which generated the series will keep its properties unchanged in the future. Thus, one may expect the model to work reasonably well even for future realizations of the series.

Let the time series model be of the form $Y_t = T_t + \varepsilon_t$, t = 1, 2,…, n, where T_t is a linear or quadratic trend (i.e. a second-order polynomial), and n be the present time point. The point prediction $\tilde{Y}_{n+h}$ of the uknown value Y_{t+h} of the series at a time point n + h, where h is positive and represents the time horizon of the point prediction, is given by:

5-12 $$\tilde{Y}_{n+h} = T_{n+h}.$$

Here, T_{n+h} is the trend evaluated at $n + h$. This value is not known, but can be estimated by the given trend regression function. The point prediction allows to estimate the future realization of the time series with a single number in a very simple and straightforward way – the future time point $n+h$ is used and inserted in the estimated trend component of the model.

Apart from point prediction, interval prediction for Y_{t+h} is constructed, as well. The interval prediction constructed at a time n for a time period $n+i$ is given by the following confidence interval:

- If the trend is linear, the 95% confidence interval is of the form

5-13 $$[\tilde{Y}_{n+i} - t_{\text{n-2}}(0{,}05)\, s\sqrt{Q_n(i)}\ ,\ \tilde{Y}_{n+i} + t_{\text{n-2}}(0{,}05)\, s\sqrt{Q_n(i)}\ ],$$

where

5-14 $$s = \sqrt{\frac{\sum_{t=1}^{n} Y_t^2 - \sum_{t=1}^{n} \hat{T}_t}{n-2}}$$

and

5-15 $$Q_n(i) = 1 + \frac{1}{n} + \frac{(n+i-\bar{t})^2}{\sum_{t=1}^{n} t^2 - n\bar{t}^2},\ \bar{t} = (n+1)/2.$$

- If the trend is quadratic, the 95% confidence interval is given by

5-16 $$[\tilde{Y}_{n+i} - t_{\text{n-3}}(0{,}05)\, s\sqrt{Q_n(i)}\ ,\ \tilde{Y}_{n+i} + t_{\text{n-3}}(0{,}05)\, s\sqrt{Q_n(i)}\ ],$$

where

5-17 $$s = \sqrt{\frac{\sum_{t=1}^{n} Y_t^2 - \sum_{t=1}^{n} \hat{T}_t}{n-3}}$$

and

5-18 $$Q_n(i) = 1 + \left[1, n+i, (n+i)^2\right] \cdot (X^T \cdot X)^{-1} \cdot \left[1, n+i, (n+i)^2\right]^T,\ X = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ . & . & . \\ 1 & n & n^2 \end{pmatrix}.$$

Any 95% confidence interval tells us that it contains the unknown realization of Y with probability = 0,95.

SUMMARY

In this chapter, we studied several important types of time series and possibilities of their use for making predictions. First, we examined the trend components of time series models, the components being expressed in the form of a polynomial or S-curve. In the latter case, a procedure of estimating the unknown parameters of the curve was demonstrated. We also analysed the seasonal component of the model, taking the regression approach again. Finally, the random component of the model and its properties were clarified for the model to be credible when exploited for predictions. One of these properties, the lack of autocorrelation, was examined further with the Durbin-Watson's statistical test. At the end of the chapter, formulas for point and interval predictions were presented for the case when the trend in the time series model is either linear or quadratic.

CONTROL TEST 5

5.1 The deterministic component of a time series model consists of (more of the following answers may be correct):

- **a.** trend component
- **b.** trend and seasonal components
- **c.** trend, seasonal and cyclical components
- **d.** seasonal and cyclical components

5.2 The periodical component of a time series model consists of:

- **a.** seasonal component
- **b.** trend and seasonal components
- **c.** trend, seasonal and cyclical components
- **d.** seasonal and cyclical components

5.3 Select an item from the left column of the following scheme, and decide what item from the right column it belongs to:

(1)	Additive time series model	(A)	has components summed together
(2)	Multiplicative t.s. model	(B)	is a line
(3)	Linear trend in a t.s. model	(C)	has components multiplied together

5.4 Complete the sentences:

- **a.** If the variance of the random part of a model is constant, the property is called _____.
- **b.** Random components of a model should be mutually _____.

5.5. The following table contains a time series. Model the series with a quadratic trend (no other component, apart from the random one, is assumed in the model). Use the least squares.

t	1	2	3	4	5	6	7	8	9	10
$\mathbf{Y_t}$	1,2	6,3	14,3	37,1	76,5	125	274	349	499	578

t	11	12	13	14	15	16	17	18	19	20
$\mathbf{Y_t}$	711	859	987	1114	1135	1349	1506	1680	1721	1890

Source: author's

5.6. Model the time series below with moving averages of length five.

t	1	2	3	4	5	6	7
$\mathbf{Y_t}$	18,683	15,236	20,552	20,988	30,598	23,22	38,375

T	8	9	10	11	12	13	14	15
$\mathbf{Y_t}$	43,698	47,813	61,403	62,002	68,386	63,904	68,247	67,818

Source: author's

5.7 Using the data in the tables below, estimate the model

$$Y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \varepsilon_t,$$

and verify the validity of the no-autocorrelation assumption with the Durbin-Watson's test. The nivel of test is five per cent.

t	1	2	3	4	5	6	7	8
$\mathbf{x_{t1}}$	3,3	3,4	3,5	3,5	3,4	3,3	3,4	3,2
$\mathbf{x_{t2}}$	5,9	6	6,2	6,3	6,3	5,9	5,9	5,8
$\mathbf{Y_t}$	25,3	23,02	19,9	20,95	18,59	16,15	15,22	17,26

t	9	10	11	12	13	14	15	16	17
$\mathbf{x_{t1}}$	3,2	3,1	3,1	3,1	3,2	3,1	3,1	3	3
$\mathbf{x_{t2}}$	5,5	5,4	5,2	4,8	4,8	4,7	4,6	4,5	4,5
$\mathbf{Y_t}$	18,98	20,09	18,65	17,79	20,84	16,69	18,33	16,79	16,48

Source: author's

SOLUTIONS

5.1 **c.**
5.2 **d.**
5.3 (1) – (A), (2) – (C), (3) – (B)
5.4 **a.** homoscedasticity **b.** uncorrelated.
5.5 The least squares method applied to the model

$$Y = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon$$

gives the estimates

$$\hat{\beta}_0 = -127,\ \hat{\beta}_1 = 38,34,\ \hat{\beta}_2 = 3,27.$$

5.6 Moving averages:

t	3	4	5	6	7	8	9	10	11	12	13
average	21,2	22,1	26,7	31,4	36,7	42,9	50,7	56,7	60,7	64,8	66,1

5.7 The least squares give the estimates $b_0 = 3,5$, $b_1 = 3,88$, $b_2 = 0,52$. The residuals are

t	1	2	3	4	5	6	7	8	9
e(t)	5,898	3,173	-0,439	0,566	-1,407	-3,256	-4,568	-1,703	0,17

t	10	11	12	13	14	15	16	17
e(t)	1,727	0,389	-0,266	2,401	-1,313	0,379	-0,719	-1,03

The sample correlation is

$$r = \frac{\sum_{t=2}^{17} e_t e_{t-1}}{\sum_{t=1}^{17} e_t^2} = \frac{40,3}{59,37} = 0,67.$$

Further, the Durbin-Watson's statistic equals

$$DW = \frac{\sum_{t=2}^{17} (e_t - e_{t-1})^2}{\sum_{t=1}^{17} e_t^2} = \frac{71,84}{59,37} = 1,21.$$

Since the model contains two parameters (the absolute term is excluded from the test), k is equal to 2. The sample size is n = 17. Therefore, the Durbin-Watson's test table provides us with values d_L = 1,015 and d_H = 1,536. Since DW lies between the two values, the test is inconclusive, as to whether there is an autocorrelation or not in the model.

6 ANALYSIS OF VARIANCE

Chapter 2 of this textbook dealt with two-sample t-tests. Chapter 4 is devoted to their extension in the form of analysis of variance - ANOVA. Analysis of variance ranks among one of the most frequently used statistical procedures in marketing as well as other areas of data analysis. The method enables one to assess the potential influence of a qualitative or quantitative variable on another quantitative variable. For example, it is possible to evaluate effects of different forms of a promotional campaign on the sales of a product. In this case, different promotional campaigns reresent different categories of the observed qualitative variable = promotional campaign. The sales are then the quantitative variable in question. The potential effect can be expressed mathematically in such a way that the expression analyses whether a change in the level of the qualitative/quantitative variable changes the population mean of the other observed quantitative variable. In this sense, ANOVA tests if there are any differences among the population means of the quantitative variable.

Mathematically speaking, the basic idea of ANOVA is given by a decomposition of what is called the **total variability of the observed variable**. The decomposition is made up of different sources of the total variability. There is more than one term forming the decomposition. Some of the terms represent the main sources of the total variability. Another term is called the **residual variability**, which reflects the influence upon the total variability of all the other minor sources. Depending on how many main sources or **factors** appear in the decomposition, we talk about one-way ANOVA, two-way ANOVA and so on.

A shrewd observer might come up with the suggestion that the two-sample t-test could be used several times instead. Such a procedure would test potential differences of various population means under scrutiny. In this case, if none of these tests ended up being significant, i.e. the null hypothesis of equal population means woud always be accepted for each pair, we could conclude that all the population means are the same, i.e we would conclude that the factor has no effect. Theoretically speaking, it is possible to proceed this way, but at the cost of credibility of this procedure. Recall that every statistical test is accompanied by errors, and if a whole series of tests is realized, the probability of these errors may cumulate to unbearable levels. This is the reason why ANOVA was developed as a special procedure to keep the probability at a reasonal level.

We shall discuss the one-way and two-way ANOVA in this chapter. ANOVA stands for „ANalysis Of VAriance“. After having studied the subject matter on ANOVA, we encourage the readers to try to resolve theproblems presented in this book on their own, and check their results against those presented at the end of the chapter.

6.1 ONE-WAY ANOVA

There are many situations when k independent data samples are available, and the samples do not come from the same population. The sample sizes are $n_1, n_2, ..., n_k$, k being equal or greater than 2. The sample mean $\bar{x}_i$ can be calculated for the i-th sample, as well as the sample variance s_i^2. In practice, these samples usually originate by classifying the population into k classes by a factor X, and then drawing randomly data of size n_i from each of these k classes. The variable X is called **factor** and its levels or categories are given beforehand, so the factor is called controllable. The categories of X are denoted $x_1, x_2, ..., x_k$.

Let the factor *X* is observed at *k* levels (categories) and may potentialy influence a statistical quantitative variable *Y*. Values of *Y* randomly obtained for the *i*-th category x_i of *X* are denoted $y_{i1}, y_{i2}, ..., y_{in_i}$. It is convenient to organize the entry data into table 31.

Table 31: ANOVA table

Factor level	Data sample for the factor level	Sample size	Sample mean	Sample variance
1	$y_{11}, y_{12}, ..., y_{1j}, ..., y_{1n_1}$	n_1	$\bar{y}_1$	s_1^2
2	$y_{21}, y_{22}, ..., y_{2j}, ..., y_{2n_2}$	n_2	$\bar{y}_2$	s_2^2
⋮	⋮	⋮	⋮	⋮
i	$y_{i1}, y_{i2}, ..., y_{ij}, ..., y_{in_i}$	n_i	$\bar{y}_i$	s_i^2
⋮	⋮	⋮	⋮	⋮
k	$y_{k1}, y_{k2}, ..., y_{kj}, ..., y_{kn_k}$	n_k	$\bar{y}_k$	s_k^2
Total		N	$\bar{y}$	s^2

The main principle of the analysis of variance is to decompose the total variability of the observed variable. The total variability, measured by the sum of squared deviations of the individual values of the variable from their average, is divided by the decomposition into a part that reflects a variability within the samples and a part which reflects a variability between the samples.

The total variability is usually measured by sample variance:

$$s^2 = \frac{\sum_i \sum_j \left(y_{ij} - \bar{y} \right)^2}{N - 1}.$$

In analysis of variance, we are interested only in the numerator of the variance, where

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} .$$

We shall denote the total sum of squares, which represents the total variability, as S_y:

6-1
$$S_y = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y} \right)^2 .$$

The symbol $S_{y,v}$ will be used for the **within-group variability**, which is also called residual variability:

6-2
$$S_{y,v} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \overline{y}_i \right)^2 .$$

The **among-group variability**, denoted $S_{y,m}$, is defined as:

6-3
$$S_{y,m} = \sum_{i=1}^{k} \left(\overline{y}_i - \overline{y} \right)^2 \cdot n_i .$$

Expressions 6-1 to 6-3 use $\overline{y}$, the sample average of all the values of *y*, as well as the subgroup averages $\overline{y}_i$ (see table 31).

Using algebraic operations, the following fundamental formula for the one-way analysis of variance can be derived:

6-4
$$S_y = S_{y,m} + S_{y,v} .$$

Anglo-Saxon scholarly literature and software may denote the just-described variabilities with other symbols, as well. For instance,

$S_y = S_D$ (*D* for Difference),
$S_{y,m} = S_T$ (*T* for Treatment),
$S_{y,v} = S_R$ (*R* for Residual).

We shall use our symbols in the rest of this chapter.

6.1.1 ANOVA HYPOTHESES

Analysis of variance is a statistical test. Therefore, we work with a pair of hypotheses: a null hypothesis and an alternative hypothesis. Before specifying the test, we emphasize that ANOVA has its conditions under which it was derived. The method assumes that each of the *k* random samples comes from a normal distribution, and the distributions have the same variance. Also, the samples were drawn independently of each other. The prerequisite of normality can be tested in more than one way, using the chi-square test, Anderson-Darling's test, Kolmogorov-Smirnov's test, Shapiro-Wilk's test, etc. Regarding the condition of constant variance, we described earlier the F-test which verifies the hypothesis of equality of two variances. In analysis of variance, more than two samples are usually worked with, and for this case, an extension of the F-test exists in the form of Bartlett's test.

Let us return to ANOVA. Assuming the factor *X* is observed at *k* levels, the following relation is considered to hold true.

6-5
$$\mu_i = \mu + \alpha_i , \; i = 1, 2, ..., k.$$

Here, μ_i is the population average of the variable *Y*, corresponding to the *i*-th level of factor *X*, μ is a constant and α_i is called *effect*. It is this effect that is supposed to express the potential difference among the population means of *Y*, the differences in means being caused by different levels of factor *X*. Now, we may ask the question whether all *k* samples came from the same populations. In other words, whether the populations the samples came from have the same means. Whether the effects α_i are all equal to zero would be yet another equivalent question. This question forms the null hypothesis of ANOVA:

6-6 $$H_0: \ \mu_1 = \mu_2 = ... = \mu_k .$$

or

6-7 $$H_0: \ \alpha_1 = \alpha_2 = ... = \alpha_k = 0 .$$

The alternative hypothesis is the negation of 6-6 or 6-7. In the first case, this means that the alternative hypothesis takes the form: H_1 – there exist indices *i* and *j*, such that $\mu_i \neq \mu_j$.

The test criterion of ANOVA is

6-8 $$T = \frac{S_{y,m} / (k-1)}{S_{y,v} / (N-k)},$$

which follows a **Fisher's distribution** with *k*-1 and *N-k* degrees of freedom. The critical value of the test $F_{k-1,N-k}(\alpha)$ for a nivel of test alpha is tabulated, or it can be obtained with the Excel function FINV(α, *k*-1, *N-k*).

To sum up, testing the null hypothesis of ANOVA is characterized by the following steps:

Step 1. Select a nivel of test α. Alpha is usually 0,1, 0,05, 0,01, or 10%, 5%, 1%, respectively.

Step 2. Calculate the test criterion *T* according to 6-8, where formulas 6-2 and 6-3 are used to get the within-group variability and among-group variability, respectively. Formulas 6-9 and 6-10 may also be used. They are more convenient if the variabilities are to be calculated on a calculator.

6-9 $$S_y = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}^2 - \frac{1}{N} \left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} \right)^2 ,$$

6-10 $$S_{y,m} = \sum_{i=1}^{k} n_i \bar{y}_i^2 - \frac{1}{N} \left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} \right)^2 ,$$

6-11 $$S_{y,v} = S_y - S_{y,m} .$$

Step 3. Compare *T* from step 2 with the critical value $F_{k-1_1,N-k}(\alpha)$. If $F < F_{k-1_1,N-k}(\alpha)$, the null hypothesis H_0 is accepted and the factor *X* can be pronounced not influential in relation to the variable *Y*. If $F \geq F_{k-1_1,N-k}(\alpha)$, the null hypothesis H_0 is rejected, meaning the factor *X* has a statistically significant influence on the variable *Y*.

If the test confirms that the factor *X* affects *Y*, we may ask which population means are different. It can be the case that only population means are different, while all the other population means are the same. There are methods that try to answer this question, one of them being devised by Scheffé and one by Tukey.

PROBLEM 1

The following table contains data obtained through several independent random samplings. The observed factor is the number of octanes used to describe the quality of car fuel (90, 91, 95, 98 octanes are usually available). Thus, the factor is monitored at four possible levels. For each of the levels, five car drivers using the fuel of the corresponding quality were randomly selected (see table 32). In this case, all samples have the same size, which is not required for one-way ANOVA. We want to know whether the quality of the fuel affects fuel consumption (car mileage). To answer the question, we shall employ ANOVA.

Table 32: Car mileage for different types of fuel

Factor levels	**90**	**91**	**95**	**98**
Samples	8,1	7,7	7,6	7,5
	8	7,8	7,6	7,8
	7,9	7,9	7,5	7,6
	7,8	7,6	7,6	7,5
	8,2	7,8	7,6	7,5

Source: author's

The nivel of test is set at 5%. Regarding the among-group variability, we must calculate the column averages (or group averages, more generally speaking). These are 8, 7,76, 7,58 and 7,58 for the first, second, third and fourth column of the table, respectively. The total average is 7,73. Using 6-3 with $n_i = 5$ for every *i*, we have $S_{y,m} = 0,594$. The within-group variability is $S_{y,v} = (8,1-8)^2 + (8-8)^2 + ... + (8,2-8)^2 + (7,7-7,76)^2 + ... + (7,5-7,58)^2 = 0,228.$

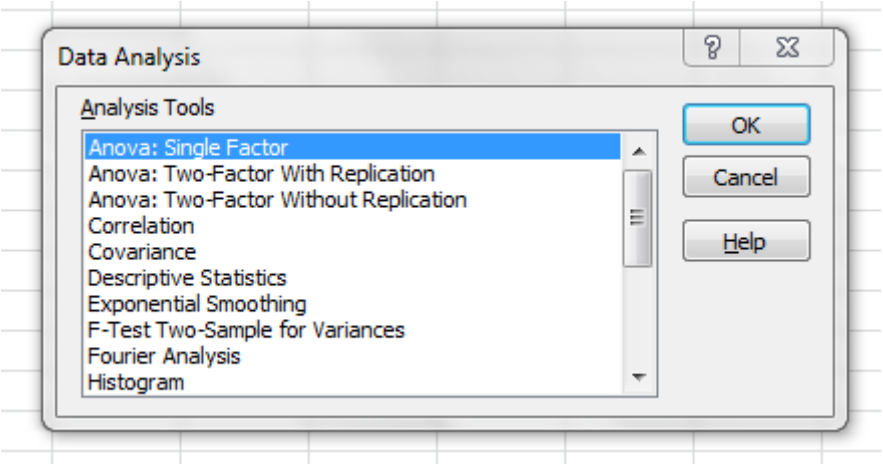
We have *N* = 20 values altogether and the number of factor levels is *k* = 4. Therefore,

$$T = \frac{S_{y,m}/(k-1)}{S_{y,v}/(N-k)} = \frac{0,594/3}{0,228/16} = 13,895.$$

The critical value *K* = FINV(0,05,3,16) = 3,2389. Since the test criterion is greater than *K*, we reject the hypothesis that fuel quality has no effect on car mileage. In other words, it seems the factor does have an influence on car mileage.

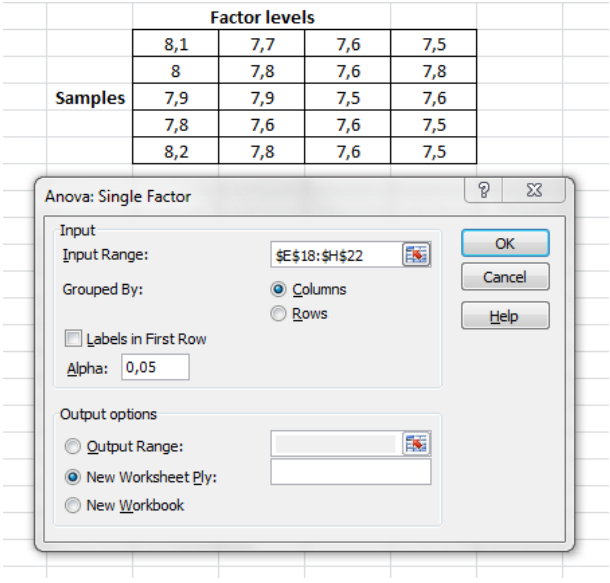
Excel: The same procedure can be performed in Excel with its Data Analysis module. The module offers one-way analysis of variance (see figure 9).

Figure 9: The dialogue window of the Data Analysis module



In the subsequent dialogue window (figure 10), it is necessary to insert as the Input Range a reference to the area of the Excel spreadsheet that contains the data samples to be worked with in ANOVA:

Figure 10: Insertion of information for ANOVA



The analyst must also confirm in the dialogue window whether each sample represents a column in the spreadsheet table, or a row (see figure 10). In our example, we work with columns. Once the nivel of test alpha is confirmed at 5%, or changed, and the location of the ANOVA output is selected, Excel returns a result of the form shown in table 33.

Table 33: ANOVA results presented by Excel

ANOVA

Source of variability	*SS*	*Difference*	*MS*	*F*	*P-value*	*F krit*
Among-groups	0,594	3	0,198	13,89474	0,0001015	3,238872
Within-groups	0,228	16	0,01425			
Total	0,822	19				

In this table, „F“ represents the test criterion, and „F krit“ stands for the critical value. As we can see, our calculations were correct.

6.1.2 A MEASURE OF DEPENDENCE

Variability of $\overline{y}_i$'s around $\overline{y}$ is caused by a dependence of *Y* on *X*. We described such variability with the among-group sum of squares $S_{y,m}$. The within-group variability $S_{y,v}$ is, on the other hand, induced by factors other than *X*. Higher $S_{y,m}$ implies a stronger dependence of *Y* on *X*. Based on 6-4, this dependence can be measured, using the *determination ratio*, denoted P^2:

6-12
$$P^2 = \frac{S_{y,m}}{S_y}.$$

The square root of P^2 is called the *correlation ratio*.

P^2 can take on any value from interval [0,1].The stronger the dependence of *Y* on *X*, the closer the characteristic is to one, and the closer the among-group sum of squares is to the total sum of squares (total variability). On the contrary, the within-group variability approaches zero in this situation. The closer the determination ratio is to zero, the smaller the part of the total variability which is accounted for by the among-group variability. In this case, the dependence of *Y* on *X* is weak.

SUMMARY

If we are interested in whether there is no statistically significant difference between two population means, we can verify such hypothesis or surmise with the two-sample t-test. Analysis of variance (ANOVA) enables us to verify the hypothesis that there is no difference between two or more population means. The procedure makes it also possible to test if different levels of one factor or more factors have any effect on another quantitative variable. Analysis of variance is based on the idea that total variability of a variable can be broken down to sub-variabilities each of which reflects its own source of variation. One type of variability is called residual variability, and is generated by sources that are not of interest to us, and are hard or impossible to identify. Depending on how many sources of the total variability we work with, we talk about one-way ANOVA, two-way ANOVA, three-way ANOVA, etc. This chapter was devoted to the first type: one-way ANOVA.

In one-way ANOVA, the total variability/sum of squares is divided into two parts: one part represents the influence of the only factor considered, while the other part is represented by the residual variability/sum of squares. We assume that the only factor *X* is observed at

k possible levels, and we formulate the null hypothesis that all samples, each of which is obtained randomly for the given level of *X*, came from the same population. To verify the null hypothesis, we use statistic 6-8, which follows a Fisher's distribution if the null hypothesis is true. The appropriate critical value is found for a given nivel of test alpha. In the end, the null hypothesis is either accepted, meaning that *X* has no effect, or we reject the null hypothesis, in which case the factor does exert an influence.

Apart from the testing itself, which provides a yes/no answer to the existence of an effect of the factor, we can also measure the amount of the effect. This is done by evaluating the determination ratio, the values of which belong to closed interval [0, 1]. The stronger the influence, the higher the value of the ratio. Its square root is called the correlation ratio.

CONTROL TEST 6

6.1 One-way ANOVA serves for (check the correct answer(s)):

- **a.** calculating frequency distribution of individual variables
- **b.** testing presence of effect of a factor on a quantitative variable
- **c.** finding probability distribution
- **d.** testing mutual correlation of statistical variables

6.2 In ANOVA, we

- **a.** test the null hypothesis that population means are the same,
- **b.** test the null hypothesis that two variables are mutually dependent,
- **c.** test the null hypothesis that value of a variable is different from a preedefined value
- **d.** test the null hypothesis that two statistical variables are mutually independent.

6.3 ANOVA uses the critical value of:

- **a.** a student's distribution,
- **b.** a Pearson's distribution,
- **c.** a Fisher's distribution,
- **d.** a normal distribution.

6.4 Determine whether the following statements are true (write T) or false (write F):

- **a.** The F-test of equal variances should be used before ANOVA.
- **b.** The determination ratio takes on values from interval [0,1].
- **c.** The smaller the among-group variability, the stronger the dependence between *X* and *Y*.
- **d.** The ANOVA test criterion must fall to a certain set for the ANOVA null hypothesis to be accepted. This set is a union of two intervals.
- **e.** The variance of ANOVA sample/group averages reflects the within-group variability.

6.5 Complete the statement:

a. If the ANOVA test criterion F falls to critical region, the variable Y can be considered to be _____ on/of X for a given nivel of test.

b. ANOVA, where the single factor is observed at l different levels, and the total number of all observations of Y equals m, works with a Fisher's distrbution with _____and _____degrees of freedom.

c. The ANOVA test criterion F is always _____ (positive/negative).

d. One-way ANOVA tests_____ Y on/of a factor X.

6.6 Complete the statement:

a. The square root of the determination ratio is called _____ _____.

b. If the ANOVA test criterion F falls to _____ _____ , the null hypothesis is rejected.

c. To find the critical region in ANOVA, we must know _____ and _____.

SOLUTIONS

6.1 **b.**
6.2 **a.**
6.3 **c.**
6.4 T, T, F, F, F
6.5 **a.** dependent, **b.** l-1 and m-l, **c.** positive **d.** independence
6.6 **a.** correlation ratio, **b.** critical region, **c.** degrees of freedom of a Fisher's distribution and the nivel of test.

7 TWO-WAY ANOVA AND LATIN SQUARES

We got acquainted with one-way ANOVA in the previous chapter. Now we shall work with more factors. We are in a situation where influence of two or three factors on a quantitative variable is examined, the factors being again qualitative or quantitative. Therefore, we shall work with two-way and three-way ANOVA. Two-way and three-way ANOVA have their own experimental plans. More on experimental plans will be presented in later chapters. These plans can be designed in such a way that not much data is necessary for ANOVA to yield credible results. The plans play an important role in statistics since the more factors appear in the analysis, the more data is needed, and the increase in the amount of data is then fairly steep.

If influence of two factors on a quantitative variable Y is observed, we talk about two-way ANOVA. As in the previous chapter, random sampling can be performed for different combinations of the two factors considered, and the data provides us later with the possibility of examining a potential influence of each of the two factors individually. An interaction of the two factors can also be regarded as another factor. We will skip interactions in our ANOVA presentation. Analogous statements are true for the three-way ANOVA case in which three major factors appear, and if it is required, two-factor interactions and a three-factor interaction may be analysed, as well.

As was already said, a lot of data is required when more factors are included in ANOVA. Therefore, it is often the case that only one observation is made for each combination of the factors. We then talk about ANOVA with a single observation in each subgroup. The case of the same number of observations in each subgroup was exploited in the previous chapter. However, while one-way ANOVA can do without this experimental plan, the case of two-way and three-way ANOVA is more complicated, and it is strongly recommended that the requirement of the same number of observations be complied with whenever possible in practice since in the opposite case, ANOVA may be carried out in more than one way, and each of the ways give generally different results. This is not the case when the number of observations in each subgroup is the same.

7.1 TWO-WAY ANOVA

In two-way ANOVA, two factors are considered. In this case, the total variability, inroduced in chapter six, is decomposed again, but this time into more terms each of which reflects an effect of the corresponding factor. As opposed to the case of one-way ANOVA, where two terms appeared in the decomposition, two-way ANOVA leads to three such terms. The additional term represents an effect of the second factor.

The decomposition of the total variability takes the following form:

7-1
$$S = S_A + S_B + S_R ,$$

where

7-2
$$S = \sum_{j=1}^{k} \sum_{i=1}^{n} (y_{ij} - \bar{y})^2 ,$$

7-3
$$S_A = k\sum_{i=1}^{n}\left(\overline{y}_i - \overline{y}\right)^2,$$

7-4
$$S_B = n\sum_{j=1}^{k}\left(\overline{y}_j - \overline{y}\right)^2$$

7-5
$$S_R = S - S_A - S_B.$$

Here, n is the number of levels/categories of factor A, k is the number of levels/categories of factor B. There are nk observations altogether with a single observation in each subgroup. The symbol $\overline{y}_i$ is used to denote the average observation when A is at its i-th level, while $\overline{y}_j$ is the average observation when B is at its j-th level. The symbol $\overline{y}$ denotes the average of all observations, as usual. The scheme of the experiment is given in table 34.

Table 34: Data scheme for two-way ANOVA

		Factor B: levels			
		B1	**B2**	**…**	**Bk**
Factor A: levels	**A1**				
	A2				
	.				
	.				
	An				

$\overline{y}_j$ y_{ij} in the i-th row, j-th column $\overline{y}_i$

The symbol $\overline{y}_i$ can be regarded as the row average in our table representation, and the symbol $\overline{y}_j$ as the column average. The sum S_A reflects the effect of A, the sum S_B represents the effect of B, the sum S_R reflects the effect of all the other factors. The total variability of all observations is described by the term S. Since two factors are involved in this analysis, two-way ANOVA consists of two statistical tests. Each of the tests examines the statistical significance of one of the factors.

7.1.1 *EFFECT OF FACTOR A*

We test the null hypothesis H_0: Factor A has no effect on the variable Y. The alternative hypothesis says the opposite: H_1: Factor A has an effect on the variable Y.

The test criterion T is of the form

7-6
$$T = \frac{S_A/(n-1)}{S_R/(nk-n-k+1)}.$$

The critical value $K = F_{n-1,nk-n-k+1}(\alpha)$ for a nivel of test alpha, i.e. it concerns a Fisher's distribution with *n*-1 and *nk*-*n*-*k*+1 degrees of freedom.

If $T \geq K$, we reject the null hypothesis. In this case, we may conclude that the factor *A* does have an effect on *Y*. If $T < K$, we accept the null hypothesis, and the factor *A* does not have an effect on *Y*.

7.1.2 *EFFECT OF FACTOR B*

We test the null hypothesis H_0: Factor *B* has no effect on the variable *Y*. The alternative hypothesis says the opposite: H_1: Factor *B* has an effect on the variable *Y*.

The test criterion *T* is of the form

7-7
$$T = \frac{S_B / (k-1)}{S_R / (nk-n-k+1)}.$$

The critical value $K = F_{k-1,nk-n-k+1}(\alpha)$ for a nivel of test alpha.

IF $T \geq K$, we reject the null hypothesis, meaning that the factor *B* affects *Y*. On the other hand, if $T < K$, the null hypothesis is accepted, and the factor does not affect *Y*.

PROBLEM 1

Two factors A, B are given. The factor A is observed at three levels, the factor B is considered at four levels. A single observation is available for each combination of the two factors. We assume the observations originated independently of each other, and they follow a normal distribution with equal variances. The nivel of test being five per cent, we want to test the potential influence of the two factors. The observations are in table 35.

Table 35: Entry data for problem 1

		B			
		B1	**B2**	**B3**	**B4**
A	**A1**	24	25	25	23
	A2	22	21	22	25
	A3	21	22	21	21

Source: author's

Table 36 is an extension of table 35, containing the row and column averages as well as the total average. Also, $n = 3$ and $k = 4$. Using formulas 7-2 through 7-5, we get

Table 36: ANOVA and various averages

		B				
		B1	**B2**	**B3**	**B4**	**averages**
A	**A1**	24	25	25	23	**24,25**
	A2	22	21	22	25	**22,5**
	A3	21	22	21	21	**21,25**
averages		**22,33333**	**22,66667**	**22,66667**	**23**	**22,66667**

Total average

$$S = \sum_{j=1}^{k}\sum_{i=1}^{n}(y_{ij} - \overline{y})^2 = 30,66,$$

$$S_A = k\sum_{i=1}^{n}\left(\overline{y}_i - \overline{y}\right)^2 = 18,166,$$

$$S_B = n\sum_{j=1}^{k}\left(\overline{y}_j - \overline{y}\right)^2 = 0,66,$$

$$S_R = S - S_A - S_B = 11,833.$$

Therefore:

To test the effect of *A*:

$$T = \frac{S_A/(n-1)}{S_R/(nk-n-k+1)} = \frac{18,166/2}{11,833/6} = 4,6.$$

To test the effect of *B*:

$$T = \frac{S_B/(k-1)}{S_R/(nk-n-k+1)} = \frac{0,66/3}{11,833/6} = 0,1126.$$

The critical values are:

for the test of *A*:

K = FINV(0,05,3-1,12-3-4+1) = 5,143.

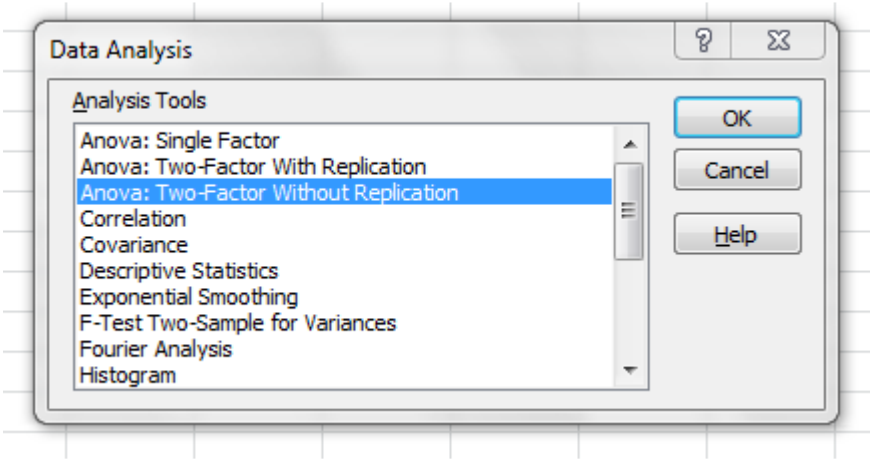
for the test of *B*:

K = FINV(0,05,4-1,12-3-4+1) = 4,757.

As we can see, none of the factors seems influential.

Excel: The same problem can be solved in Excel, using the Data Analysis module. In this module, Anova: two factors without repetition is selected (figure 11).

Figure 11: The dialogue window of the Data Analysis module in Excel



After selecting ANOVA in the dialogue window, a reference is made to the area of the spreadsheet containing the entry data, and the nivel of test is confirmed at five per cent or changed. Excel then returns the following result (table 37):

Table 37: Two-way ANOVA results provided by Excel

ANOVA

Source of variability	*SS*	*Difference*	*MS*	*F*	*P-vaue*	*F krit*
Rows	18,16666667	2	9,083333	**4,605634**	0,06137	**5,143253**
Columns	0,666666667	3	0,222222	**0,112676**	0,949497	**4,757063**
Error	11,83333333	6	1,972222			
Total	30,66666667	11				

The interpretation of the table is the same as in the case of one-way ANOVA. The second column contains the terms making up the total variability in 7-1. The *F* symbol is the test criterion for the two tests of ANOVA and *F krit* stands for the corresponding critical values.

7.2 THREE-WAY ANOVA (LATIN SQUARES)

A special case of a three-factor procedure, called Latin squares, belongs to analysis of variance, as well. We shall describe the procedure at the end of this chapter. Latin squares rank among classical methods of experimental design. The name of the procedure dates back to the eighteenth century when L. Euler (1707 – 1783) presented to the Petrohrad-based academy a problem on 36 commissioned officers: the task was to position officers of 6 different ranks from 6 different regiments on a square in such a way that each row and column of the square contained officers of all ranks from all regiments. More generally speaking:

Let us have objects with two properties of interest: A and B (example: A = rank, B = regiment). Each property may take on n different forms (example: n = 6, 6 different military ranks, such as private, corporal, sergeant, captain, major and colonel; 6 different regiments). The task is to set up the n^2 objects, each having a different A and B properties, so that each row and column was occupied by objects that do not have the same A and B properties

(example: the first row is occupied by the private from the first regiment, the corporal from the second regiment, etc.). Such a scheme is called the **Latin square of order *n***. The well-known mathematical result discovered by Euler himself says that at least one Latin square of order *n* exists for any integer *n*. We will use Latin squares for three-way analysis of variance.

Let us have three factors an effect of which on another variable *Y* is conceivable at the moment. Since three factors are involved in the analysis, it is hard to represent the whole experiment with a two-dimensional table. Another problem is that the number of data increases significantly as more factors are included in the analysis. However, it is possible to consider each combination of the factor levels, and utilize a single observation for each of these combinations. Such a procedure allows us to represent the experiment with a table. The heading of the table will describe different levels of two factors, while the interior of the table will contain a level of the third factor together with the single observation realized by an experiment. The levels of the third factor will be selected in such a way that the whole scheme results in a Latin square.

Let us denote the three factors *A*, *B* and *C*. When talking about a Latin square of order $n = 3$, we may describe our experiment in the form of table 38

Table 38: Three-way ANOVA scheme in the form of a Latin square

a	**b**	**c**
b	**c**	**a**
c	**a**	**b**

One side of the square represents the three levels of factor *A*. The adjoining side of the square represents the three levels of factor *B*. The interior of the table contains the levels of the third factor *C*. To read the square correctly, we say, for instance, that when both the factor *A* and *B* are at their first level, the factor *C* is as well at its first level (this corresponds to the element [1,1] of the table). For this combination of the levels, a single observation of the variable *Y* is realized and inscribed into the table. The analogous procedure holds true for the other elements of the table. One of the merits of this experiment is that we need only 9 observations instead of 27 we would have needed if we had wanted to use all the possible combinations of the factors. At the same time, the design of the experiment is such that the analysis of variance will give credible results.

The total variability *S* in three-way ANOVA has the form

7-8 $$S = S_A + S_B + S_C + S_R,$$

where

7-9 $$S_A = n\sum_{i=1}^{n}(\bar{y}_{i\bullet\bullet} - \bar{y})^2$$

reflects the effect of factor *A*, and $\bar{y}_{i\bullet\bullet}$ is the average of *Y* when *A* is at its *i*-th level,

7-10
$$S_B = n\sum_{j=1}^{n}(\bar{y}_{\bullet j\bullet} - \bar{y})^2$$

reflects the effect of factor *B*, and $\bar{y}_{\bullet j\bullet}$ is the average of *Y* when *B* is at its *j*-th level,

and

7-11
$$S_C = n\sum_{k=1}^{n}(\bar{y}_{\bullet\bullet k} - \bar{y})^2$$

reflects the effect of factor *C*, where $\bar{y}_{\bullet\bullet k}$ is the average of *Y* when *C* is at its *k*-th level.

Finally,

7-12
$$S = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y})^2$$

is the total sum of squares. The expression contains individual observations of *Y*, i.e. the terms y_{ijk}, and also the average of all observations $\bar{y}$.

Also,

7-13
$$S_R = S - S_A - S_B - S_C .$$

is the residual sum of squares.

Three-way ANOVA comprises three statistical tests. Each of these tests relates to one of the factors. Also, in each of the tests, the null hypothesis has the form: H_0 – **the tested factor is insignificant**. The alternative hypothesis is: H_1 **the tested factor has an effect**. The tests criteria for each test are summarized in table 39:

Table 39: Three-way ANOVA

Source of variability	Sum of squares	Degrees of freedom	Estimate of variance	Test criterion
Factor A	S_A	df_A=n-1	$MS_A=S_A / df_A$	$F_A=MS_A / MS_R$
Factor B	S_B	df_B=n-1	$MS_B=S_B / df_B$	$F_B=MS_B / MS_R$
Factor C	S_C	df_C=n-1	$MS_C=S_C / df_C$	$F_C=MS_C / MS_R$
Residuals	S_R	df_R=(n-1)(n-2)	$MS_R=S_R / df_R$	
Total	S	$df_T=n^2$-1		

If $F_A \geq F_{n-1,(n-1)(n-2)}(\alpha)$, the null hypothesis is rejected, and factor A is considered significant. If the opposite inequality holds, the null hypothesis is accepted.

If $F_B \geq F_{n-1,(n-1)(n-2)}(\alpha)$, the null hypothesis is rejected, and factor B is considered significant. If the opposite inequality holds, the null hypothesis is accepted.

If $F_C \geq F_{n-1,(n-1)(n-2)}(\alpha)$, the null hypothesis is rejected, and factor C is considered significant. If the opposite inequality holds, the null hypothesis is accepted.

PROBLEM 2

Fuel emission *Y* is studied, and its potential dependence on the following three factors:

Factor 1 = *petrol ingredient (A, B, C, D)*,
Factor 2 = *car driver (I, II, III, IV)*,
Factor 3 = vehicle used *(1, 2, 3, 4)*.

The result of the corresponding experiment is in table 40

Table 40: Entry data for the three-way ANOVA in problem 2

driver\vehicle	*1*	*2*	*3*	*4*
I	*A : 21*	*B : 26*	*D : 20*	*C : 25*
II	*D : 23*	*C : 26*	*A : 20*	*B : 27*
III	*B : 15*	*D : 13*	*C : 16*	*A : 16*
IV	*C : 17*	*A : 15*	*B : 20*	*D : 20*

We are testing the potential effect of the individual factors on *Y*, the nivel of test being five per cent.

i	$\bar{y}_{i\bullet\bullet}$
1=A	18
2=B	22
3=C	21
4=D	19

j	$\bar{y}_{\bullet j\bullet}$
1=I	23
2=II	24
3=III	15
4=IV	18

k	$\bar{y}_{\bullet\bullet k}$
1	19
2	20
3	19
4	22

According to formulas 7-9 through 7-13, we have

$$S_1 = n\sum_{i=1}^{n}(\bar{y}_{i\bullet\bullet} - \bar{y})^2 = 40.$$

$$S_2 = n\sum_{j=1}^{n}(\bar{y}_{\bullet j\bullet} - \bar{y})^2 = 216.$$

$$S_3 = n\sum_{k=1}^{n}(\bar{y}_{\bullet\bullet k} - \bar{y})^2 = 24.$$

$$S = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y})^2 = 296.$$

$$S_R = S - S_A - S_B - S_C = 16.$$

The test criteria are:

For factor 1: $$T = \frac{(40/3)}{(16/6)} = 5.$$

For factor 2: $$T = \frac{(216/3)}{(16/6)} = 27.$$

For factor 3: $$T = \frac{(24/3)}{(16/6)} = 3.$$

The critical value is the same in all three tests: K = FINV(0,05,3,6) = 4,757. This means the factors 1 and 2 are statistically significant, whereas type of vehicle used is not.

SUMMARY

This chapter described three-way ANOVA, in particular its special form called Latin squares, and also two-way ANOVA with a single observation in each subgroup. We explained the purpose of these methods and the mathematical technique behind it. The reader became familiar with the terms two-factor and three-factor analysis of variance, Latin squares, decomposition of total variability and ANOVA table. The following problems allow the reader to practise the methods, including one-way ANOVA from the previous chapter.

CONTROL TEST 7

1) The nivel of test being 5 per cent, test if parsley yields depend on type of fertilizer used. All the necessary observations are in table 41.

Table 41: Entry data for ANOVA

Fertilizer	**Yields (1kg/10m^2)**					
A	40	42	45	40	44	47
B	76	75	82	68		
C	60	58	62	64	70	

Source: author's

2) Evaluate intensity of dependence between the parsley yields and the type of fertilizer, using the determination ratio characteristic.

3) Six drivers were randomly selected, each of them experiencing a ride with different gasoline. Test whether gasoline consumption depends on type of gasoline used and/or driver. The nivel of test is 5%. The data are in table 42.

Table 42: Entry data for ANOVA

	Driver						
Gasoline	**A**	**B**	**C**	**D**	**E**	**F**	**Averages**
Aral	7,5	6,9	7,9	7,3	6,9	7,8	**7,38**
Shell	7,6	7,2	7,5	8	7,3	8,2	**7,63**
Benzina	7,2	8,1	7,8	7,6	7,8	6,9	**7,57**
Slovnaft	7	7,3	7,2	7,5	8,2	7,7	**7,48**
Averages	7,33	7,38	7,6	7,6	7,55	7,65	***7,5***

Source: author's

SOLUTIONS

1) There are three types of fertilizers, i.e. $k = 3$, and the corresponding samples of observations are of the size $n_1 = 6,\ \ n_2 = 4\ \ n_3 = 5$, respectively. The total number of observations is $N = 15$.

Tested is the hypothesis $H_0: \ \mu_1 = \mu_2 = \mu_3$, i.e. **parsley yields do not depend on type of fertilizer used.** To perform the test, we make the following preliminary calculations:

- *Conditional averages* $\overline{y}_i = \dfrac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$, for $i = 1, 2..., k$, where y_{ij} are observations.
- *Total average* $\overline{y} = \dfrac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \dfrac{1}{k}\sum_{i=1}^{k}\overline{y}_i$,
- *Among-group variability* $S_{y,m} = \sum_{i=1}^{k} n_i \left(\overline{y}_i - \overline{y}\right)^2$, where: n_i is the number of observations in the i-th group, $\overline{y}_i$ is the sample average in the i-th data group.
- *Within-group variability* $S_{y,v} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y}_i\right)^2$.
- *Total variability* $S_y = S_{y,m} + S_{y,v}$.

We get:

$$\overline{y}_1 = \frac{40 + 42 + ... + 47}{6} = 43,$$
$$\overline{y}_2 = \frac{76 + 75 + ... + 68}{4} = 75,25,$$
$$\overline{y}_3 = 62,8,$$

$$\bar{y} = \frac{43 + 75,25 + 62,8}{3} = 60,35.$$

$$S_{y,m} = \sum_{i=1}^{k} n_i \left(\bar{y}_i - \bar{y}\right)^2 = 6\left(43 - 60,35\right)^2 + 4\left(75,25 - 60,35\right)^2 + 5\left(62,8 - 60,35\right)^2 = 2724,188.$$

$$\begin{aligned} S_{y,v} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_i\right)^2 &= \left(40 - 43\right)^2 + \left(42 - 43\right)^2 + \ldots + \left(47 - 43\right)^2 + \\ &+ \left(76 - 75,25\right)^2 + \left(75 - 75,25\right)^2 + \ldots + \left(68 - 75,25\right)^2 + \\ &+ \left(60 - 62,8\right)^2 + \left(58 - 62,8\right)^2 + \ldots + \left(70 - 62,8\right)^2 = 223,55. \end{aligned}$$

The results are summarized in table 43:

Table 43: ANOVA output

Source of variability	**Sums of squares**	**Degrees of freedom**	**Averaged sums of squares**	**Test criterion F**
Factor *x* (among-group variability)	2724,188	$k - 1 = 2$	1362,1	73,12
Residual (within-group) variability	223,55	$N - k = 12$	18,63	
Total variability	2947,74	$N - 1 = 14$		

The test criterion $T = 73,12$, the critical value $F_{2,12}(0,05) = 3,89$, the critical region is $C = [3,89; +\infty)$. Since *T* belongs to the critical region, we reject the null hypothesis. **Parsley yields depend on type of fertilizer used.**

2) To answer the question: „How strong is the relationship between type of fertilizer used and parsley yields?“, we calculate the correlation ratio

$$P = \sqrt{\frac{S_{y,m}}{S_y}},$$

where

$S_{y,m}$ is the among-group variability,

S_y is the total variability.

We get $P = \sqrt{\frac{2724,188}{2947,74}} = \sqrt{0,92} = 0,96$.

Raising the result to the second power, we obtain the determination ratio $P^2 = 0,922$. A value of the determination ratio which is close to one signals a strong dependence of parsley yields on type of fertilizer used.

3) The table with entry data already contains the subgroup averages needed for testing dependence of fuel consumption Y on type of gasoline X_1 and driver X_2. There are four levels of X_1, $k = 4$, and six levels of X_2, $r = 6$. In the case of X_1, we test the null hypothesis: factor X_1 is not influential, versus the alternative hypthesis: X_1 is influential. The hypotheses for the second factor X_2 are analogous.

The individual sums of squares are:

$$S_{X_1} = r\sum_{i=1}^{4}\left(\overline{y}_{i.} - \overline{y}\right)^2 = 6\left[\left(7,38-7,5\right)^2 + ... + \left(7,48-7,5\right)^2\right] = 0,21.$$

$$S_{X_2} = k\sum_{j=1}^{6}\left(\overline{y}_{.j} - \overline{y}\right)^2 = 4\left[\left(7,33-7,5\right)^2 + ... + \left(7,65-7,5\right)^2\right] = 0,358.$$

The most straightforward way to calculate the residual sum of squares S_R is to calculate first the total variability S, and then evaluate the residual sum of squares from equation $S = S_R + S_{X_1} + S_{X_2}$.

$$S = \sum_{i=1}^{4}\sum_{j=1}^{6}\left(y_{ij} - \overline{y}\right)^2 = \left(7,5-7,5\right)^2 + \left(6,9-7,5\right)^2 + ... + \left(8,2-7,5\right)^2 + \left(7,7-7,5\right)^2 = 3,79.$$

Thus, $S_R = 3,22.$

The test criterion for the first factor is $T = \dfrac{\dfrac{S_{X_1}}{k-1}}{\dfrac{S_R}{(k-1)(r-1)}} = \dfrac{\dfrac{0,21}{3}}{\dfrac{3,22}{15}} = 0,33.$

The critical value is $F_{3,15}(0,05) = 3,29$. Since 0,33 < 3,29, we cannot reject the null hypothesis. Therefore, **type of gasoline used seems to have no effect on fuel consumption.**

The test criterion for the second factor is $T = \dfrac{\dfrac{S_{X_2}}{r-1}}{\dfrac{S_R}{(k-1)(r-1)}} = \dfrac{\dfrac{0,36}{5}}{\dfrac{3,22}{15}} = 0,33.$

The critical value is $F_{5,15}(0,05) = 2,9$. Since 0,34 < 2,9, we accept the null hypothesis again, therefore **type of person driving the car does not affect fuel consumption either.**

8 FULL FACTORIAL EXPERIMENTAL PLANS

This chapter covers the foundations of design of experiments, a branch of statistics dominated in industrial applications. To experiment means to change working conditions so that the best possible working procedures are found, and more knowledge is acquired about the product and related working process. The best possible working procedures can be viewed in the following sense: denoting Y the observed quality characteristic of a product (or we can work with more such characteristics $Y_1, Y_2, ..., Y_k$) and denoting A, B, C,… the factors which *potentially* affect the product, their levels being $A_1, A_2, A_3, ...$ for factor A, $B_1, B_2, B_3, ...$ for factor B and so on, design of experiments aims to determine which of the factors are influential and what their optimal levels are with respect to optimal levels of the quality characteristics.

8.1 FOUNDATIONS OF EXPERIMENTING AND ITS APPLICATIONS

In our context, experimenting means analysing various combinations of the levels of those factors that are thought to affect the observed quality characteristic of a product. The characteristic is a response, a result of an experiment. The response is related to a certain combination of the factor levels. Analysing the relations is how one can attain the objectives outlined in the previous paragraph.

The outlined objectives may be achieved in more than one way. Some of the appropriate procedures were already described in the previous chapters. For example, to determine which factor is influential, analysis of variance can be exploited for this purpose. Regression analysis could also be used to define suitable levels of the factors with respect to the process output, represented in the regression equation by the dependent variable. The problem with these methods is that they often require a lot of data. Also, the methods are based on a set of conditions that must be met for the particular method to work correctly. Some of the conditions are even impossible to verify with a prescribed high probability, such as the form of the regression model governing the relation between several variables. For these and other reasons, design of experiments as a separate branch originated decades ago, to help solve reasonably the drawbacks just explained. We shall start the description of the discipline by explaining the meaning of several elementary terms the discipline works with.

Factor: a parameter or independent variable affecting the observed quality characteristic of a product. We denote a factor with a capital letter, such as A, and its levels with the same letter with a lower index, such as A_1 (the first level of factor A).

Two elementary types of factors exist:

a) **regulated factor** is a variable thought to affect the observed quality characteristic. The levels of the variable can be set up and maintained, and it is desirable to do so.
b) **noise factor** is a factor that has an adverse effect on the quality characteristic. The levels of the factor cannot be set up and maintained during the experiment, or it is not desirable to do so.

Interaction of factors: a combined effect of two or more factors. In this case, the effect of one factor of the interaction generally depends on the effect of another factor of the interaction. The interaction of factors A and B is denoted as AB.

Design of experiments is applied in areas, such as: simulation, product design and development, process design and development, testing and validation, solving production quality problems, measurement system analysis and improvement.

8.2 EXPERIMENTAL PROCEDURE

The steps that must be taken to detect influential factors and their optimal levels form an experimental procedure. These steps are: planning the experiment (using brainstorming, for instance), designing the experiment, realizing the experiment, analysing the the results of the experiment.

Planning the experiment means to set up an experimental team in the first place. Representatives of the departments that design the product and the process leading to the product should be members of the team. The team should have 2-15 members. The team members ought to participate in brainstorming sessions which will determine what quality or output characteristics of the product will be monitored, and which factors should be considered together with their starting levels.

The planning phase results in defining the objective of the experiment, the objective being related to the product under scrutiny, and it should also define the characteristics of the product, based on which it will be possible to make a judgement on whether the objectives were achieved. Another result of the planning phase is a list of factors that could potentially play a role in defining the product quality.

These information are used in the second phase – in designing the experiment, which results in an experimental plan. Such a plan is a table consisting of individual experimental runs. Each run, which is represented by a row in the table, tells the experimenter the levels the factors should take on during the experimental run. Later, we shall demonstrate this procedure in an example.

Realizing the experiment takes place either separately at a laboratory or at the production line. The latter case is more challenging, of course, since the production capacity on the one hand, and the requirement to analyse the production through an experiment on the other hand, can lead to a clash of interests. Therefore, in the latter case, night shifts or weekends usually provide an opportunity to run the experiment outside laboratory.

Analysing the results of the experiment means to seek a combination of the input factors that will optimize (at least approximately) the observed characteristic of the product. In the final step, the optimal factor set-up is verified by subsequent experiments and/or simulations.

PROBLEM 1 (*Spring*)

This problem demonstrates how to set up a full factorial experimental plan, using coded variables.

What we are now interested in is how much pressure an industrial spring can withstand, the spring being compressed by a machine tool until it breaks down. The following factors are considered in the experiment.

L = length of the spring,
G = width of the spring,
T = material of the spring.

It is to be determined which factors influence the service life of the spring.

SOLUTION

Let us construct a table of the factor levels considered: we shall use two levels for each factor, therefore the corresponding experimental plan is called a *two-level plan* (there are also three-level plans).

Table 44: Factors and their levels

Factor	Symbol	Lower level	Upper level
		-	+
Spring length	*L*	10 cm	15 cm
Spring width	*G*	5 mm	7 mm
Spring material	*T*	*A*	*B*

Source: author's

There is more than one way how to build an experimental plan that will prescribe individual experimental runs. The so-called **full factorial plan** is one of the most frequently used schemes:

Table 45: Full factorial plan

Run	*L*	*G*	*T*	*Y*
1	10	5	A	
2	15	5	A	
3	10	7	A	
4	15	7	A	
5	10	5	B	
6	15	5	B	
7	10	7	B	
8	15	7	B	

We talk about a full factorial plan because such a plan contains all possible combinations of the factor levels. The symbol Y will be used to denote the result of an experiment, i.e. the response to a specific combination of the factor levels.

It is more suitable to prescribe an experimental plan, using the following symbols: the lower and the upper level of a factor is denoted -1 and +1, respectively. Table 45 then takes the form of table 46

Table 46: Full factorial experimental plan using coded variables

Run	*L*	*G*	*T*	*Y*
1	-1	-1	-1	
2	+1	-1	-1	
3	-1	+1	-1	
4	+1	+1	-1	
5	-1	-1	+1	
6	+1	-1	+1	
7	-1	+1	+1	
8	+1	+1	+1	

The conversion of the original variables to the coded variables, but not only their upper and lower levels as we do here, can be done according to equation

8-1
$$x_c = \frac{x_0 - \frac{x_{\max} + x_{\min}}{2}}{\frac{x_{\max} - x_{\min}}{2}},$$

where

x_0 is the variable in the original physical units,

x_c is the coded variable,

$x_{\max}$ is the upper level of x,

$x_{\min}$ is the lower level of x.

For instance, the conversion of factor *L* leads to

$$L_c = \frac{10 - \frac{15+10}{2}}{\frac{15-10}{2}} = -1,$$

or the upper level of factor *G*, $G = 7$, is

$$G_c = \frac{7 - \frac{7+5}{2}}{\frac{7-5}{2}} = +1.$$

A full factorial experimental plan containing *k* factors represents $n = 2^k$ experimental runs. For example, if $k = 3$ factors are considered in an experiment, there will be $n = 2^3 = 8$ experimental runs altogether. Therefore, there will be eight rows in the table of the experimental plan.

Experimental plan prescribes how and under which conditions to proceed with the experiment. After applying the plan, values of the observed variable *Y* resulting from the experiment can be recorded. In our example, each experimental run has been carried out twice, and the results are recorded in table 47.

Table 47: Full factorial plan and measurements

Run	Factor	Factor	Factor	Output	Output	Average
	L	*G*	*T*	Y_1	Y_2	$\bar{Y}$
1	-	-	-	77	81	79
2	+	-	-	98	96	97
3	-	+	-	76	74	75
4	+	+	-	90	94	92
5	-	-	+	63	65	64
6	+	-	+	82	86	84
7	-	+	+	72	74	73
8	+	+	+	92	88	90

Source: author's

Table 47 ends the preparatory and experimental work. What follows now are calculations which are to determine the influential factors. These factors may also include interactions *LG*, *LT*, *GT* and *LGT*, therefore interactions are usually included in experimental plans, as well. This can be done after the experiment ran its course. The levels of the interactions are calculated additionally by multiplying the elements of the table that lie in the same row and columns of those factors which form the interaction. For instance, to calculate the levels of the interaction *LG*, the signs, or ones with the corresponding sign, are taken from a given row and the columns marked „*L*" and „*G*", and then they are multiplied. Table 48 represents the result of this procedure.

Table 48: Full factorial plan and interactions

Run	*L*	*G*	*T*	*LG*	*LT*	*GT*	*LGT*
1	-	-	-	+	+	+	-
2	+	-	-	-	-	+	+
3	-	+	-	-	+	-	+
4	+	+	-	+	-	-	-
5	-	-	+	+	-	-	+
6	+	-	+	-	+	-	-
7	-	+	+	-	-	+	-
8	+	+	+	+	+	+	+

8.3 EFFECT OF A FACTOR AND ITS SIGNIFICANCE

Effect of a factor is the change in quality characteristic *Y*, induced by changing the level of that factor from -1 to +1. To calculate the effect, we shall use the *sign method* which, in the table of an experiment, multiplies each value from the column of *Y* by the corresponding

number one of the analysed factor (the one from the same row of the table), the multiples are summed together, and the sum is divided in the end by one half of the number of experimental runs (number of rows in the table). For instance, to calculate the effect of factor *L* in our example, we get

$$efekt(L) = \frac{1}{4}(-79 + 97 - 75 + 92 - 64 + 84 - 73 + 90) = 18.$$

The effect of *T* is

$$efekt(T) = \frac{1}{4}(-79 - 97 - 75 - 92 + 64 + 84 + 73 + 90) = -8.$$

Analogous procedure is used for other factors including their interactions:

$$efekt(LG) = \frac{1}{4}(79 - 97 - 75 + 92 + 64 - 84 - 73 + 90) = -1.$$

The effects of all the factors from our example are recorded in the last row of table 49.

Table 49: Full factorial plan and factor effects

Run	***L***	***G***	***T***	***LG***	***LT***	***GT***	***LGT***	$\bar{Y}$
1	-	-	-	+	+	+	-	79
2	+	-	-	-	-	+	+	97
3	-	+	-	-	+	-	+	75
4	+	+	-	+	-	-	-	92
5	-	-	+	+	-	-	+	64
6	+	-	+	-	+	-	-	84
7	-	+	+	-	-	+	-	73
8	+	+	+	+	+	+	+	90
Effect	**18**	**1,5**	**-1**	**-8**	**0,5**	**6**	**-0,5**	

Source: author's

Now, we would like to find out which of the effects is significant. To do that statistically, variance of the factor effect σ_e^2, the effect being a random variable since we work with a data sample, must be estimated. The variance is the same for all factor effects under suitable conditions:

8-2
$$\sigma_e^2 = \frac{4\sigma^2}{N},$$

where *N* is the number of experimental runs (including their repetitions if there are any). In our example, $N = 16$ as the table has eight rows, i.e. there are eight runs, and each run is implemented twice. If runs are repeated, we can calculate

8-3
$$s^2 = \frac{\upsilon_1 s_1^2 + ... + \upsilon_k s_k^2}{\upsilon_1 + ... + \upsilon_k},$$

where $\upsilon_i = n_i - 1$, n_i is the number of repetitions of the i –th experimental run, and s_i^2 is sample variance of Y corresponding to the i-th experimental run. Now, the estimate of 8-2 is

8-4
$$s_e^2 = \frac{4s^2}{N}.$$

8.3.1 STATISTICAL TEST OD FACTOR SIGNIFICANCE

Estimate 8-4 is used to test significance of factor effect:

1. The null hypothesis is $\mathbf{H_o}$: Factor effect is insignificant, the alternative is $\mathbf{H_1}$**:** Factor effect is significant.

2. The test criterion is of the form
$$t = \frac{efekt}{s_e}.$$

3. The critical value $K = t_{n_1+n_2+...+n_k-n}(\alpha)$, where $n_1,...,n_k$ are the numbers of repetitions of the first, second,…, k-th experimental run, respectively; in our example, $n_i = 2$ for each i; n is the number of experimental runs without the repetitions, i.e. the number of rows in the experimental plan. As we can see, the critical value concerns a Student's distribution.

4. If $|t| \geq t_{n_1+n_2+...+n_k-n}(\alpha)$, the null hypothesis is rejected, and the factor is considered significant. In the opposite case, the factor is regarded as insignificant.

PROBLEM 2

Let us return to the *Spring* example. We have:

$$t_{n_1+n_2+...+n_k-n}(\alpha) = t_{16-8}(0,05) = 2,306.$$

Using 8-3, we also get:

$$s^2 = \frac{8+2+2+8+2+8+2+8}{8} = 5,$$

and this result inserted in 8-4 leads to

$$s_e^2 = \frac{4s^2}{N} = \frac{4 \cdot 5}{16} = 1,25, \text{ tj. } s_e = 1,12.$$

This allows us to evaluate the test criteria for all the factors including interactions. The criteria are in table 50.

Table 50: Test of factor significance

Run	Y_1	Y_2	s_i^2	Effect	*t*
1	77	81	8		
2	98	96	2	L = 18	**16,07**
3	76	74	2	G = 1,5	**1,34**
4	90	94	8	LG = -1,0	**-0,89**
5	63	65	2	T = -8,0	**-7,14**
6	82	86	8	LT = 0,5	**0,45**
7	72	74	2	GT = 6,0	**5,36**
8	92	88	8	LGT =-0,5	**-0,45**

Source: author's

The critical value K = TINV(0,05,8) = 2,306, which exceeds in absolute value the test criterion for the factors *L*, *T* and *GT*. Therefore, the three factors are significant, the remaining factors are not, as far as their effect on the spring service life is concerned.

8.3.2 GRAPHICAL ASSESSMENT OF FACTOR SIGNIFICANCE

If the experimental runs are not repeated, the test above cannot be used. In such cases, a graphical method for detecting the influential factors exists. The method, as suggested by its name, is based on constructing a graph, on the horizontal axis of which factor effects are designated. The vertical axis of the graph records the following values:

8-5
$$P_i = \frac{100(i-0,5)}{m},$$

where i = 1,2, ..., m, m being the number of all the factors of the experiment including interactions. More precisely, the graph is a set of points [efect$_{(i)}$, P$_i$], where efect$_{(i)}$ is the i-th smallest effect among all the calculated effects. The points of the graph that seem to lie ouside the central line running through the middle section of the graph suggest the significant factors. If the graph takes the form of an S-curve, which is the case under certain conditions, then some of the points of the graph will turn away from the approximately linear middle section of the S-curve. Those are the points that signal which factors are influential.

When using the graphical method, it is convenient to set up a table similar to table 51 which contains the factor effects sorted in the ascending order.

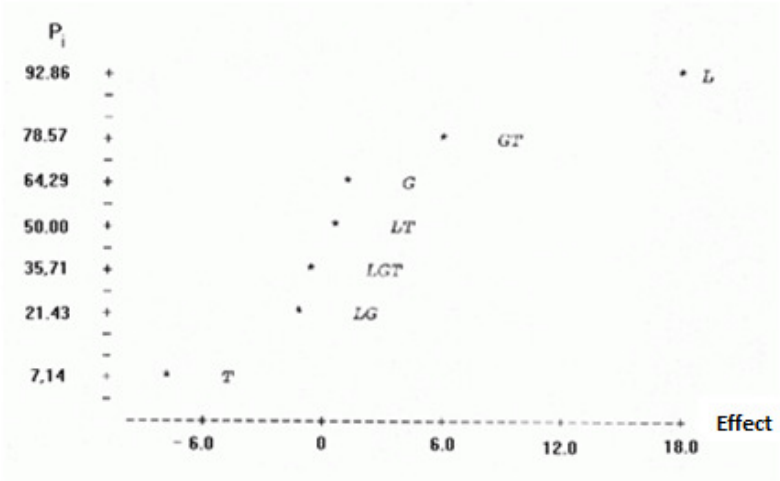
Table 51: Auxiliary data for graphical assessment of factor significance

Number	1	2	3	4	5	6	7
Effect	-8,0	-1,0	-0,5	0,5	1,5	6,0	18
Factor	T	LG	LGT	LT	G	GT	L
P_i	7,14	21,42	35,71	50	64,29	78,57	92,86

Source: author's

The second and the fourth row of table 51 are coordinates of the points that form the graph (figure 12).

Figure 12: Points determining significance of factors



The graph shows that it is the points of the factors the effects of which turn out to be significant that lie outside the line running through the middle section of the graph. These points relate to factors *L*, *T* and *GT*.

8.3.3 GRAPH OF INTERACTIONS

Significant interactions are usually accompanied by graphs that allow for a discussion on the optimal levels of the factors making up the interactions. For instance, the interaction *GT* can be accompanied by a graph which will outline the effect of the factor *G* on *Y*, depending on the level of the factor *T*. To do that, we can scrutinize the full experimental plan, and select the values of *Y* from the plan, corresponding to different levels of *G* and *T*. Table 52 contains these values, together with the average value of *Y*.

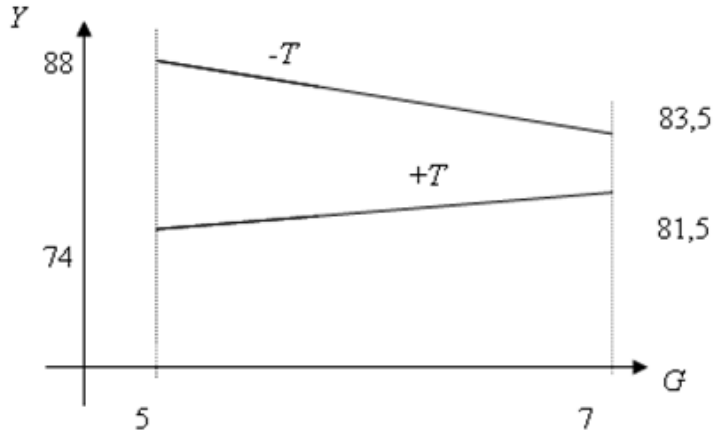
Table 52: Responses of *Y* to various levels of *G* and *T*

G	***T***	**Response 1**	**Response 2**	**Average *Y***
-	-	79	97	88
+	-	75	92	83,5

G	***T***	**Response 1**	**Response 2**	**Average *Y***
-	+	64	84	74
+	+	73	90	81,5

Connecting the points $[5, 88], [7, 83,5]$, we get the average response of *Y* to a change in *G* provided the factor *T* is fixed at its lower level. Connecting the points $[5, 74], [7, 81,5]$, we get the average response of *Y* to a change in *G* provided the factor *T* is fixed at its upper level (see figure 13).

Figure 13: Interaction *GT* and its effect on *Y*



As suggested by the graph, to maximize *Y*, for instance, it is best to have *T* at its lower level. We also see that the interaction has a certain effect on *Y* because the slope of the line, describing the effect of *G*, changes with a change in the level of *T*.

8.4 REGRESSION MODEL OF THE 2^3 EXPERIMENT

Once the effects of main factors and their interactions are detected, it is possible to construct a regression model of the experiment. The model will describe dependence of *Y* on the factors. Incomplete quadratic model of the 2^3 experiment, which works with factors *A*, *B*, *C*, is of the form

8-6 $$\hat{y} = b_0 + b_1 A + b_2 B + b_3 C + b_{12} AB + b_{13} AC + b_{23} BC + b_{123} ABC\ .$$

It is a model containing all major factors and their interactions as explanatory variables, but not the quadratic terms A^2, B^2,… of the major factors. Each of the coefficients $b_1, b_2, ..., b_{123}$ can be calculated as one half of the effect of the factor the coefficient belongs to in the model. The absolute term of the model is $b_0 = \overline{Y}$. These are exactly the values one would get by applying the least squares method to the matrix of regressors, represented by the full experimental plan.

In our *Spring* example, we have

$$\hat{y} = 81{,}75 + 9L - 4T + 3GT.$$

There are many reasons why such a model is constructed. It is constructed

1. to determine local minima/maxima of the factors involved,

2. to determine the direction of the so-called dynamic experimental planning, which shifts the experiment to a new subset in the domain of $\hat{y}$, the new subset serving as an area for a new experiment. The shift is usually carried out, using the gradient of the model found.
3. to make local predictions of the quality characteristic $\hat{y}$.

SUMMARY

The reader has been acquainted with foundations of design of experiments in this chapter. It is possible to create a full experimental plan, or a fractional/partial experimental plan which will be analysed in the next chapter. Fractional plans are constructed if conditions of the experiment are such that not all experimental runs can be performed (the whole experiment would be too costly, for instance). Each factor participating in the experiment may potentially affect *Y*, a variable of interest. An effect is present if a change in the level of the factor leads to a change in *Y*. The presence of the effect can be verified either by a graphical method or a statistical test. Such procedures were described in the chapter.

Planned experiments follow an experimental plan which prescribes individual experimental runs to be carried out and the order of the runs in which they are to be realized. There are two terms that must be distinguished in connection with experimenting: *experimental run*, which leads to measurements of *Y*, the variable of interest, under a *given* set of conditions; these conditions are represented by a specific row in the experimental plan; *experiment*, which is the set of all experimental runs.

The aim of experimental planning is to determine which factors have a statistically significant effect on a quality characteristic *Y*, and to determine the optimal level of the significant factors so that the variable *Y* is optimized and/or stabilized. Stability of *Y* means that the variable remains optimal or close to its optimal state under various conditions (environment, product treatment, etc.). If this is the case, we talk about product robustness.

The following terms were described in the chapter: experimental plan, experiment, experimental run, factor effect, test of factor significance, model of experiment.

The following examples provide further assistance in the study of this subject matter.

PROBLEM 3

A full experimental plan was constructed for two factors *A* and *B*. Each experimental run was realized twice. The result of the experiment is in table 53.

Table 53: Result of an experiment with two factors

A	*B*	Y_1	Y_2
-	-	5	6
+	-	5	5
-	+	7	6
+	+	5	4

Source: author's

Calculate:

a. Effects of the factors *A*, *B* and *AB*,
b. Write the equation of the incomplete quadratic model for this experiment,
c. Estimate variance of the factor effects,
d. Test significance of *A*, *B* and *AB* (the nivel of test is 5%).

SOLUTION

a. First, we shall attach average responses of *Y* to table 53. By doing so, we get table 54.

Table 54: Average response to different combinations of factor levels

A	*B*	Y_1	Y_2	$\overline{Y}$
-	-	5	6	5,5
+	-	5	5	5
-	+	7	6	6,5
+	+	5	4	4,5

The effects are:

$$e_A = \frac{1}{2}(-5,5+5-6,5+4,5) = -1,25,$$
$$e_B = \frac{1}{2}(-5,5-5+6,5+4,5) = 0,25,$$
$$e_{AB} = \frac{1}{2}(5,5-5-6,5+4,5) = -0,75.$$

b. The equation of the model is:

$$\hat{y} = 5,375 - \frac{1,25}{2}A + \frac{0,25}{2}B - \frac{0,75}{2}AB.$$

c. The estimated variance is:

$$s_e^2 = \frac{4s^2}{N},$$

where

$$s^2 = \frac{0,25+0+0,25+0,25}{4} = 0,1875.$$

Thus,

$$s_e^2 = \frac{4 \cdot 0,1875}{8} = 0,094.$$

The estimated standard deviation is $s_e = 0,31$.

d. We test the hypotheses:

H_0 : Factor (effect) is not statistically significant;

H_1 : Factor (effect) is statistically significant.

The test criterion is $t = \frac{efekt}{s_e}$. We get:

$$t_A = -4,03\,,\; t_B = 0,8\,,\; t_{AB} = -2,41\,.$$

The test criteria are compared to the following critical value:

$$t_{N-n}(0,05) = t_{8-4}(0,05) = 2,776\,,$$

where *N* is the number of all experimental runs including their repetitions, and *n* is the number of experimental runs excluding their repetitions.

Since $|-4,03| > 2,776, |0,8| < 2,776, |-2,41| < 2,776$, *A* is significant, the other factors are not.

PROBLEM 4

Using the graphical method, find out which factor is significant. The factor effects and other data are contained in table 55.

Table 55: Entry data for graphical assessment of factor significance

i	1	2	3	4	5	6	7
Effect	**-8**	**-1**	**-0,5**	**0,5**	**1,5**	**6**	**18**
Factor	*C*	*AB*	*ABC*	*AC*	*B*	*BC*	*A*
P_i							

Source: author's

SOLUTION

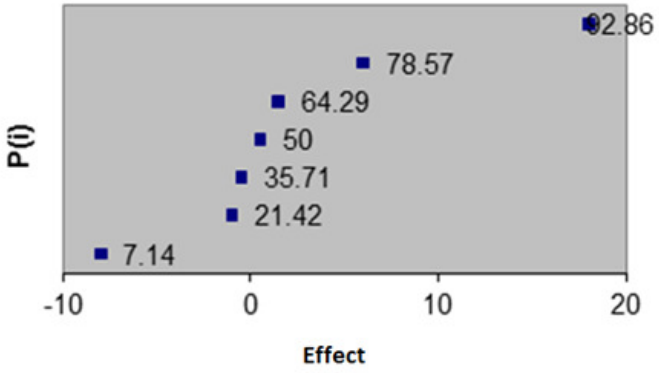
Expanding table 55 by calculating $P_i = \frac{100(i-0,5)}{m}$, $i = 1,2,...,m$, m = the number of all factors including their interactions, i.e. $m = 7$ in this case, we have:

Table 56: P_i's for graphical evaluation of factor effects

i	1	2	3	4	5	6	7
Effect	**-8**	**-1**	**-0,5**	**0,5**	**1,5**	**6**	**18**
Factor	*C*	*AB*	*ABC*	*AC*	*B*	*BC*	*A*
P_i	**7,14**	**21,42**	**35,71**	**50**	**64,29**	**78,57**	**92,86**

The resulting graph is:

Figure 14: Graphical evaluation of effects



The central line running through the middle section of the graph does not seem to contain the „points“: 92,86; 7,14; 78,57. This suggests that the factors *A*, *C* and *BC* are influential.

CONTROL TEST 8

Yes/No answers:

8.1 Experimental plan defines the order in which experimental runs are carried out?
8.2 A full plan working with 4 major factors consists of 8 experimental runs?
8.3 Factor effect can take on positive values only?
8.4 When testing factor significance, the corresponding critical value is related to a Fisher's distribution?
8.5 When testing factor significance with the graphical method, those factors which lie outside the central line of the graph are regarded as significant?

Complete the statement:

8.6 Experiment is a system of _____
8.7 A full plan with three major factors has _____experimental runs.
8.8 The null hypothesis of the test of factor significance is: _____
8.9 The graphical method of testing factor significance is used when _____ is/are not available.
8.10 The graph constructed for testing factor significance requires calculation of P(i) which is given by the formula _____
8.11 Complete the table so that it represents a full experimental plan:

Run	*A*	*B*
1		
2		
3		
4		

8.12 The table below represents a full plan for factors *A* and *B*. Each experimental run has been realized twice.

A	*B*	Y_1	Y_2
-	-	2,3	2,6
+	-	3,1	2,9
-	+	3	3,5
+	+	1,9	2,2

Source: author's

Calculate:

a. the effect of the factors *A*, *B* and *AB*,
b. the model of the experiment,
c. the estimate of the factor effect variance.

8.13 Test whether the effects of *A*, *B* and *AB* from 8.12 are significant (nivel of test = 5%).
8.14 Draw the graph of the interaction *AB* from 8.12. Depict the effect of *A* on *Y*, depending on the level of *B*. What level of *B* maximizes *Y*?

SOLUTIONS

8.1 yes
8.2 no
8.3 no
8.4 no
8.5 yes
8.6 runs
8.7 $2^3 = 8$
8.8 insignificant
8.9 repetition of individual runs
8.10. $P_i = \frac{100(i-0,5)}{m}$, where $i = 1,2,...,m$ and *m* is the number of all the factors.
8.11.

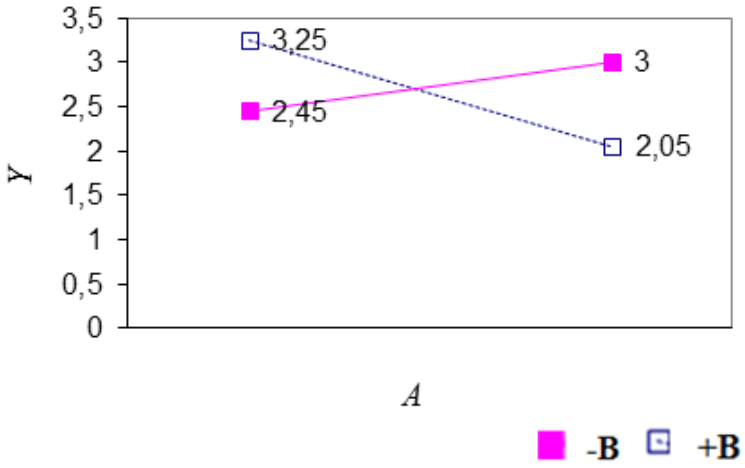
Run	*A*	*B*
1	-	-
2	+	-
3	-	+
4	+	+

8.12.

a. effect(*A*) = – 0,325; effect(*B*) = – 0,075; effect(*AB*)= – 0,875.
b. $Y = 2,69 - 0,1625A - 0,0375B - 0,4375AB$.
c. $s^2 = 0,029;\ s_e = 0,12.$

8.13. Only *AB* is significant, since $|-7,29| > 2,776$.

8.14. The graph shows that the maximal value of *Y* is achieved for the upper level of *B*, and the interaction proves to have an effect on *Y*, as the two lines resemble a cross.



9 TWO-LEVEL FRACTIONAL PLAN

The previous chapter presented full factorial plans. It is, however, not always possible to construct such a plan, for different reasons including financial limitations. In such cases, fractional plans are used instead. Fractional plan is a partial plan which does not work with all possible combinations of the factors observed. As we shall learn in this chapter, different degrees of full plan reduction exist, the so-called half plans being among them. It is also the half plans we shall devote to, in particular.

Full factorial plans prescribe experimental runs for each factor, whereas fractional plans set down runs only for a subset of the factors, for *main factors*, and the remaining factors (secondary factors) are calculated as combinations of the main factors. The calculation is later used to define runs for the secondary factors. In this way, the total number of experimental runs can be reduced, which leads to a fractional plan.

If 2^k denotes the number of experimental runs of a full experiment, k being the number of its factors excluding interactions, then 2^{k-p} denotes the corresponding fractional plan, p being the degree of the reduction in this case. For instance, if we want to cut down 128 runs of a 2^7 full experiment by one half to

$$\frac{2^7}{2} = 2^{7-1},$$

we get a half plan with $n = 2^{7-1} = 64$ runs. This is the smallest possible degree of reduction of the full plan. Plans originating from full plans by applying the smallest degree of reduction are called *half plans*.

The degree of reduction p can be greater than one, such as four, leading to a 2^{7-4} fractional plan, which will have $n = 8$ runs. If the number of factors exluding interactions is $k = 7$, the degree of reduction equal to 4 would be the highest possible, based on the rule that the number of runs should be at least as high as the number of all factors. In the opposite case, too few data is available to perform a meaningful analysis. To give another example, if $k = 15$, the highest possible degree of reduction is 11 since $2^{15-11} = 16$. The degree of reduction equal to 12 would lead to $2^{15-12} = 8$, which is too small a number. The degrees of reduction between two and the highest possible degree define *central plans*. For example, 2^{7-1} and 2^{7-4} belong to this category.

To sum up, fractional factorial plans can be divided into

a. Half plans, representing the smallest degree of reduction of full plans,
b. Plans, representing the highest possible degree of reduction of full plans,
c. Central plans.

Let us take a closer look at the half plans now.

9.1 HALF PLANS

We shall demonstrate in an example the effect of reducing a full plan by one half. To understand better what is going on, however, we need to define other fundamental terms from the theory of experimental plans. We do so now.

We shall use the symbol I to denote the factor whose column in the experimental plan contains only ones. We call this variable *identity factor*. We also define multiplication of two factors: such an algebraic operation applied to two factors yields as its result a factor whose column in the experimental plan contains values obtained by multiplying the ones from the corresponding rows of the plan and from the columns of the two factors that appear in the multiplication. The multiplication possesses the following basic algebraic properties:

$$A.A = I$$
$$A.I = I.A = A$$
$$(A.B).C = A.(B.C)$$
$$A.B = B.A$$

Suppose now that A, B, C, D, E are factors for which a half plan is to be constructed. To do so, one must select four of these factors (the main factors) for which the full plan will be created. These can be, for instance, the factors A, B, C, D. The remaining factor(s), the factor E in this case, will be defined as a combination (multiplication) of the main factors: let us define $E = ABCD$. In this way, instead of working with the 2^5 full plan, we shall work with the 2^{5-1} half plan. Only one half of all two-level combinations of all the factors is set up before the experiment is carried out, whereas the remaining two-level combinations are assigned by the multiplication.

Now, not every combination of the factors is appropriate. Every combination forms a *word*. Such a word consists of letters. Number of letters defines the length of the word. The equality $E = ABCD$ is called the *plan generator*. The 2^{k-p} factorial plans contain p generators. Since

$$E.E = E.\ ABCD,$$

we get

$$I = ABCDE.$$

Words that yield the indentity factor I are called *defining equations*. The shortest word among the defining equations is the *resolution of the plan*. The length of the shortest word is designated by the corresponding Roman figure in the symbol of the plan: for instance, 2_V^{5-1} applies to our example. Using the defining equations, we can find factor pairs (interactions, in general) with the same columns in the experimental plan. Such pairs are called interchangeable, and they play a role in the entire data analysis based on experimental planning. To illustrate the idea, if the plan generator is $E = ABCD$, the defining equation is $I = ABCDE$. In this case, an interchangeable pair for the interaction DE, for instance, is obtained by multiplying the defining equation by DE:

$$I = ABCDE \quad /.\ DE$$
$$DE.I = DE.ABCDE.$$

Hence,

$$DE = ABC.$$

The two interactions have the same column of ones in the experimental plan. The following problem shows the principles of working with half plans.

PROBLEM 1 (Dyestuff)

The amount of dyestuff *Y* left in a piece of fabric is observed. The amount depends on five factors: *A = pH, B = temperature, C = concentration of the solution used for tinting, D = finishing temperature of tinting, E = finishing time.* We are to construct a half experimental plan, and use it to detect influential factors. The necessary entry data is in table 57.

Table 57: Entry data for problem 1

Factor	Symbol	-1	+1
pH	*A*	4,5	5,5
temperature	*B*	$70^0\,C$	$80^0\,C$
concentrationn	*C*	1 g/l	3 g/l
finishing temp.	*D*	$170^0\,C$	$190^0\,C$
finishing time	*E*	50s.	70s.

Source: author's

SOLUTION

The main factors are chosen to be the factors *A*, *B*, *C*, *D*, while *E* is selected to be the secondary factor. Based on the half plan, constructed as described above and depicted by table 58, the experiment is performed, leading to the following values of *Y*:

Table 58: The output of the experiment

Run	*A*	*B*	*C*	*D*	*E =ABCD*	*Y*
1	-	-	-	-	+	6,4
2	+	-	-	-	-	9,9
3	-	+	-	-	-	8,1
4	+	+	-	-	+	6,6
5	-	-	+	-	-	9,0
6	+	-	+	-	+	5,3
7	-	+	+	-	+	-5,1
8	+	+	+	-	-	-1,0
9	-	-	-	+	-	10,6
10	+	-	-	+	+	12,7
11	-	+	-	+	+	12,9
12	+	+	-	+	-	11,2
13	-	-	+	+	+	2,4
14	+	-	+	+	-	9,7
15	-	+	+	+	-	4,1
16	+	+	+	+	+	4,0

The factor effects are calculated here the same way as for full experimental plans. For instance, the effect of *D* is

$$efekt(D) = \frac{1}{8}\left(-6,4 - 9,9 - ... - (-1) + 10,6 + 12,7 + ... + 4\right) = 4,8\,.$$

The effects of other factors are obtained similarly. However, since we work with a half plan, interchangeable pairs exist in this case. As we said before, such pairs are factors that have the same column of ones in the experimental plan. The effect calculated for one of these factors does not belong to that factor alone anymore! Now it represents the influence of all the interchangeable factors together. This is illustrated in table 59. If the effect was not attributed to all the interchangeable factors, the effect would lose its original interpretation it had in the case of the full experimental plan.

Table 59: Effects in a half plan

Factor	Effect
A + BCDE	0,0
B + ACDE	-4,4
C + ABDE	-5,0
D + ABDE	4,8
E + ABCD	-0,8
AB + CDE	0,2
AC + BDE	-0,6
AD + BCE	-0,6
AE + BCD	0,5
BC + ADE	-4,2
BD + ACD	1,1
BE + ACD	-0,2
CD + ABE	0,7
CE + ABD	-0,5
DE + ABC	2,4

For instance, the first zero effect represents the collective influence of the factors *A* and *BCDE*, the interaction which is interchangeable for *A*.

9.2 GRAPHICAL EVALUATION OF FACTOR EFFECT

The graphical method we used to verify significance of factor effects in the case of full experimental plans can be exploited for half plans, as well. This means we calculate the P_i's:

Table 60: Sorted effects and their P_i's

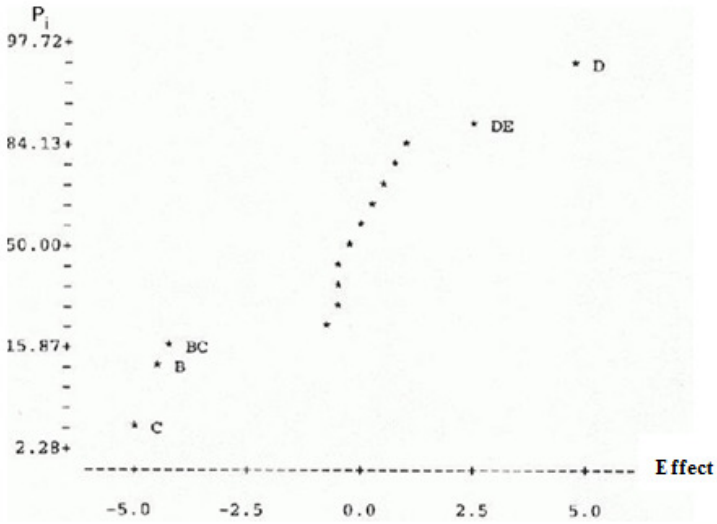
i	1	2	3	4	5	6
Factor	*C + ABDE*	*B + ACDE*	*BC + ADE*	*E + ABCD*	*AD + BCE*	*AC + BDE*
Effect	-5	-4,4	-4,2	-0,8	-0,6	-0,6
P_i	3,3	10	16,6	23,3	30	36,6

i	7	8	9	10	11	12
Factor	*CE + ABD*	*BE + ACD*	*A + BCDE*	*AB + CDE*	*AE + BCD*	*CD + ABE*
Effect	-0,5	-0,2	-0,0	0,2	0,5	0,7
P_i	43,3	50	56	63,3	70	76,6

i	13	14	15
Factor	*BD + ACD*	*DE + ABC*	*D + ABCE*
Effect	1,1	2,4	4,8
P_i	83,3	90	96,6

Once the P_i's are calculated, we construct the familiar graph (figure 15).

Figure 15: Graphical evaluation of factor effect significance



As we can see, the half plan brought us results similar to those obtained with the full experimental plan. This means we get similar results with fewer experimental runs. However, the results are not generally the same, and a certain amount of information has been lost after all, because the calculated effect belongs to the combined influence of several factors, such as *A + BCDE*, *B + ACDE* etc. The fact that a particular effect belongs to *A + BCDE* does not mean that exactly one half of the effect is caused by *A* and the other half is caused by *BCDE*! Generally speaking, it is not known what part of the total effect belongs to *A* or *BCDE* in this case. None the less, it is known that the longer the word representing the interaction/factor, the smaller its contribution to the total effect. Therefore, it is advisable that the fractional plans be generated in such a way that the interchangeable pairs are very long interactions.

PROBLEM 2

Let us have five factors *A*, *B*, *C*, *D*, *E*, where factor *E* is to be generated as a secondary factor. There is more than one way of generating *E*. Let us compare the consequences of two scenarios:

a. $E = AB$,
b. $E = ABCD$.

SOLUTION

The defining equations are:

a. $I = ABE$
b. $I = ABCDE$

In a), we have a 2_{III}^{5-1} plan, whereas in b), we have a 2_{V}^{5-1} plan. The plan b) is better because its resolution is *V*, which leads to the following discovery: when seeking the interchangeable pairs for *A*, for instance, we have

a. $A = BE$
b. $A = BCDE$

In the second case, the interchangeable pair has more factors (it is represented by a longer word), which results in the interaction BCDE contributing less to the total effect of the factor A+BCDE. Interchangeable interactions, represented by words with at least three letters, account for such a small part of the total effect that their contribution to the total effect is often neglected. This, of course, facilitates the interpretation of the final effect.

SUMMARY

We have learnt how it is possible to set up a fractional plan – it is the full plan constructed for a subset of the set of all factors, while the levels of the remaining one-letter factors are generated (calculated). This procedure reduces the total number of experimental runs, making the total experiment cheaper and faster. We've also learnt that the graphical method used to detect influential factors of a full plan can be exploited for the same purpose in the case of half plans. Finally, we explained that it is important to select a proper generation of the levels of the secondary factors. A proper generator leads to favourable interchangeable pairs for each factor, facilitating the conclusion over how large effects belong to each factor.

What follows is a set of illustrative examples.

PROBLEM 3

A half plan was constructed for factors *A*, *B*, *C* and *D*:

- **a.** Complete the table below,
- **b.** Using the graphical method, determine which factor is significant.

The necessary data are in table 61.

Table 61: Effects and their probabilities

Factor	Effect	i	P_i
$A + BCD$	1	3	35,7143
$B + ACD$	-0,5	2	21,4286
$C + ABD$	-4	1	
$D + ABC$	3	4	50
$AB + CD$	9	6	78,5714
$AC + BD$	6	5	64,2857
$AD + BC$	17	7	92,8571

Source: author's

SOLUTION

a) Using the equation

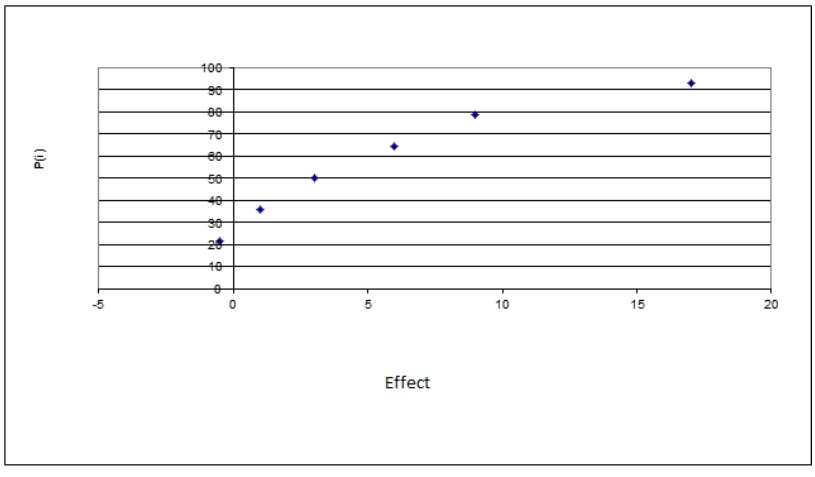
$$P_i = \frac{100(i-0,5)}{m},$$

we get

$$P_1 = \frac{100(1-0,5)}{7} = 7,14.$$

b) The following graph 16 implies that the interactions AD and BC and also the factor C could be considered significant.

Figure 16: Graphical evaluation of factor significance



PROBLEM 4

A half plan has been constructed for factors A, B, C, D, the generator of the plan being $D = ABC$. The output of the experiment is in table 62.

a. Calculate effects of the factors including the three-factor and four-factor interactions.
b. Using the interactions from a), estimate variance of the factor effects (this is a new procedure to be explained!)
c. Write the defining equation and interchangeable pairs.
d. Assess graphically the factor effects.

Table 62: Experimental output

A	*B*	*C*	*D = ABC*	*Y*	*ABC*	*ABD*	*BCD*	*ACD*	*ABCD*
-	-	-	-	77					
+	-	-	+	67					
-	+	-	+	64					
+	+	-	-	51					
-	-	+	+	64					
+	-	+	-	53					
-	+	+	-	73					
+	+	+	+	67					

Source: author's

SOLUTION

a) Inserting the remaining signs + and – (or plus ones and minus ones) in the table, we get the effects:

$$e_A = \frac{1}{4}(-77+67-64+51-64+53-73+67) = -10 = e_{BCD};$$
$$e_B = -1,5 = e_{ACD};\ e_C = -0,5 = e_{ABD};\ e_D = 2 = e_{ABC};\ e_{ABCD} = 129.$$

b) If there is only one measurement (experimental output) for each combination of the factors (for each row of the table), s^2 needed to estimate the variance is calculated as the average of the second powers of the effects belonging to the longest interactions:

$$s^2 = \frac{2^2 + (-0,5)^2 + (-10)^2 + (-1,5)^2 + 129^2}{5} = 3349,5$$
$$s_e^2 = \frac{4.s^2}{n} = 1674,75 \quad \Rightarrow \quad s_e = 40,9.$$

c) Since $D = ABC$, the defining equation is

$$I = ABCD.$$

The interchangeable pairs are *AB*, *CD*; *AC*, *BD*; *AD*, *BC*.

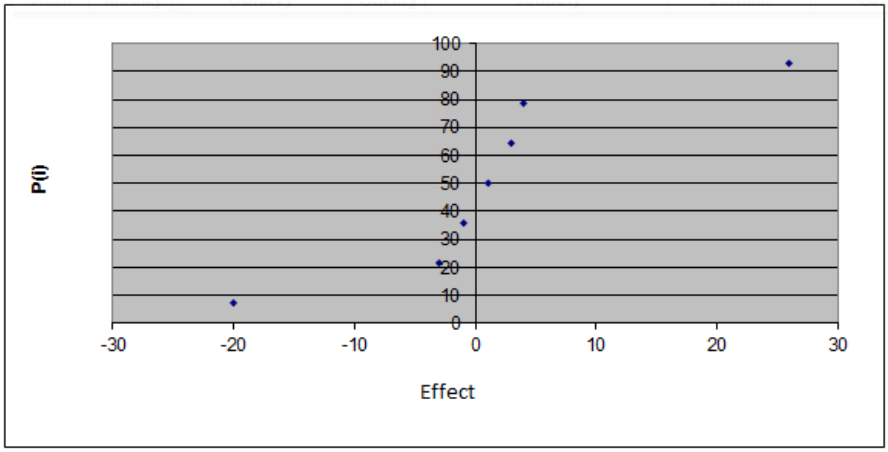
d) Calculating the P_i's, as in table 63,

Table 63: Factor effects and their P_i's

Factor	Effect	*i*	P_i
A + BCD	-20	1	7,14
B + ACD	-3	2	21,42
C + ABD	-1	3	35,71
D + ABC	4	6	78,57
AB + CD	1	4	50
AC + BD	3	5	64,28
AD + BC	26	7	92,85

we construct the graph (figure 17).

Figure 17: Graphical evaluation of factor effects



The graph shows that the effects of *AD + BC* and *A+BCD* lie outside the central line. These effects are deemed significant. In the latter case, we can restrict ourselves to the factor *A*, since the part of the total effect of *A+BCD* belonging to *BCD* will be small enough to be neglected.

CONTROL TEST 9

Yes/No answers:

9.1 Factorial plans contain all combinations of factor levels?

9.2 The commutative property of multiplication known from the theory of real numbers does not hold true in the case of factor multiplication?

9.3 If the generator of the plan is *D = BC*, the defining equation is *I=BCD*?

9.4 If interactions *ABC* and *DE* have the same columns in the experimental plan, the effect calculated from one of these columns belongs to both interactions?

9.5 When a half plan is constructed, one of the factors is defined as an interaction of other factors?

Complete the statement:

9.6 Secondary factors are expressed as a _____ of the main factors.

9.7 Factorial plans can be divided into _____plans, _____plans and _____ plans.

9.8 Given a factor *A*, the following holds: *AI* = *IA* = ____, where *I* is the identity factor.

9.9 Two factors with the same column of ones in an experimental plan are called_____.

9.10 In full plan, effects of factors and their interactions _____ the same as in the related half plan.

9.11 Construct the half plan for factors *A*, *B*, *C*, *D*, generated by *B*=*ACD*. Calculate the effect of *C* provided the following two measurements for each factor combination were obtained from the experiment:

The first series of measurements: 10,11,14,12,12,10,13,14,
The second series of measurements: 11,12,12,8,14,12,13,14.

9.12 A half plan was used for factors *A, B, C, D*.

a) Complete the table below,
b) Detect significant factors with the graphical method.

	Effects	*i*	P_i
A + *BCD*	1		
B + *ACD*	-8		
C + *ABD*	-10		
D + *ABC*	4		
AB + *CD*	9		
AC + *BD*	7		
AD + *BC*	5		

Source: author's

SOLUTIONS

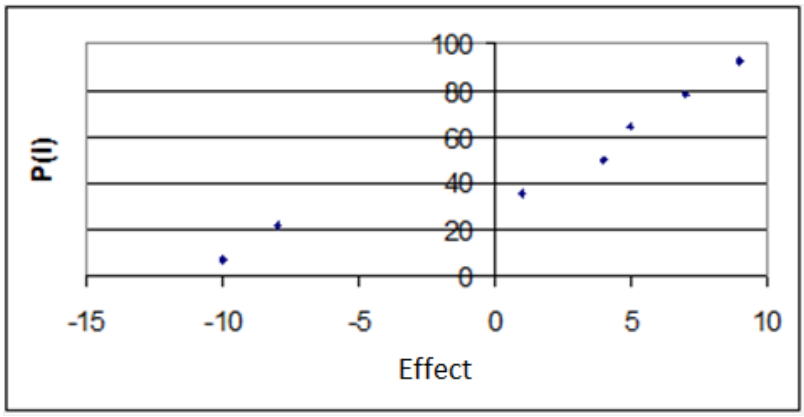
9.1 no
9.2 no
9.3 yes
9.4 yes
9.5 not necessarily
9.6 combination
9.7 half plans, central plans, highest-reduction plans (saturated plans)
9.8 *A*
9.9 interchangeable
9.10 are not
9.11 $e_C = 1$

9.12 a.

	Effects	i	P_i
$A + BCD$	1	3	35,71
$B + ACD$	-8	2	21,42
$C + ABD$	-10	1	7,14
$D + ABC$	4	4	50
$AB + CD$	9	7	92,85
$AC + BD$	7	6	78,57
$AD + BC$	5	5	64,28

b. Significant factors: *B, C* (figure 18):

Figure 18: Graphical evaluation of factor effects



10 TAGUCHI'S METHODS – LOSS FUNCTIONS

Taguchi's methods, based on the research by Genichi Taguchi, include online methods used during production, and offline methods reserved for pre-production stages. The former is the contents of the chapters 10 and 11, and it relies heavily on *loss functions*. In this chapter, we shall explain the logic behind these functions and the way they are constructed and used. The first part of the chapter defines loss function, and presents its properties. The second part of the chapter works with different kinds of loss functions, which is related to different kinds of what is called tolerance interval. The end of the chapter presents examples and a control test with questions and answers.

Taguchi's methods based on loss functions try to measure financial losses experienced by product users due to producers' inability to fabricate a product that would precisely comply with users' demands. Most often there is always at least a slight imprecision in the production due to its physical nature, no matter how much the production is surveilled and controlled.

Introduction of loss functions brought a new concept to how problems with quality are viewed. Earlier, the standard approach had been such that as long as the observed quality characteristic of a product lied within a tolerance interval, the characteristic not necessarily being equal to its desired optimal value, the product users would not bear any losses incurred by quality imprecisions. Taguchi disagreed with this view of the problem, and introduced simple mathematical functions that suprisingly turned out to be precise enough to measure the losses that occur even when the slightest deviation of the product quality characteristic from its optimal level exists.

Let us emphasize that the loss-function approach is only one of many forms of looking at the process or product. Whereas loss functions quantify the process quality, another question is how to improve the quality if it is detected to be inadequate, for example by a loss function. There are many ways how to solve problems within a process: *Six Sigma* methodology is one technique based on statistical methods (regression, in particular); analyzing process by *simulation* (Zgodavová and Bober, 2012) is another technique which may be used after process quality characteristics are properly measured or quantified (Zgodavová, 2010), and key concepts are exactly defined (Brannmark et al., 2012). The objective of process analyses is to create an optimal or close-to-optimal process set-up, and keep it that way, regardless of variations of factors which could destabilize the optimal set-up, i.e. keep the process robust (Siva, 2012).

10.1 DEFINITION AND PROPERTIES OF LOSS FUNCTIONS

Before defining the functions and presenting their graphs and properties, let us mention some fundamental conditions which are considered to be met in order for the functions to be used correctly and properly:

1. Every product bears a certain quality characteristic (such as its size, weight, mechanical property, etc.), and the quality of the whole product is judged based on that particular quality characteristic.
2. A target (optimal) value *T* is given for the quality characteristic from 1).
3. Lack of product quality is measured by deviation of the observed product quality characteristic from its target value *T*.
4. Any deviation of the characteristic from *T* brings a financial loss that the product user must bear because of the necessity to increase expenses on the product maintenance, repairments, etc.

One of the simpler loss functions is of the form

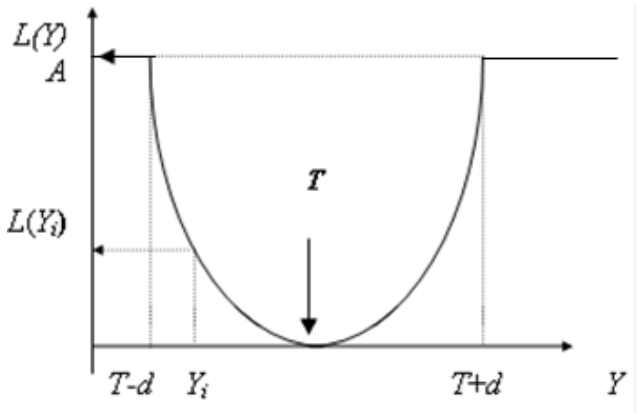
10-1
$$L(Y) = k(Y-T)^2 \text{ for } Y \in (T-d, T+d),$$
$$= A \text{ otherwise,}$$

where

T = target value of the quality characteristic,
d = tolerance
A = maximal loss due to poor quality
Y = truly achieved value of the quality characteristic (which is a random variable)
$L(Y)$ = financial loss given by the specific value of Y
k = constant to be determined.

Figure 19 shows the loss function just described. Above the tolerance interval $(T-d, T+d)$, it is a parabole, whereas outside the interval, it is a constant function.

Figure 19: Loss function



If $Y \le T-d$ or $Y \ge T+d$, in other words, if $d \le |Y-T|$, then $L(Y) = A$.

We can write:

10-2
$$A = kd^2.$$

Since d and A are usually known, 10-2 is used to determine k: $k = A / d^2$.

PROBLEM 1

Write the loss function equation for $d = 5$ and $A = 2$.

SOLUTION

We have $k = 2/5^2 = 0,08$, therefore $L(Y) = 0,08(Y-T)^2$.

The variable Y is considered to be a random variable, usually following approximately a normal distribution $N(E(Y),\sigma^2)$. We are often more interested in the average loss $E(L)$ rather than the individual loss. The average loss is calculated according to:

10-3
$$E(L) = E\left[k(Y-T)^2\right] = kE(Y-T)^2 = k\sigma^2,$$

Provided that $E(Y) = T$. The symbol σ^2 denotes the variance of Y, as usual.

However, if $E(Y) \neq T$, then $E(L) = k\sigma^2 + k(E(Y)-T)^2$.

Therefore, several equations are used in connection with loss functions:

a. the defining equation $L(Y) = k(Y-T)^2$,
b. the equation determining the constant k: $A = kd^2$,
c. the equation for average loss $E(L) = k\sigma^2$ or $E(L) = k\sigma^2 + k(E(Y)-T)^2$.

Quality costs can be enumerated in a much more complex manner involving all possible kinds of losses induced by the not-optimal product quality, such as expenses on repairs, expenses on product control, losses due to imprecise quality measurements, etc. We shall work with this concept in chapter 11.

There are also loss functions dependent on more quality characteristics, i.e. they are functions of several variables. Some of the characteristics do not even have to be quantitative.

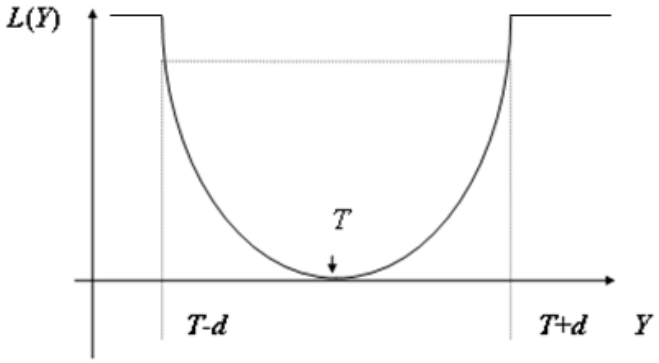
10.2 LOSS FUNCTIONS FOR DIFFERENT TYPES OF TOLERANCES

The loss function described in figure 19 is not the only one. There are more kinds of loss functions, depending on what tolerance interval we work with. What follows is a classification of some of the most fundamental loss functions. Each of the functions is accompanied by the corresponding graphical representation.

We distinguish the following types of tolerances:

a) Symmetric N-tolerance

Figure 20: Symmetric N-tolerance

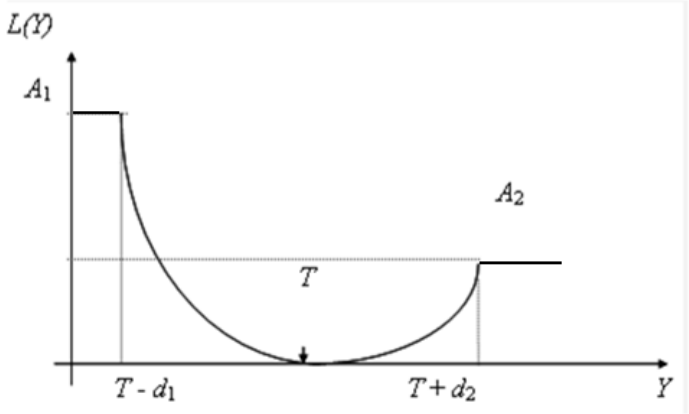


In this case, we write $T \pm d$, where d = tolerance. The interval (T-d, T+d) is called the *tolerance interval*. The tolerance is symmetric in the sense that the target value lies in the center of the tolerance interval. If the quality characteristic observed is smaller that the lower tolerance limit T-d, the financial loss incurred equals A, and the same is true for the case when the characteristic is greater than the upper tolerance limit T+d.

b) Nonsymmetric N-tolerance

In this case, the loss function looks as described by figure 21.

Figure 21: Nonsymmetric N-tolerance



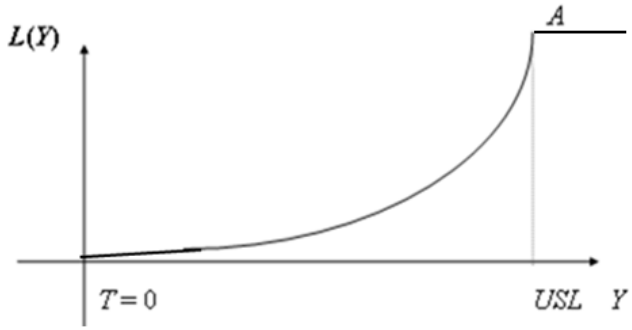
Here, the tolerance interval is (T - d_1, T + d_2). We see that there are two tolerances d_1 and d_2, which are generally different. There are also two generally different maximal losses A_1 and A_2, depending on whether the quality characteristic is too high or too low. The different maximal losses and different tolerances define two generally different curves above the horizontal axis, one of the curves being above and to the right of T, and the other being above and to the left of T.

On the interval (T-d_1, T), we work with equation 10-1, where $k = k_1$. On the interval (T, T+d_2) we work again with equation 10-1, however $k = k_2$ in this case. In the former case, $k_1 = A_1 / d_1^2$, in the latter case, $k_2 = A_2 / d_2^2$.

c) S-type tolerance (S for Small)

For this type of tolerance, the following is true: the smaller the observed quality characteristic Y, the better. The target value $T = 0$. Figure 22 describes the shape of an S-type tolerance loss function.

Figure 22: S-tolerance



To give an example of this situation, surface roughness can be the quality characteristic Y. Another example is air pollution. A certain upper tolerance/specification limit is acceptible, and beyond that limit, the losses reach their maximum. On interval (0, USL), we work with equation 10-1, where $k = A/\mathrm{USL}^2$; beyond USL, the loss function is constant.

d) L-type tolerance (L for Large)

In this case, the opposite is true: the bigger the characteristic Y, the better. The optimal/target value is $T = \infty$. Here, the loss function is of the form

10-4 $$E(L) = A \cdot d^2 \cdot s^2,$$

where

$$s^2 = E(1/Y^2),$$

i.e. the expected value (the „average“) of the random variable $1/Y^2$.

Equations a) -c) give an *individual loss*. If we want to determine the average loss, we calculate the average of the individual losses. Since the population average is usually unknown, we estimate it with the sample average. The same is true for the case d) where the unknown population average $E(1/Y^2)$ is replaced with its estimate $n^{-1}\sum Y^{-2}$.

Before presenting examples, let us summarize the essentials.

SUMMARY

We have described all major types of loss functions. We know that a quality characteristic is observed for each product. An optimal or target value is given for the characteristic, and if the optimal value is not achieved, certain financial loss is brought upon the product user. Depending on what the target value and tolerance are, we distinguish the following types of tolerances:

1. N-type tolerance: symmetric and unsymmetric,
2. S-type tolerance: smaller Y is better, $T = 0$.
3. L-type tolerance: larger T is better, $T = \infty$.

Different loss functions correspond to the cases 1-3.

PROBLEM 2

In a crankshaft production, length and diameter of crankshafts are observed. The diameter is supposed to be 25mm $\pm$ 1mm, while the length has a prescription of 100mm $\pm$ 2mm. If the diameter falls outside the tolerance interval, an individual loss of 40 crowns is generated; for the length, the individual loss is 30 crowns. Ten crankshafts have been taken out of the production line randomly for examination. These are the measurements of their length and diameter:

Diameter (in mm):

25,1; 25; 25; 24,9; 25,1; 25; 24,9; 25; 25,1; 24,9.

Length (in mm):

99,9; 99,9; 99,8; 100,2; 100; 100; 100,1; 98; 99,9; 100,2.

Compare the quality of two operations: one that yields a certain diameter of the crankshaft, and the other resulting in a length of the crankshaft.

SOLUTION:

Diameter: T_1 =25, $A_1 = 40$, $d_1 = 1$.

$$s^2 = \frac{1}{10}\left[(25{,}1-25)^2 + (25-25)^2 + \ldots + (24{,}9-25)^2\right] = 0{,}006\,.$$

$$Estimated\ E(L) = \frac{40}{1^2}0{,}006 = 0{,}24 \text{ crowns per unit.}$$

Length: $T_2 = 100$, $A_2 = 30$, $d_2 = 2$.

$$s^2 = \frac{1}{10}\left[(99{,}9-100)^2 + \ldots + (100{,}2-100)^2\right] = 0{,}02\,.$$

$$Estimated\ E(L) = \frac{30}{2^2} 0,02 = 0,15 \text{ crowns per unit.}$$

We can conclude that the length of the crankshaft is produced at a higher quality than the diameter of this product. The total average loss is estimated to be 0,24 + 0,15 = 0,39 crowns per unit (per crankshaft).

PROBLEM 3

Drums to be used in washing machines are supposed to be 30 cm wide (= diameter). The tolerances are specified as follows: 30cm-1cm, 30cm+4cm. If the diameter is smaller than the lower tolerance limit, a loss of 50 crowns is recorded by the consumer. If the diameter exceeds the upper tolerance limit, the loss is 100 crowns. Two companies produce the same drums. Compare their quality if the following data samples on their production are available:

Table 64: Production data

Company	Deviations from the target value (!)
A	0; 0; -1; 3; 0; 4; 2; -1; 0; 1; 2; 4
B	-1; -1; 0; 0; 0; 3; 2; -1; 1; 2; 0

Source: author's

SOLUTION:

The parameters are: $A_1 = 50$, $A_2 = 100$, $d_1 = 1$, $d_2 = 4$.

Company A:

$$Estimated\ E(L_1) = \frac{1}{12}\left\{\frac{50}{1^2}\left[(-1)^2+(-1)^2\right]+\frac{100}{4^2}\left(3^2+4^2+2^2+1^2+2^2+4^2\right)\right\}$$

$$Estimated\ E(L_1) = 34,375 \text{ crowns per unit.}$$

Company B:

$$Estimated\ E(L_2) = \frac{1}{11}\left\{\frac{50}{1^2}\left[(-1)^2+(-1)^2+(-1)^2\right]+\frac{100}{4^2}\left(3^2+2^2+1^2+2^2\right)\right\}$$

$$Estimated\ E(L_2) = 23,864 \text{ crowns per unit.}$$

The production of company B seems to give a higher quality. Let us take a closer look at the logic behind the calculations performed (the case A, the second case is analogous): we are estimating the average loss by the sample average. Therefore, the data are summed together and divided by the sample size, which is 12. Each of the terms in the summation represents an individual loss, i.e. a functional value of the loss function used. Since the tolerance we work

with is unsymmetric, each part of the loss function has its own defining equation. Each of these equations lead to a different value of the constant *k*. In one case, the constant equals 50, in the other case, it is equal to 100/16.

PROBLEM 4

Ballbearings are produced by two different companies: Company A produces it with specification $T \pm 0,4$, while the other company B must cling to specification $T \pm 1$. Fifty thousand ballbearings is produced every day. Each ballbearing costs 0,60 crowns. If the tolerance interval is not achieved, the corresponding production unit is scrapped. A spot check at the two companies led to the following data samples:

Company A: deviations from the optimal size:

-0,3; 0,1; 0,2; 0; 0; -0,2; -0,1; 0; 0,4; 0,1; -0,1; 0, 0; 0,1; -0,2.

Company B: deviations from the optimal size:

0; 0; 1; -0,8; -0,8; 0, 0,6; 0,7; 0; -0,3; -0,2; 0; 0; 1; 0,2.

Compare the production quality of the two companies.

SOLUTION

1. Company A: $A = 0,6$; $d = 0,4$.

We have

$$s^2 = \frac{1}{15}\left[(-0,3)^2 + 0,1^2 + ... + (-0,2)^2\right] = \frac{0,42}{15} = 0,028 .$$

and

$$Estimated\ E(L) = \frac{A}{d^2} s^2 = \frac{0,6}{0,4^2}.0,028 = 0,105 \text{ crowns per unit.}$$

The daily loss is 50 000·0,105 = 5 250 crowns.

2. Company B: $A = 0,6$; $d = 1$.

In this case,

$$s^2 = \frac{1}{15}\left[0^2 + ... + 0,2^2\right] = 0,287 .$$

and

$$Estimated\ E(L) = \frac{0,6}{1^2} 0,287 = 0,172 \text{ crowns per unit.}$$

The daily loss in the second case is 50 000·0,172 = 8 600 crowns.

PROBLEM 6

The surface of a piston is adjusted during production so that the surface roughness does not exceed 10 mm. The smoother the surface, the better. If the upper roughness limit is exceeded, it is re-worked with a machine tool for 200 crowns. Two different workers at two different firms have the same job: they are in charge of how smooth the piston surface is. Compare the quality of their work when the following data samples are available (table 65).

Table 65: Entry data

Worker	Surface smoothness
1	0, 1, 9, 6, 10, 2, 3, 0, 9
2	3, 2, 4, 4, 5, 2, 4, 6, 5, 3

Source: author's

SOLUTION

The parameters are: $A = 200$, $d = 10$ (S-type tolerance)

1st Worker:

$$s^2 = \frac{1}{9}\left((0\text{-}0)^2 + (1\text{-}0)^2 + (9\text{-}0)^2 + \ldots + (9\text{-}0)^2 \right) = 34{,}67.$$

$$Estimated\ E(L) = \frac{A}{d^2} s^2 = \frac{200}{10^2} 34{,}67 = 69{,}34 \text{ crowns per unit.}$$

2nd Worker:

$$s^2 = \frac{1}{10}(3^2 + 2^2 + 4^2 + \ldots + 5^2 + 3^2) = 16.$$

$$Esitmated\ E(L) = \frac{200}{10^2} 16 = 32 \text{ crowns per unit.}$$

The second worker's skills are more than twice as good as those of the first worker.

PROBLEM 7

Rock-climbing equipment makers are required to produce ropes with stiffness of at least 300 kg. If the lower limit is not achieved, the rope must be re-stiffened at the cost of 50 crowns per metre. Compare two technologies of rope-making if the following production data is available

Table 66: Production data

Technology	Rope stiffness
1	305, 350, 350, 410, 310, 300, 350, 400
2	305, 301, 308, 306, 300, 320, 310, 310, 320, 325

Source: author's

SOLUTION

The parameters are: $A = 50$, $d = 300$. It is the L - type tolerance.

1st technology.

Based on 10-4, we have:

$$s_1^2 = \frac{1}{8}\left(\frac{1}{305^2} + \frac{1}{350^2} + ... + \frac{1}{400^2}\right) = 8,62 \cdot 10^{-6}.$$

The average loss is

$$Est.\, E(L_1) = 50 \cdot 300^2 \cdot 8,62 \cdot 10^{-6} = 38,79 \text{ crowns per metre.}$$

2nd technology

Variance:

$$s_2^2 = \frac{1}{10}\left(\frac{1}{305^2} + \frac{1}{301^2} + ... + \frac{1}{325^2}\right) = 1,03 \cdot 10^{-5}.$$

Average loss:

$$Est.\, E(L_2) = 50 \cdot 300^2 \cdot 1,03 \cdot 10^{-5} = 46,76 \text{ crowns per metre.}$$

CONTROL TEST 10

Yes/No questions:

10.1 Loss function can be described mathematically as $L(Y) = k(Y - T)^2$?

10.2 The higher Y, the better…this is the S –type tolerance?

10.3 If a product feature has its optimal value, a lower-than-optimal quality of the product manifests itself by deviations of the feature value from the optimal value?

10.4 With the N-type tolerance, the optimal value is smaller than the target value?

10.5 Loss functions are in part a parabola?

Complete the statement:

10.6 Any deviation from the target value T brings _____ .

10.7 A certain product_____is usually observed, based on which we judge the quality of the product.

10.8 A part of loss function is mathematically a _____.

10.9 Based on what is considered the target value T, we distinguish these types of tolerances:_____, _____, _____.

10.10 When working with the S – tolerance, the target value T = _____.

10.11 A product diameter and weight are observed. The diameter is to be $T_1 = 20\text{cm} \pm 1$ and the weight is to be $T_2 = 100\text{g} \pm 2$. If the diameter falls outside the tolerance interval, it costs 20 crowns to repair the product or scrap it; for the weight, the same cost is 30 crowns. Ten product units have been randomly drawn from the production line:

Their diameter was:

20,1; 20; 20; 19,9; 20,1; 20; 19,9; 20, 20,1; 19,9.

Their weight was:

99,9; 99,9; 99,8; 100,2; 100; 100; 100,1; 9,8; 99,9; 100,2.

Compare the production quality in terms of the ability of the company to keep the target value of the diameter and the weight.

10.12 In filter production, it is required that the throughput of the filter be 10% at the most. Two filter makers were inspected, and these are the results of the inspection:

Maker	% throughput
A	3, 9, 9, 7, 1
B	8, 8, 1, 1, 2, 5

Source: author's

If the throughput tolerance is exceeded, the costs of the maker A rise by 600 crowns, whereas in the case of the maker B, the costs increase by 700 crowns. Which producer yields a better-quality filter?

SOLUTIONS

10. 1 yes
10. 2 no
10. 3 yes
10. 4 no
10. 5 yes
10. 6 losses
10. 7 characteristic
10. 8 $L(Y) = k(Y - T)^2$

10. 9 *N* (nominal), *S* (smaller), *L* (larger)

10. 10 0

10. 11 Diameter: $E(L) = 0{,}12$ crowns per unit; Weight: $E(L) = 0{,}15$ crowns per unit.

10. 12 A: $E(L) = 265{,}2$ crowns per unit; B: $E(L) = 185{,}5$ crowns per unit.

11 TAGUCHI'S METHODS: TOTAL QUALITY COSTS

In the previous chapter, we worked with loss functions that measured financial losses of customers, resulting from product lower quality. The losses were feature-specific. In this chapter, we shall include in the losses other types of expenses that are not directly linked to a specific product feature. Inclusion of these expenses leads to a **total** quality costs function. In chapter 11, we shall talk about control charts, as well. This is a very important part of quality management, as it monitors stability of production processes. The chapter is divided into four parts: the first part discusses quality cost monitoring, the second part introduces the total quality cost function when 100% control of the production is realized, the third part of the text provides the reader with the total quality cost function when the production is controlled after every batch of n production units, and the final part of the chapter describes basic types of control charts.

11.1 QUALITY COST MONITORING

The term „quality cost" can mean more than one thing, nevertheless, it is mostly connected to expenses on ensuring or improving quality, as well as expenses of nonproductive nature, such as those resulting from making nonconforming products. From the practical point of view, it is convenient to divide the quality costs into three categories:

- quality costs of the producer,
- quality costs of the customer,
- quality costs of the whole society.

We shall focus on the first category.

Producers must invest in prevention, production evaluation and defect removal, so that appropriate quality is achieved in all production stages, i.e. in the product development, product manufacturing, product installation and product use. Monitoring these „investments" allows for product improvement. There are different ways of monitoring the expenses:

1) monitoring based on PAF models,
2) monitoring based on process models,
3) monitoring through the Taguchi's approach.

Ad1) **PAF models** (**P**revention, **A**ppraisal, **F**ailure)

This model is based on dividing company costs into four categories:

Costs resulting from internal defects (these defects originate within company before its final product reaches the customer),

Costs resulting from external defects (these include customer complaints, repairs, handling costs, discounts, expenses due to lawsuits, market share losses, etc.),

Evaluation costs (these are mainly expenditures on measuring customer satisfaction, measuring equipment, software, certification, laboratory testing, etc.).

Prevention costs (these are expenses which should rise continuously; they include expenditures on exploring customer demands, management system development, education of employees and others).

Ad2) **Process models**

Process models represent a higher degree of monitoring which keeps track of costs related not to each product but processes. The costs involve expenses on converting process inputs into process outputs according to a plan, as well as expenses on resolving inconsistencies that were not supposed to originate in the process at all.

Ad 3) **Taguchi's methods**

These methods use mathematics to describe relations between total quality costs and different factors that contribute to the costs. The exact nature of mathematics enables to optimize the costs, which is, of course, an advantage of this approach. We shall devote ourselves to Taguchi's methods in the following two sections 11.2 and 11.3.

11.2 TAGUCHI'S APPROACH – THE CASE OF 100% PROCESS CONTROL

The total quality costs are calculated in this case according to equation

11-1
$$L = \frac{Q}{R} + \frac{A}{d^2} s_0^2,$$

where

Q = yearly expenses on 100 % control,
R = yearly production (number of product units made),
d = tolerance within which the product remains satisfactory in terms of its quality,
A = losses due to exceeding the tolerance d,

$$s_0^2 = \frac{1}{n-1}\left[(y_2 - y_1)^2 + (y_3 - y_2)^2 + ... + (y_n - y_{n-1})^2 \right],$$

the y's being measurements of the observed product quality characteristic.

PROBLEM 1

An automated control (i.e. a 100% control) at a factory costs 25 000 crowns a year. Each year, four million product units leave the factory. The tolerance for the quality characteristic observed is 9 and the company's costs rise by 5 crows each time the tolerance is exceeded. Calculate the total quality costs if a random sampling showed the „variability" of the characteristic to be $s_0^2 = 1$.

SOLUTION

We have:

Q = 25 000 Kč
R = 4 000 000 ks,
d = 9,
A = 5 Kč,
$s_0^2 = 1$.

Thus, according to 11-1, we get:

$$L = \frac{25000}{4000000} + \frac{5}{9^2} \cdot 1 = 0,068 \text{ crowns per unit.}$$

The total quality costs are 4 000 000·0,068 = 272 000 crowns per year.

11.3 THE CASE OF PROCESS CONTROL AFTER N UNITS

If n product units are made between two controls, the total quality costs are calculated according to formula

11-2

$$L = \frac{B}{n} + \frac{C}{u} + \frac{A}{d^2}\frac{D^2}{3} + \frac{A}{d^2}\frac{D^2}{u}\left(\frac{n+1}{2} + z\right) + \frac{A}{d^2}s_m^2,$$

where

A = loss due to exceeding tolerance d,
B = product control costs,
C = production machinery repair costs,
n = control interval,
u = average number of units produced between two controls,
d = tolerance within which the product remains satisfactory in terms of its quality (the tolerance is defined by the customer),
D = tolerance defined by the producer (it is usually more demanding than what the customer demands),
z = number of product units made during the control,
$\frac{B}{n}$ = control costs per unit,
$\frac{C}{u}$ = repair costs per unit,
$\frac{A}{d^2}\frac{D^2}{3}$ = costs resulting from imprecise production,
$\frac{A}{d^2}\frac{D^2}{u}\left(\frac{n+1}{2} + z\right)$ = costs due to producing defective units,
$\frac{A}{d^2}s_m^2$ = costs due to imprecise measurements.

Equation 11-2 is a result of Taguchi's long-time experience, and it was mathematically defined, not derived. However, three terms in the equation are derived from loss functions.

The question we usually ask ourselves is: How often should the production control take place so that the total quality costs were minimal? What tolerance should the company define for itself to minimize its total quality costs? The answers to the questions can be obtained by standard optimization procedures which seek local minima of a function: the first-order derivatives with respect to *n* and *D* are calculated (the two problems are solved separately), and the derivatives are put equal to zero. This necessary extremum condition leads to an optimal control interval of

11-3
$$n^* = \sqrt{\frac{2uB}{A}}\frac{d}{D}.$$

and an optimal tolerance

11-4
$$D^* = \sqrt[4]{\frac{3CD^2d^2}{Au}}.$$

11.4 CONTROL CHARTS

Control charts rank among major statistical tools for production process regulation. The charts were introduced by Walter Shewhart in the 1920s. Their aim is to monitor a characteristic of a process in time, and give a signal if a problem in the process occurs. If such a deterioration of the process takes place, the process owner reacts to the situation, and makes the necessary adjustments to the process. Thus, the charts serve as a problem prevention. Values of the observed characteristic are measured on the y-axis of the chart, whereas its x-axis records points in time at which the characteristic is observed or measured. The time series values of the characteristic should not exceed certain limits, nor should they form an improbable pattern. In either case, the chart signals a systematic impact on the process which has nothing to do with the natural character of the process. Such an impact may result, for instance, from a defect developed in a machine that is used during the process. Although nearly any process characteristic can be observed in time, it should satisfy some basic requirements if it is to be used in the framework of statistics. In our case, such a characteristic should at least approximately follow a normal distribution.

Two properties of the process characteristic (or the process in question) are observed:

- **a.** its ability to keep itself close to a pre-defined target value,
- **b.** its variability around the target value.

Therefore, two control charts are usually constructed, each of which observes either the property a) or b). Perhaps the most common charts are: the chart for the average $\bar{x}$ of the characteristic and the chart for the range of the characteristic *R*; the pair is denoted CC$(\bar{x}, R)$; or the chart for the average $\bar{x}$ of the characteristic and the chart for its standard deviation *s*, i.e. the pair CC$(\bar{x}, s)$. Let us take a closer look at the first pair.

To construct a CC$(\bar{x}, R)$, we take the following steps:

a. We gather data about the characteristic at time points $t = 1,2, \ldots, m$ (the first column of table 67).

b. The average and range $R = x_{\max} - x_{\min}$ are calculated for each of the samples, i.e. for each point in time t.

c. An optimal value (central line), an upper limit and a lower limit are calculated to be used in the chart as a reference (see below).

d. We plot the time averages in one chart to get the CC$(\bar{x})$ chart. Likewise, we plot the individual time ranges in another chart to get the CC(R) graph.

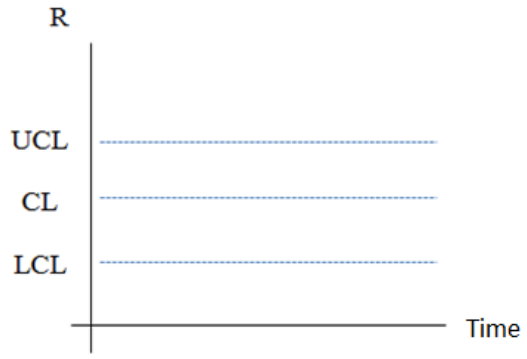
The time series depicted by either chart should stay below the calculated upper limit and above the calculated lower limit. Also, it should fluctuate more or less randomly around the central line of the chart. If it doesn't seem to be the case, the process owner must check the status of the process.

Tabulka 67: Samples and characteristics for CC($\bar{x}$, R) charts

Data	x-axis	y-axis	
	i = time	average $\bar{x}_i$	range R_i
$x_{11}, x_{12}, \ldots, x_{1n}$	1	$\bar{x}_1$	R_1
$x_{21}, x_{22}, \ldots, x_{2n}$	2	$\bar{x}_2$	R_2
$x_{31}, x_{32}, \ldots, x_{3n}$	3	$\bar{x}_3$	R_3
:	:	:	:
$x_{m1}, x_{m2}, \ldots, x_{mn}$	m	$\bar{x}_m$	R_m

Figure 23 outlines the fundamental limits of the CC(R) chart:

Figure 23: Limits and lines of the CC(R) control chart



UCL = upper control limit, LCL = lower control limit, CL = central line.

For the CC($\bar{x}$) chart, the limits are calculated as follows::

$$LCL = \bar{\bar{x}} - A_2\bar{R},$$
$$UCL = \bar{\bar{x}} + A_2\bar{R},$$
$$CL = \bar{\bar{x}},$$

where

$$\bar{\bar{x}} = \frac{1}{m}\sum_{i=1}^{m}\bar{x}_i$$

is the average of the individual sample averages from different time points. Further,

$$\bar{R} = \frac{1}{m}\sum_{i=1}^{m}R_i\,,$$

is the average of the individual ranges from different time points.

The limits for the CC(*R*) chart are:

$$UCL = D_4\bar{R}\,,$$
$$LCL = D_3\bar{R}\,,$$
$$CL = \bar{R}\,.$$

The constants A_2, D_3 and D_4 are in table 68 (the table has zeros where the data's missing).

Table 68: Constants for *LCL* and *UCL* limits of control charts

n	A_2	A_3	B_3	B_4	D_3	D_4	d_2
2	1,88	2,550		3,267		3,267	1,128
3	1,023	1,954		2,568		2,574	1,693
4	0,720	1,628		2,265		2,282	2,039
5	0,577	1,427		2,089		2,114	2,326
6	0,483	1,267	0,330	1,970		2,004	2,534
7	0,419	1,182	0,118	1,892	0,076	1,924	2,704
8	0,373	1,029	0,185	1,815	0,136	1,864	2,847
9	0,337	1,032	0,239	1,761	0,184	1,816	2,970
10	0,308	0,975	0,284	1,716	0,223	1,777	3,078
11	0,285	0,927	0,321	1,679	0,256	1,744	3,173
12	0,265	0,886	0,354	1,646	0,283	1,717	3,258
13	0,249	0,850	0,382	1,618	0,307	1,693	3,336
14	0,235	0,817	0,406	1,594	0,328	1,672	3,407
15	0,223	0,789	0,428	1,572	0,347	1,653	3,472
16	0,212	0,763	0,448	1,552	0,363	1,637	3,532
17	0,203	0,739	0,465	1,534	0,378	1,622	3,558
18	0,194	0,718	0,482	1,518	0,391	1,608	3,640
19	0,187	0,698	0,497	1,503	0,403	1,597	3,689
20	0,080	0,680	0,510	1,490	0,415	1,585	3,735
21	0,173	0,663	0,523	1,477	0,425	1,575	3,778
22	0,167	0,647	0,534	1,466	0,434	1,566	3,719
23	0,162	0,633	0,545	1,455	0,443	1,557	3,858
24	0,157	0,619	0,555	1,445	0,451	1,548	3,859
25	0,153	0,606	0,565	1,435	0,459	1,541	3,931

Source: author's

PROBLEM 2

Calculate the limits *UCL*, *LCL* and *CL* of CC($\bar{x}$) and CC(*R*) if the following data is given

Table 69: Data samples for CC$(\bar{x}, R)$

i					$\bar{x}$	*R*
1	9,9	9,9	11	9,8	10,15	1,2
2	9,1	9,8	9,9	11,2	10	2,1
3	9,6	9,4	10,7	9,9	9,9	1,3
4	10,4	9,4	9,2	9,9	9,725	1,2
5	9,9	10,6	9,6	10	10,03	1
6	10,3	9,8	9,7	9,9	9,925	0,6
7	10,2	11,1	9,6	10	10,23	1,5
suma					69,96	8,9

Source: author's

SOLUTION

Control chart CC($\bar{x}$):

$$\bar{\bar{x}} = \frac{1}{m}\sum_{i=1}^{m}\bar{x}_i = 9,994 .$$

$$\bar{R} = \frac{1}{m}\sum_{i=1}^{m}R_i = 1,271 .$$

$$LCL = \bar{\bar{x}} - A_2\bar{R} = 9,994 - 0,720 \cdot 1,271 = 0,915 .$$

$$UCL = \bar{\bar{x}} + A_2\bar{R} = 9,994 + 0,720 \cdot 1,271 = 10,909 .$$

$$CL = \bar{\bar{x}} = 9,994 .$$

Control chart CC(*R*):

$$UCL = D_4\bar{R} = 2,282 \cdot 1,271 = 2,9 .$$

$$LCL = D_3\bar{R} = 0 \cdot 1,271 = 0 .$$

$$CL = \bar{R} = 1,271 .$$

Subsequently, for each point in time, the corresponding individual average would be plotted on the vertical axis of the control chart CC($\bar{x}$), the time point being plotted on the horizontal axis of the chart, and the same is true for the control chart CC(*R*), in which the individual range corresponding to the particular point in time would be plotted on the vertical axis. In the end, the charts are evaluated: the basic rule is that none of the points plotted in the chart is either above the UCL limit or below the LCL limit of the chart. If it does happen, the process must be checked as to what occured in the process at the moment when the point exceeded the chart limits. It is quite probable that something unnatural or systematic interfered the natural course of the process.

SUMMARY

The chapter presented the total quality cost function, and also served as an introduction to control charts. The cost function was not derived mathematically, it was defined, based on experience of Genuchi Taguchi, a Japanese engineer. If production process is not controlled each time its product unit is made, it is convenient to figure out how often the control should be carried out and how precise it should be so that it didn't cost too much. The total quality cost function can be used for these purposes.

Control charts are a fundamental tool for statistical process regulation. Their objective is to monitor process characteristics, and signal that anomalies occur in the process. There are different types of control charts, depending on the nature of the process characteristic observed. The reader became acquainted with two basic control charts: one of them monitors time development of the average value of the process characteristic, while the other records time development of the range of the same characteristic.

What follows is a problem related to the total quality cost functions.

PROBLEM 3

A pressing machine produces a set of 8 pressed parts at once. It cost 0,5 crowns to produce each of these parts. The factory controls the pressing by checking the whole set, and if one of the parts is faulty, the whole set is scrapped, the machine is stopped and adjusted at a cost of 70 crowns. The tolerance defined by the customer for the size of each part is 4, while the factory has its own tolerance of 10. Four hundred and eighty parts are produced every hour, and the number of working hours is 2000 per year. The control of the machine lasts 2 minutes, and it costs 10 crowns. Imprecision of the measuring equipment used during the control is not considered here. The time interval between two controls is 4 hours in average. We are to calculate the total quality costs, and determine the optimal control regime, including its contribution to the quality cost reduction.

SOLUTION

The parameters are:

$A = 8 \cdot 0{,}5 = 4$ crowns
$B = 10$ crowns
$C = 70$ crowns
$D_o = 4$
$d = 10$
$n_o = 480$ units
$z = \frac{480}{60}.2 = 16$
$u_o = 4.480 = 1920$ units

Inserting u_o, n_o and D_o in 11-2, we have the current quality costs of

$$L_0 = \frac{10}{480} + \frac{70}{1920} + \frac{4}{10^2}\frac{4^2}{3} + \frac{4}{10^2}\left(\frac{480+1}{2} + 16\right) \cdot \frac{4^2}{1920} = 0{,}356 \text{ crowns per unit.}$$

To optimize the control, we have

a. According to 11-3

$$n^* = \sqrt{\frac{2u_o B}{A}}\,\frac{d}{D_o} = \sqrt{\frac{2\cdot 1920\cdot 10}{4}}\,\frac{10}{4} = 244,9 \approx 240 \text{ units},$$

i.e. the control should be performed approximately every 30 minutes.

b. Based on 11-4, the optimal tolerance is

$$D^* = \sqrt[4]{\frac{3CD_o^2 d^2}{Au_o}} = \sqrt[4]{\frac{3\cdot 70\cdot 4^2\cdot 10^2}{4\cdot 1920}} = 2,57 \approx 2\,.$$

c. Cost savings:

$$L = \frac{B}{n^*} + \frac{C}{u} + \frac{A}{d^2}\frac{D^{*2}}{3} + \frac{A}{d^2}\left(\frac{n^*+1}{2} + z\right)\cdot\frac{D^{*2}}{u}$$

$$L = \frac{10}{240} + \frac{70}{480} + \frac{4}{10^2}\cdot\frac{2^2}{3} + \frac{4}{10^2}\left(\frac{240+1}{2} + 16\right)\cdot\frac{2^2}{480} = 0,287 \text{ crowns per unit.}$$

The cost reduction is

$$L_0 - L = \ 0{,}356 - 0{,}287 = 0{,}069 \text{ crowns per unit},$$

that is 0,069·480·2000 = 66 240 crowns in savings per year.

CONTROL TEST 11

Yes/No answers:

11.1 If a production process is controlled each time its product unit is made, the total quality costs are $L = \frac{Q}{R} + \frac{A}{d^2}s_0^2$?

11.2 The fundamental formula for quality cost evaluation is mathematically derived?

11.3 When working with a process characteristic, we are interested in how well the characteristic clings to its target value and how it fluctuates around the target value?

11.4 The range R of a data sample is calculated as $R = x_{\max} - \bar{x}$?

11.5 $\mathrm{CC}(\bar{x}, R)$ represents a chart for the average and a chart for the range of a process characteristic?

Complete the statement:

11.6 In equation $L = \frac{Q}{R} + \frac{A}{d^2}s_0^2$, $s_0^2 =$_____.

11.7 If process control is not performed each time a product unit is made, we are interested in how often _____ and _____.

11.8 _____ _____ is a main tool for statistical process regulation.

11.9 Two elementary control charts are: for the average and _____; and for the average and _____ _____.

SOLUTIONS

11.1 yes
11.2 no
11.3 yes
11.4 no
11.5 yes
11.6 $s_0^2 = \frac{1}{n-1}\left[(y_2 - y_1)^2 + (y_3 - y_2)^2 + ... + (y_n - y_{n-1})^2\right]$
11.7 to control, how precisely to control
11.8 control charts
11.9 range; standard deviation.

CONCLUSION

This textbook has presented selected but frequently used statistical methods which include procedures applied in industry. The logic of the text was based on the fact that industrial procedures draw on statistical terms and techniques, which means, the terms and techniques have to be presented before they can be applied either in industry or other sectors of economy. Classical statistical methods include, but are not restricted to, regression, correlation analysis, hypothesis testing, time series analysis, analysis of variance, and descriptive statistics. These topics were covered in chapters 1-8, whereas the remaining chapters described the fundamentals of the design of experiments, which is closely related to regression, Taguchi's loss functions and control charts. The structure of the text followed the well-established scheme according to which the subject matter explained is accompanied by examples, and the end of the chapter presents relevant questions.

The textbook best serves as an outline of major statistical methods, providing the reader with main ideas and principles of the methods. The extent and depth of the presented topics comply with the subject matter contained in the course Statistical methods for economists, taught at the Faculty of Business Administration of the Silesian University. There are other literary resources, as well, which cover each topic of this textbook. These resources focus specifically only on some of the methods, and therefore elaborate the ideas behind the methods further, as compared to what is presented in this textbook. The reader is encouraged to examine other external scholarly texts, as well. Some of the literary sources are listed on the following page.

REFERENCES

[1]ANTONY, J.: *Design of Experiments for Engineers and Scientists*, 8th edition, Butterworth-Heinemann, 2003, ISBN: 0-7506-4709-4.
[2] BISSELL, B.: *Statistical methods for SPC and TQM.* 1.vyd. London: Chapman and Hall, 1994, ISBN 9780412394409.
[3] BRÄNNMARK M., LANGSTRAND J., JOHANSSON S., HALVARSSON A., ABRAHAMSSON L., WINKEL J.: *Researching Lean*: *Methodological Implications of Loose Definitions*, Quality, Innovation, Prosperity, Vol. XVI/2-2012, p. 35-48, ISSN 1335-1745, DOI: 10.12776/qip.v16i2.67.
[4] McClave, J., SINCICH, T.: *Statistics,* 12th edition, Pearson Education Ltd., 2014, ISBN: 978-1-292-02265-9.
[5] ROY, R.K.: *Design of Experiments Using Taguchi Approach,* 1st edition, John Wiley and Sons, 2001, ISBN: 0-471-36101-1.
[6] SIVA V.: *Improvement in Product Development: Use of Back-End Data to Support Upstream Efforts of Robust Design Methodology,* Quality, Innovation, Prosperity, Vol. XVI/2-2012, p. 84-102, ISSN: 1335-1745, DOI: 10.12776/qip.v16i2.65.
[7] TAGUCHI G., CHOWDHURY, S., WU, Y.: *Taguchi's Quality Engineering Handbook*, 1st edition, John Wiley and Sons, 2005, ISBN: 0-471-41334-8.
[8] TOŠENOVSKÝ, J., NOSKIEVIČOVÁ, D.: *Statistické metody pro zlepšování jakosti.* 1.vyd. Ostrava: Montanex, a.s., 2001, ISBN 80-7225-040-X.
[9] TOŠENOVSKÝ, J., DUDEK, M.: *Základy statistického zpracování dat.*1.vyd. Ostrava: VŠB, 2001, ISBN 80-248-0006-3.
[10] WITTE, R.S., WITTE, J.S.: *Statistics,* 9th edition, John Wiley and Sons, 2010, ISBN: 978-470-39222-5.
[11] ZGODAVOVÁ, K., BOBER, P.: *An Innovative Approach to the integrated Management System Development: SIMPRO-IMS Web-Based Environment*, Quality, Innovation, Prosperity, Vol. XVI/2-2012, p. 59-70, ISSN: 1335-1745, DOI: 10.12776/qip.v16i2.69.
[12] ZGODAVOVÁ, K.: *Complexity of Entities and its Metrological Implications*, Proceedings of the 21st International DAAAM Symposium, p. 365-367, 2010, ISSN: 1726-9679.

APPENDIX 1 – TABLE FOR DURBIN-WATSON'S TEST

Table for Durbin – Watson's test: **alpha = 1%**, dL = lowel limit, dU = upper limit, *n* = sample size, k' = number of model regressors without the absolute term.

	$k'=1$		$k'=2$		$k'=3$		$k'=4$		$k'=5$		$k'=6$		$k'=7$		$k'=8$		$k'=9$		$k'=10$	
n	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*
6	0.390	1.142	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
7	0.435	1.036	0.294	1.676	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
8	0.497	1.003	0.345	1.489	0.229	2.102	—	—	—	—	—	—	—	—	—	—	—	—	—	—
9	0.554	0.998	0.408	1.389	0.279	1.875	0.183	2.433	—	—	—	—	—	—	—	—	—	—	—	—
10	0.604	1.001	0.466	1.333	0.340	1.733	0.230	2.193	0.150	2.690	—	—	—	—	—	—	—	—	—	—
11	0.653	1.010	0.519	1.297	0.396	1.640	0.286	2.030	0.193	2.453	0.124	2.892	—	—	—	—	—	—	—	—
12	0.697	1.023	0.569	1.274	0.449	1.575	0.339	1.913	0.244	2.280	0.164	2.665	0.105	3.053	—	—	—	—	—	—
13	0.738	1.038	0.616	1.261	0.499	1.526	0.391	1.826	0.294	2.150	0.211	2.490	0.140	2.838	0.090	3.182	—	—	—	—
14	0.776	1.054	0.660	1.254	0.547	1.490	0.441	1.757	0.343	2.049	0.257	2.354	0.183	2.667	0.122	2.981	0.078	3.287	—	—
15	0.811	1.070	0.700	1.252	0.591	1.464	0.488	1.704	0.391	1.967	0.303	2.244	0.226	2.530	0.161	2.817	0.107	3.101	0.068	3.374
16	0.844	1.086	0.737	1.252	0.633	1.446	0.532	1.663	0.437	1.900	0.349	2.153	0.269	2.416	0.200	2.681	0.142	2.944	0.094	3.201
17	0.874	1.102	0.772	1.255	0.672	1.432	0.574	1.630	0.480	1.847	0.393	2.078	0.313	2.319	0.241	2.566	0.179	2.811	0.127	3.053
18	0.902	1.118	0.805	1.259	0.708	1.422	0.613	1.604	0.522	1.803	0.435	2.015	0.355	2.238	0.282	2.467	0.216	2.697	0.160	2.925
19	0.928	1.132	0.835	1.265	0.742	1.415	0.650	1.584	0.561	1.767	0.476	1.963	0.396	2.169	0.322	2.381	0.255	2.597	0.196	2.813
20	0.952	1.147	0.863	1.271	0.773	1.411	0.685	1.567	0.598	1.737	0.515	1.918	0.436	2.110	0.362	2.308	0.294	2.510	0.232	2.714
21	0.975	1.161	0.890	1.277	0.803	1.408	0.718	1.554	0.633	1.712	0.552	1.881	0.474	2.059	0.400	2.244	0.331	2.434	0.268	2.625
22	0.997	1.174	0.914	1.284	0.831	1.407	0.748	1.543	0.667	1.691	0.587	1.849	0.510	2.015	0.437	2.188	0.368	2.367	0.304	2.548
23	1.018	1.187	0.938	1.291	0.858	1.407	0.777	1.534	0.698	1.673	0.620	1.821	0.545	1.977	0.473	2.140	0.404	2.308	0.340	2.479
24	1.037	1.199	0.960	1.298	0.882	1.407	0.805	1.528	0.728	1.658	0.652	1.797	0.578	1.944	0.507	2.097	0.439	2.255	0.375	2.417
25	1.055	1.211	0.981	1.305	0.906	1.409	0.831	1.523	0.756	1.645	0.682	1.766	0.610	1.915	0.540	2.059	0.473	2.209	0.409	2.362
26	1.072	1.222	1.001	1.312	0.928	1.411	0.855	1.518	0.783	1.635	0.711	1.759	0.640	1.889	0.572	2.026	0.505	2.168	0.441	2.313
27	1.089	1.233	1.019	1.319	0.949	1.413	0.878	1.515	0.808	1.626	0.738	1.743	0.669	1.867	0.602	1.997	0.536	2.131	0.473	2.269
28	1.104	1.244	1.037	1.325	0.969	1.415	0.900	1.513	0.832	1.618	0.764	1.729	0.696	1.847	0.630	1.970	0.566	2.098	0.504	2.229
29	1.119	1.254	1.054	1.332	0.988	1.418	0.921	1.512	0.855	1.611	0.788	1.718	0.723	1.830	0.658	1.947	0.595	2.068	0.533	2.193
30	1.133	1.263	1.070	1.339	1.006	1.421	0.941	1.511	0.877	1.606	0.812	1.707	0.748	1.814	0.684	1.925	0.622	2.041	0.562	2.160
31	1.147	1.273	1.085	1.345	1.023	1.425	0.960	1.510	0.897	1.601	0.834	1.698	0.772	1.800	0.710	1.906	0.649	2.017	0.589	2.131
32	1.160	1.282	1.100	1.352	1.040	1.428	0.979	1.510	0.917	1.597	0.856	1.690	0.794	1.788	0.734	1.889	0.674	1.995	0.615	2.104
33	1.172	1.291	1.114	1.358	1.055	1.432	0.996	1.510	0.936	1.594	0.876	1.683	0.816	1.776	0.757	1.874	0.698	1.975	0.641	2.080
34	1.184	1.299	1.128	1.364	1.070	1.435	1.012	1.511	0.954	1.591	0.896	1.677	0.837	1.766	0.779	1.860	0.722	1.957	0.665	2.057
35	1.195	1.307	1.140	1.370	1.085	1.439	1.028	1.512	0.971	1.589	0.914	1.671	0.857	1.757	0.800	1.847	0.744	1.940	0.689	2.037
36	1.206	1.315	1.153	1.376	1.098	1.442	1.043	1.513	0.988	1.588	0.932	1.666	0.877	1.749	0.821	1.836	0.766	1.925	0.711	2.018
37	1.217	1.323	1.165	1.382	1.112	1.446	1.058	1.514	1.004	1.586	0.950	1.662	0.895	1.742	0.841	1.825	0.787	1.911	0.733	2.001
38	1.227	1.330	1.176	1.388	1.124	1.449	1.072	1.515	1.019	1.585	0.966	1.658	0.913	1.735	0.860	1.816	0.807	1.899	0.754	1.985
39	1.237	1.337	1.187	1.393	1.137	1.453	1.085	1.517	1.034	1.584	0.982	1.655	0.930	1.729	0.878	1.807	0.826	1.887	0.774	1.970
40	1.246	1.344	1.198	1.398	1.148	1.457	1.098	1.518	1.048	1.584	0.997	1.652	0.946	1.724	0.895	1.799	0.844	1.876	0.789	1.956
45	1.288	1.376	1.245	1.423	1.201	1.474	1.156	1.528	1.111	1.584	1.065	1.643	1.019	1.704	0.974	1.768	0.927	1.834	0.881	1.902
50	1.324	1.403	1.285	1.446	1.245	1.491	1.205	1.538	1.164	1.587	1.123	1.639	1.081	1.692	1.039	1.748	0.997	1.805	0.955	1.864
55	1.356	1.427	1.320	1.466	1.284	1.506	1.247	1.548	1.209	1.592	1.172	1.638	1.134	1.685	1.095	1.734	1.057	1.785	1.018	1.837
60	1.383	1.449	1.350	1.484	1.317	1.520	1.283	1.558	1.249	1.598	1.214	1.639	1.179	1.682	1.144	1.726	1.108	1.771	1.072	1.817
65	1.407	1.468	1.377	1.500	1.346	1.534	1.315	1.568	1.283	1.604	1.251	1.642	1.218	1.680	1.186	1.720	1.153	1.761	1.120	1.802
70	1.429	1.485	1.400	1.515	1.372	1.546	1.343	1.578	1.313	1.611	1.283	1.645	1.253	1.680	1.223	1.716	1.192	1.754	1.162	1.792
75	1.448	1.501	1.422	1.529	1.395	1.557	1.368	1.587	1.340	1.617	1.313	1.646	1.284	1.682	1.256	1.716	1.227	1.746	1.199	1.785
80	1.466	1.515	1.441	1.541	1.416	1.568	1.390	1.595	1.364	1.624	1.338	1.653	1.312	1.683	1.285	1.714	1.259	1.745	1.232	1.777
85	1.482	1.528	1.458	1.553	1.435	1.578	1.411	1.603	1.386	1.630	1.362	1.657	1.337	1.685	1.312	1.714	1.287	1.743	1.262	1.773
90	1.496	1.540	1.474	1.563	1.452	1.587	1.429	1.611	1.406	1.636	1.383	1.661	1.360	1.687	1.336	1.714	1.312	1.741	1.288	1.769
95	1.510	1.552	1.489	1.573	1.468	1.596	1.446	1.618	1.425	1.642	1.403	1.666	1.381	1.690	1.358	1.715	1.336	1.741	1.313	1.767
100	1.522	1.562	1.503	1.583	1.482	1.604	1.462	1.625	1.441	1.647	1.421	1.670	1.400	1.693	1.378	1.717	1.357	1.741	1.335	1.765
150	1.611	1.637	1.598	1.651	1.584	1.665	1.571	1.679	1.557	1.693	1.543	1.708	1.530	1.722	1.515	1.737	1.501	1.752	1.486	1.767
200	1.664	1.684	1.653	1.693	1.643	1.704	1.633	1.715	1.623	1.725	1.613	1.735	1.603	1.746	1.592	1.575	1.582	1.768	1.571	1.779

	$k'=11$		$k'=12$		$k'=13$		$k'=14$		$k'=15$		$k'=16$		$k'=17$		$k'=18$		$k'=19$		$k'=20$	
n	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*
16	0.060	3.446	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
17	0.084	3.286	0.053	3.506	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
18	0.113	3.146	0.075	3.358	0.047	3.557	—	—	—	—	—	—	—	—	—	—	—	—	—	—
19	0.145	3.023	0.102	3.227	0.067	3.420	0.043	3.601	—	—	—	—	—	—	—	—	—	—	—	—
20	0.178	2.914	0.131	3.109	0.092	3.297	0.061	3.474	0.038	3.639	—	—	—	—	—	—	—	—	—	—
21	0.212	2.817	0.162	3.004	0.119	3.185	0.084	3.358	0.055	3.521	0.035	3.671	—	—	—	—	—	—	—	—
22	0.246	2.729	0.194	2.909	0.148	3.084	0.109	3.252	0.077	3.412	0.050	3.562	0.032	3.700	—	—	—	—	—	—
23	0.281	2.651	0.227	2.822	0.178	2.991	0.136	3.155	0.100	3.311	0.070	3.459	0.046	3.597	0.029	3.725	—	—	—	—
24	0.315	2.580	0.260	2.744	0.209	2.906	0.165	3.065	0.125	3.218	0.092	3.363	0.065	3.501	0.043	3.629	0.027	3.747	—	—

	$k' = 11$		$k' = 12$		$k' = 13$		$k' = 14$		$k' = 15$		$k' = 16$		$k' = 17$		$k' = 18$		$k' = 19$		$k' = 20$	
n	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*
25	0.348	2.517	0.292	2.674	0.240	2.829	0.194	2.982	0.152	3.131	0.116	3.274	0.085	3.410	0.060	3.538	0.039	3.657	0.025	3.766
26	0.381	2.460	0.324	2.610	0.272	2.758	0.224	2.906	0.180	3.050	0.141	3.191	0.107	3.325	0.079	3.452	0.055	3.572	0.036	3.682
27	0.413	2.409	0.356	2.552	0.303	2.694	0.253	2.836	0.208	2.976	0.167	3.113	0.131	3.245	0.100	3.371	0.073	3.490	0.051	3.602
28	0.444	2.363	0.387	2.499	0.333	2.635	0.283	2.772	0.237	2.907	0.194	3.040	0.156	3.169	0.122	3.294	0.093	3.412	0.068	3.524
29	0.474	2.321	0.417	2.451	0.363	2.582	0.313	2.713	0.266	2.843	0.222	2.972	0.182	3.098	0.146	3.220	0.114	3.338	0.087	3.450
30	0.503	2.283	0.447	2.407	0.393	2.533	0.342	2.659	0.294	2.785	0.249	2.909	0.208	3.032	0.171	3.152	0.137	3.267	0.107	3.379
31	0.531	2.248	0.475	2.367	0.422	2.487	0.371	2.609	0.322	2.730	0.277	2.851	0.234	2.970	0.196	3.087	0.160	3.201	0.128	3.311
32	0.558	2.216	0.503	2.330	0.450	2.446	0.399	2.563	0.350	2.680	0.304	2.797	0.261	2.912	0.221	3.026	0.184	3.137	0.151	3.246
33	0.585	2.187	0.530	2.296	0.477	2.408	0.426	2.520	0.377	2.633	0.331	2.746	0.287	2.858	0.246	2.969	0.209	3.078	0.174	3.184
34	0.610	2.160	0.556	2.266	0.503	2.373	0.452	2.481	0.404	2.590	0.357	2.699	0.313	2.808	0.272	2.915	0.233	3.022	0.197	3.126
35	0.634	2.136	0.581	2.237	0.529	2.340	0.478	2.444	0.430	2.550	0.383	2.655	0.339	2.761	0.297	2.865	0.257	2.969	0.221	3.071
36	0.658	2.113	0.605	2.210	0.554	2.310	0.504	2.410	0.455	2.512	0.409	2.614	0.364	2.717	0.322	2.818	0.282	2.919	0.244	3.019
37	0.680	2.092	0.628	2.186	0.578	2.282	0.528	2.379	0.480	2.477	0.434	2.576	0.389	2.675	0.347	2.774	0.306	2.872	0.268	2.969
38	0.702	2.073	0.651	2.164	0.601	2.256	0.552	2.350	0.504	2.445	0.458	2.540	0.414	2.637	0.371	2.733	0.330	2.828	0.291	2.923
39	0.723	2.055	0.673	2.143	0.623	2.232	0.575	2.323	0.528	2.414	0.482	2.507	0.438	2.600	0.395	2.694	0.354	2.787	0.315	2.879
40	0.744	2.039	0.694	2.123	0.645	2.210	0.597	2.297	0.551	2.386	0.505	2.476	0.461	2.566	0.418	2.657	0.377	2.748	0.338	2.838
45	0.835	1.972	0.790	2.044	0.744	2.118	0.700	2.193	0.655	2.269	0.612	2.346	0.570	2.424	0.528	2.503	0.488	2.582	0.448	2.661
50	0.913	1.925	0.871	1.987	0.829	2.051	0.787	2.116	0.746	2.182	0.705	2.250	0.665	2.318	0.625	2.387	0.586	2.456	0.548	2.526
55	0.979	1.891	0.940	1.945	0.902	2.002	0.863	2.059	0.825	2.117	0.786	2.176	0.748	2.237	0.711	2.298	0.674	2.359	0.637	2.421
60	1.037	1.865	1.001	1.914	0.965	1.964	0.929	2.015	0.893	2.067	0.857	2.120	0.822	2.173	0.786	2.227	0.751	2.283	0.716	2.338
65	1.087	1.845	1.053	1.889	1.020	1.934	0.986	1.980	0.953	2.027	0.919	2.075	0.886	2.123	0.852	2.172	0.819	2.221	0.786	2.272
70	1.131	1.831	1.099	1.870	1.068	1.911	1.037	1.953	1.005	1.995	0.974	2.038	0.943	2.082	0.911	2.127	0.880	2.172	0.849	2.217
75	1.170	1.819	1.141	1.856	1.111	1.893	1.082	1.931	1.052	1.970	1.023	2.009	0.993	2.049	0.964	2.090	0.934	2.131	0.905	2.172
80	1.205	1.810	1.777	1.844	1.150	1.878	1.122	1.913	1.094	1.949	1.066	1.984	1.039	2.022	1.011	2.057	0.983	2.097	0.955	2.135
85	1.236	1.803	1.210	1.834	1.184	1.866	1.158	1.898	1.132	1.931	1.106	1.965	1.080	1.999	1.053	2.033	1.027	2.068	1.000	2.104
90	1.264	1.798	1.240	1.827	1.215	1.856	1.191	1.886	1.166	1.917	1.141	1.948	1.116	1.979	1.091	2.012	1.066	2.044	1.041	2.077
95	1.290	1.793	1.267	1.821	1.244	1.848	1.221	1.876	1.197	1.905	1.174	1.934	1.150	1.963	1.126	1.993	1.102	2.023	1.079	2.054
100	1.314	1.790	1.292	1.816	1.270	1.841	1.248	1.868	1.225	1.895	1.203	1.922	1.181	1.949	1.158	1.977	1.136	2.006	1.113	2.034
150	1.473	1.783	1.458	1.799	1.444	1.814	1.429	1.830	1.414	1.847	1.400	1.863	1.385	1.880	1.370	1.897	1.355	1.913	1.340	1.931
200	1.561	1.791	1.550	1.801	1.539	1.813	1.528	1.824	1.518	1.836	1.507	1.847	1.495	1.860	1.484	1.871	1.474	1.883	1.462	1.896

Table for Durbin – Watson's test: **alpha = 5%,** dL = lower limit, dU = upper limit, n = sample size, k' = number of model regressors without the absolute term.

	$k'=1$		$k'=2$		$k'=3$		$k'=4$		$k'=5$		$k'=6$		$k'=7$		$k'=8$		$k'=9$		$k'=10$	
n	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU
6	0.610	1.400	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
7	0.700	1.356	0.467	1.896	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
8	0.763	1.332	0.559	1.777	0.368	2.287	—	—	—	—	—	—	—	—	—	—	—	—	—	—
9	0.824	1.320	0.629	1.699	0.455	2.128	0.296	2.588	—	—	—	—	—	—	—	—	—	—	—	—
10	0.879	1.320	0.697	1.641	0.525	2.016	0.376	2.414	0.243	2.822	—	—	—	—	—	—	—	—	—	—
11	0.927	1.324	0.758	1.604	0.595	1.928	0.444	2.283	0.316	2.645	0.203	3.005	—	—	—	—	—	—	—	—
12	0.971	1.331	0.812	1.579	0.658	1.864	0.512	2.177	0.379	2.506	0.268	2.832	0.171	3.149	—	—	—	—	—	—
13	1.010	1.340	0.861	1.562	0.715	1.816	0.574	2.094	0.445	2.390	0.328	2.692	0.230	2.985	0.147	3.266	—	—	—	—
14	1.045	1.350	0.905	1.551	0.767	1.779	0.632	2.030	0.505	2.296	0.389	2.572	0.286	2.848	0.200	3.111	0.127	3.360	—	—
15	1.077	1.361	0.946	1.543	0.814	1.750	0.685	1.977	0.562	2.220	0.447	2.472	0.343	2.727	0.251	2.979	0.175	3.216	0.111	3.438
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157	0.502	2.388	0.398	2.624	0.304	2.860	0.222	3.090	0.155	3.304
17	1.133	1.381	1.015	1.536	0.897	1.710	0.779	1.900	0.664	2.104	0.554	2.318	0.451	2.537	0.356	2.757	0.272	2.975	0.198	3.184
18	1.158	1.391	1.046	1.535	0.933	1.696	0.820	1.872	0.710	2.060	0.603	2.257	0.502	2.461	0.407	2.667	0.321	2.873	0.244	3.073
19	1.180	1.401	1.074	1.536	0.967	1.685	0.859	1.848	0.752	2.023	0.649	2.206	0.459	2.396	0.456	2.589	0.369	2.783	0.290	2.974
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991	0.692	2.162	0.595	2.339	0.502	2.521	0.416	2.704	0.336	2.885
21	1.221	1.420	1.125	1.538	1.026	1.669	0.927	1.812	0.829	1.964	0.732	2.124	0.637	2.290	0.547	2.460	0.461	2.633	0.380	2.806
22	1.239	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.940	0.769	2.090	0.677	2.246	0.588	2.407	0.504	2.571	0.424	2.734
23	1.257	1.437	1.168	1.543	1.078	1.660	0.986	1.785	0.895	1.920	0.804	2.061	0.715	2.208	0.628	2.360	0.545	2.514	0.465	2.670
24	1.273	1.446	1.188	1.546	1.101	1.656	1.013	1.775	0.925	1.902	0.837	2.035	0.751	2.174	0.666	2.318	0.584	2.464	0.506	2.613
25	1.288	1.454	1.206	1.550	1.123	1.654	1.038	1.767	0.953	1.886	0.868	2.012	0.784	2.144	0.702	2.280	0.621	2.419	0.544	2.560
26	1.302	1.461	1.224	1.553	1.143	1.652	1.062	1.759	0.979	1.873	0.897	1.992	0.816	2.117	0.735	2.246	0.657	2.379	0.581	2.513
27	1.316	1.469	1.240	1.556	1.162	1.651	1.084	1.753	1.004	1.861	0.925	1.974	0.845	2.093	0.767	2.216	0.691	2.342	0.616	2.470
28	1.328	1.476	1.255	1.560	1.181	1.650	1.104	1.747	1.028	1.850	0.951	1.958	0.874	2.071	0.798	2.188	0.723	2.309	0.650	2.431
29	1.341	1.483	1.270	1.563	1.198	1.650	1.124	1.743	1.050	1.841	0.975	1.944	0.900	2.052	0.826	2.164	0.753	2.278	0.682	2.396
30	1.352	1.489	1.284	1.567	1.214	1.650	1.143	1.739	1.071	1.833	0.998	1.931	0.926	2.034	0.854	2.141	0.782	2.251	0.712	2.363
31	1.363	1.496	1.297	1.570	1.229	1.650	1.160	1.735	1.090	1.825	1.020	1.920	0.950	2.018	0.879	2.120	0.810	2.226	0.741	2.333
32	1.373	1.502	1.309	1.574	1.244	1.650	1.177	1.732	1.109	1.819	1.041	1.909	0.972	2.004	0.904	2.102	0.836	2.203	0.769	2.306
33	1.383	1.508	1.321	1.577	1.258	1.651	1.193	1.730	1.127	1.813	1.061	1.900	0.994	1.991	0.927	2.085	0.861	2.181	0.795	2.281
34	1.393	1.514	1.333	1.580	1.271	1.652	1.208	1.728	1.144	1.808	1.080	1.891	1.015	1.979	0.950	2.069	0.885	2.162	0.821	2.257
35	1.402	1.519	1.343	1.584	1.283	1.653	1.222	1.726	1.160	1.803	1.097	1.884	1.034	1.967	0.971	2.054	0.908	2.144	0.845	2.236
36	1.411	1.525	1.354	1.587	1.295	1.654	1.236	1.724	1.175	1.799	1.114	1.877	1.053	1.957	0.991	2.041	0.930	2.127	0.868	2.216
37	1.419	1.530	1.364	1.590	1.307	1.655	1.249	1.723	1.190	1.795	1.131	1.870	1.071	1.948	1.011	2.029	0.951	2.112	0.891	2.198

	$k'=1$		$k'=2$		$k'=3$		$k'=4$		$k'=5$		$k'=6$		$k'=7$		$k'=8$		$k'=9$		$k'=10$	
n	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU
38	1.427	1.535	1.373	1.594	1.318	1.656	1.261	1.722	1.204	1.792	1.146	1.864	1.088	1.939	1.029	2.017	0.970	2.098	0.912	2.180
39	1.435	1.540	1.382	1.597	1.328	1.658	1.273	1.722	1.218	1.789	1.161	1.859	1.104	1.932	1.047	2.007	0.990	2.085	0.932	2.164
40	1.442	1.544	1.391	1.600	1.338	1.659	1.285	1.721	1.230	1.786	1.175	1.854	1.120	1.924	1.064	1.997	1.008	2.072	0.945	2.149
45	1.475	1.566	1.430	1.615	1.383	1.666	1.336	1.720	1.287	1.776	1.238	1.835	1.189	1.895	1.139	1.958	1.089	2.002	1.038	2.088
50	1.503	1.585	1.462	1.628	1.421	1.674	1.378	1.721	1.355	1.771	1.291	1.822	1.246	1.875	1.201	1.930	1.156	1.986	1.110	2.044
55	1.528	1.601	1.490	1.641	1.452	1.681	1.414	1.724	1.374	1.768	1.334	1.814	1.294	1.861	1.253	1.909	1.212	1.959	1.170	2.010
60	1.549	1.616	1.514	1.652	1.480	1.689	1.444	1.727	1.408	1.767	1.372	1.808	1.335	1.850	1.298	1.894	1.260	1.939	1.222	1.984
65	1.567	1.629	1.536	1.662	1.503	1.696	1.471	1.731	1.438	1.767	1.404	1.805	1.370	1.843	1.336	1.882	1.301	1.923	1.266	1.964
70	1.583	1.641	1.554	1.672	1.525	1.703	1.494	1.735	1.464	1.768	1.433	1.802	1.401	1.837	1.369	1.873	1.337	1.910	1.305	1.948
75	1.598	1.652	1.571	1.680	1.543	1.709	1.515	1.739	1.487	1.770	1.458	1.801	1.428	1.834	1.399	1.867	1.369	1.901	1.339	1.935
80	1.611	1.662	1.586	1.688	1.560	1.715	1.534	1.743	1.507	1.772	1.480	1.801	1.453	1.831	1.425	1.861	1.397	1.893	1.369	1.925
85	1.624	1.671	1.600	1.696	1.575	1.721	1.550	1.747	1.525	1.774	1.500	1.801	1.474	1.829	1.448	1.857	1.422	1.886	1.396	1.916
90	1.635	1.679	1.612	1.703	1.589	1.726	1.566	1.751	1.542	1.776	1.518	1.801	1.494	1.827	1.469	1.854	1.445	1.881	1.420	1.909
95	1.645	1.687	1.623	1.709	1.602	1.732	1.579	1.755	1.557	1.778	1.535	1.802	1.512	1.827	1.489	1.852	1.465	1.877	1.442	1.903
100	1.654	1.694	1.634	1.715	1.613	1.736	1.592	1.758	1.571	1.780	1.550	1.803	1.528	1.826	1.506	1.850	1.484	1.874	1.462	1.898
150	1.720	1.746	1.706	1.760	1.693	1.774	1.679	1.788	1.665	1.802	1.651	1.817	1.637	1.832	1.622	1.847	1.608	1.862	1.594	1.877
200	1.758	1.778	1.748	1.789	1.738	1.799	1.728	1.810	1.718	1.820	1.707	1.831	1.697	1.841	1.686	1.852	1.675	1.863	1.665	1.874

	$k'=11$		$k'=12$		$k'=13$		$k'=14$		$k'=15$		$k'=16$		$k'=17$		$k'=18$		$k'=19$		$k'=20$	
n	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU
16	0.098	3.503	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
17	0.138	3.378	0.087	3.557	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
18	0.177	3.265	0.123	3.441	0.078	3.603	—	—	—	—	—	—	—	—	—	—	—	—	—	—
19	0.220	3.159	0.160	3.335	0.111	3.496	0.070	3.642	—	—	—	—	—	—	—	—	—	—	—	—
20	0.263	3.063	0.200	3.234	0.145	3.395	0.100	3.542	0.063	3.676	—	—	—	—	—	—	—	—	—	—
21	0.307	2.976	0.240	3.141	0.182	3.300	0.132	3.448	0.091	3.583	0.058	3.705	—	—	—	—	—	—	—	—
22	0.349	2.897	0.281	3.057	0.220	3.211	0.166	3.358	0.120	3.495	0.083	3.619	0.052	3.731	—	—	—	—	—	—
23	0.391	2.826	0.322	2.979	0.259	3.128	0.202	3.272	0.153	3.409	0.110	3.535	0.076	3.650	0.048	3.753	—	—	—	—
24	0.431	2.761	0.362	2.908	0.297	3.053	0.239	3.193	0.186	3.327	0.141	3.454	0.101	3.572	0.070	3.678	0.044	3.773	—	—

	$k' = 11$		$k' = 12$		$k' = 13$		$k' = 14$		$k' = 15$		$k' = 16$		$k' = 17$		$k' = 18$		$k' = 19$		$k' = 20$	
n	***dL***	***dU***	***dL***	***dU***	***dL***	***dU***	***dL***	***dU***	***dL***	***dU***	***dL***	***dU***	***dL***	***dU***	***dL***	***dU***	***dL***	***dU***	***dL***	***dU***
25	0.470	2.702	0.400	2.844	0.335	2.983	0.275	3.119	0.221	3.251	0.172	3.376	0.130	4.494	0.094	3.604	0.065	3.702	0.041	3.790
26	0.508	2.649	0.438	2.784	0.373	2.919	0.312	3.051	0.256	3.179	0.205	3.303	0.160	3.420	0.120	3.531	0.087	3.632	0.060	3.724
27	0.544	2.600	0.475	2.730	0.409	2.859	0.348	2.987	0.291	3.112	0.238	3.233	0.191	3.349	0.149	3.460	0.112	3.563	0.081	3.658
28	0.578	2.555	0.510	2.680	0.445	2.805	0.383	2.928	0.325	3.050	0.271	3.168	0.222	3.283	0.178	3.392	0.138	3.495	0.104	3.592
29	0.612	2.515	0.544	2.634	0.479	2.755	0.418	2.874	0.359	2.992	0.305	3.107	0.254	3.219	0.208	3.327	0.166	3.431	0.129	3.528
30	0.643	2.477	0.577	2.592	0.512	2.708	0.451	2.823	0.392	2.937	0.337	3.050	0.286	3.160	0.238	3.266	0.195	3.368	0.156	3.465
31	0.674	2.443	0.608	2.553	0.545	2.665	0.484	2.776	0.425	2.987	0.370	2.996	0.317	3.103	0.269	3.208	0.224	3.309	0.183	3.406
32	0.703	2.411	0.638	2.517	0.576	2.625	0.515	2.733	0.457	2.840	0.401	2.946	0.349	3.050	0.299	3.153	0.253	3.252	0.211	3.348
33	0.731	2.382	0.668	2.484	0.606	2.588	0.546	2.692	0.488	2.796	0.432	2.899	0.379	3.000	0.329	3.100	0.283	3.198	0.239	3.293
34	0.758	2.355	0.695	2.454	0.634	2.554	0.575	2.654	0.518	2.754	0.462	2.854	0.409	2.954	0.359	3.051	0.312	3.147	0.267	3.240
35	0.783	2.330	0.722	2.425	0.662	2.521	0.604	2.619	0.547	2.716	0.492	2.813	0.439	2.910	0.388	3.005	0.340	3.099	0.295	3.190
36	0.808	2.306	0.748	2.398	0.689	2.492	0.631	2.586	0.575	2.680	0.520	2.774	0.467	2.868	0.417	2.961	0.369	3.053	0.323	3.142
37	0.831	2.285	0.772	2.374	0.714	2.464	0.657	2.555	0.602	2.646	0.548	2.738	0.495	2.829	0.445	2.920	0.397	3.009	0.351	3.097
38	0.854	2.265	0.796	2.351	0.739	2.438	0.683	2.526	0.628	2.614	0.575	2.703	0.522	2.792	0.472	2.880	0.424	2.968	0.378	3.054
39	0.875	2.246	0.819	2.329	0.763	2.413	0.707	2.499	0.653	2.585	0.600	2.671	0.549	2.757	0.499	2.843	0.451	2.929	0.404	3.013
40	0.896	2.228	0.840	2.309	0.785	2.391	0.731	2.473	0.678	2.557	0.626	2.641	0.575	2.724	0.525	2.808	0.477	2.892	0.430	2.974
45	0.988	2.156	0.938	2.225	0.887	2.296	0.838	2.367	0.788	2.439	0.740	2.512	0.692	2.586	0.644	2.659	0.598	2.733	0.553	2.807
50	1.064	2.103	1.019	2.163	0.973	2.225	0.927	2.287	0.882	2.350	0.836	2.414	0.792	2.479	0.747	2.544	0.703	2.610	0.660	2.675
55	1.129	2.062	1.087	2.116	1.045	2.170	1.003	2.225	0.961	2.281	0.919	2.338	0.877	2.396	0.836	2.454	0.795	2.512	0.754	2.571
60	1.184	2.031	1.145	2.079	1.106	2.127	1.068	2.177	1.029	2.227	0.990	2.278	0.951	2.330	0.913	2.382	0.874	2.434	0.836	2.487
65	1.231	2.006	1.195	2.049	1.160	2.093	1.124	2.138	1.088	2.183	1.052	2.229	1.016	2.276	0.980	2.323	0.944	2.371	0.908	2.419
70	1.272	1.986	1.239	2.026	1.206	2.066	1.172	2.106	1.139	2.148	1.105	2.189	1.072	2.232	1.038	2.275	1.005	2.318	0.971	2.362
75	1.308	1.970	1.277	2.006	1.247	2.043	1.215	2.080	1.184	2.118	1.153	2.156	1.121	2.195	1.090	2.235	1.058	2.275	1.027	2.315
80	1.340	1.957	1.311	1.991	1.283	2.024	1.253	2.059	1.224	2.093	1.195	2.129	1.165	2.165	1.136	2.201	1.106	2.238	1.076	2.275
85	1.369	1.946	1.342	1.977	1.315	2.009	1.287	2.040	1.260	2.073	1.232	2.105	1.205	2.139	1.177	2.172	1.149	2.206	1.121	2.241
90	1.395	1.937	1.369	1.966	1.344	1.995	1.318	2.025	1.292	2.055	1.266	2.085	1.240	2.116	1.213	2.148	1.187	2.179	1.160	2.211
95	1.418	1.929	1.394	1.956	1.370	1.984	1.345	2.012	1.321	2.040	1.296	2.068	1.271	2.097	1.247	2.126	1.222	2.156	1.197	2.186
100	1.434	1.923	1.416	1.948	1.393	1.974	1.371	2.000	1.347	2.026	1.324	2.053	1.301	2.080	1.277	2.108	1.253	2.135	1.229	2.164
150	1.579	1.892	1.564	1.908	1.550	1.924	1.535	1.940	1.519	1.956	1.504	1.972	1.489	1.989	1.474	2.006	1.458	2.023	1.443	2.040
200	1.654	1.885	1.643	1.896	1.632	1.908	1.621	1.919	1.610	1.931	1.599	1.943	1.588	1.955	1.576	1.967	1.565	1.979	1.554	1.991

Název:	**Statistical Methods for Economists**
Autor:	**Ing. Filip Tošenovský, Ph.D.**
Vydavatel:	Slezská univerzita v Opavě Obchodně podnikatelská fakulta v Karviné
Určeno:	studentům SU OPF Karviná
Počet stran:	162
Vydání:	on-line
ISBN:	**978-80-7510-033-7**