General and Economic Statistics

Julianna Csugány

GENERAL AND ECONOMIC STATISTICS

Julianna Csugány

Líceum Kiadó
Eger, 2015

Edited by:
Julianna Csugány

Series editor
Lajos Kis–Tóth PhD, Professor

Reader
Tamás Tánczos PhD.

Authors
Julianna Csugány

ISBN 978-615-5509-65-0

Contents

1. INTRODUCTION

1.1 GOALS, COMPETENCIES, AND COURSE REQUIREMENTS

1.1.1 Goals

Understanding economic and social processes involves processing a large amount of information. That is what the apparatus of statistics can make more efficient. Numerical analyses allow you to succinctly describe data, to draw well-founded conclusions, and to mathematically justify observed tendencies and interrelationships. Statistical work begins with gathering and organizing data, which is followed by data analysis. In the final stage, results are interpreted, sometimes also illustrated, and conclusions are drawn.

In the General and economic statistics course, students will learn about and acquire the knowledge of

- basic concepts and technical terminology of statistics
- ways of statistical data collection and organization
- types of ratios and the logic of their derivation
- application areas where practical use is made of ratios, especially the labor market, and financial and demographic indicators
- the tools of descriptive statistics, the properties and interpretation of the characteristics of basic distribution
- possibilities of the comparison of complex ratios (grand means) based on standardization
- the method of value-weighted index calculation
- areas of the practical application of index calculation, especially measuring the performance of the national economy and external trade statistics
- means of the analysis of the relationship between attributes, association, correlation, and methods of regression calculation
- techniques of time series analysis which make forecasting possible
- basic principles and forms of graphical representation of results

1.1.2 Competencies

In order for students to be able to complete the course in General and economic statistics, they must acquire the following competencies, which vary thematically.

Students should

- be familiar with and correctly use the basic concepts of statistics
- know the relevant statistical data sources, from which they can gather secondary information for the study of social and economic phenomena
- be familiar with forms of statistical data collection and organization and be able to use the knowledge acquired
- acquire the method of deriving ratios and be able to carry out simple analyses with various forms of ratios
- be able to identify and resolve economic and social problems, which are amenable to statistical analysis involving the use of ratios
- be able to use the apparatus of descriptive statistics and interpret characteristics of basic distributions
- be able to compare complex ratios (grand means) based on standardization, and quantify differences and the effects of factors causing change
- acquire the method of value-weighted index calculation and be able to determine the change of quantity and value for a product or multiple products in a relative fashion
- be able to identify areas where the methods of index calculation are applicable
- acquire the methods of studying the relationships between attributes, and be able to distinguish stochastic relations
- be able to use and interpret indices of associations, complex relationships and correlations
- be able to characterize a relationship between quantitative attributes with a binary linear regression function
- be able to employ time series analysis techniques for making forecasts
- acquire basic principles of graphical representation and be able to use various kinds of diagrams to illustrate the results of analyses.

1.1.3 Course requirements

For the successful completion of the course in General and economic statistics, students need to

- be familiar with basic concepts of statistics and forms of data organization, thus, types and properties of statistical rows and tables
- be able to select the appropriate sampling technique for the study of a particular phenomenon
- have an overview of the system of ratios, be able to name the types of ratios encountered in everyday life
- know the major derivable ratios that serve as labor-market, financial, and demographic indicators and the way to determine them
- have an overview of the statistical apparatus, know the properties of indices and be able to interpret the results
- know methods of comparing complex ratios based on standardization, understand relationships between factors that cause differences or change, and be able to interpret the components
- be familiar with the method of defining value in statistical terms and recognize forms in which it appears, be familiar with the methods of calculating price, value, and volume indices, be able to interpret them and understand interrelationships
- be familiar with areas of the practical application of index calculation, indices that are used to measure the performance of companies and the national economy, their content and forms of calculation
- be familiar with indices in associations and complex relationships, know when to use which kind of index and be able to interpret the computed results
- be familiar with and interpret indices suitable for the identification of correlative relationships
- possess basic knowledge of regression calculation, be able to mention examples of different types of regression, be able to interpret the parameters of a binary linear regression, and decide which type of function matches best the empirical values
- have acquired the basics of deterministic time series analysis, be able to describe the factors of a time series, be familiar with the

methods of determining trends and seasonality and understand the relationships between them

- be familiar with the basic principles of graphical representation, types of graphical diagrams and their applications.

1.2 CONTENT OF THE COURSE

The course in General and economic statistics is composed of five closely related modules. The first module is concerned with elementary concepts of statistics and methods of data collection and organization. The second, third and fourth modules introduce students to methods of statistical analysis on a thematic basis, initially focusing on simple analysis techniques and then moving on to issues in the analysis of heterogeneous populations and statistical methods of demonstrating interrelationships and tendencies.

The final module discusses questions of the presentation of the results of analyses and the possibilities for the practical application of the knowledge and skills acquired during the course.

The Appendices at the end of the course material include a bibliography and a summary of diagrams, tables, and media elements. The final section, called Tests, includes questions and problems that can be used by students to test their understanding and skills in the application of the content of the course.

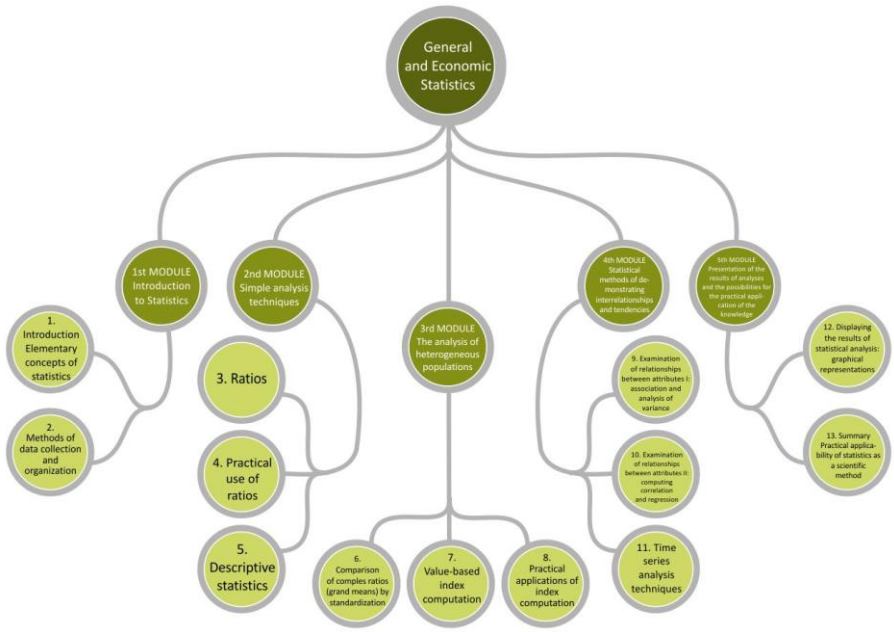


Figure 1 The logical structure of the course in General and economic statistics

1.3 STUDY TIPS

A mind map at the beginning of each unit helps you navigate through the course and understand the general structure of the methodological tools. The self-test questions and problems which conclude each chapter should help you better acquire the knowledge and skills, and complete the course successfully.

1.4 BASIC CONCEPTS OF STATISTICS

We need to clarify the concepts that we will regularly use, before we begin to discuss statistical methods, because familiarity with the meaning and use of basic concepts is essential to the application of the tools of statistical analysis.

A population is an entire collection of the elements of a phenomenon of interest.

A population may be *finite* or *infinite* in regard to the number of elements it contains. We rarely encounter infinite populations in practice. Populations may differ in terms of whether the data relate to a point or a period of time.

Data about a *static population* (called *stock data*) reflect an unchanging state. Such data may be interpreted with respect to a particular point in time or in reference to observations about a past period of time, such as, for example, data about the population of a country. Data about a *dynamic population* (called *flow data*), by contrast, reflect a process and are interpreted as referring to a particular period of time, such as, for example, the traffic in a street on a particular day. Statistical populations can also be subcategorized on the basis of whether their elements can be unambiguously identified. If the elements of a statistical population can be clearly discriminated, then it is called a *discrete population*. The population of a country, for example, is a discrete statistical population, as its elements, the people, can be clearly discriminated and identified. A population whose elements are defined arbitrarily, i.e., whose elements are not clearly discriminated, is called a *continuous population*. Choice of the basic unit in a continuous population is arbitrary, as in an analysis of soft drink consumption, for example, where the consumed amount may be expressed either in liters or in deciliters, i.e., the choice of the measure of capacity is arbitrary.

Elements of a statistical population may have different properties, which offer a basis of comparison and classification.

Properties of elements of a statistical population are called attributes.

A statistical attribute may be *shared* or *distinctive* depending on whether it applies to every element of the population or only to a subset of elements. An attribute, depending on what type of characteristics it expresses, may be temporal, areal, quantitative or qualitative. A *temporal attribute* is a feature of elements defined in terms of a point or period of time. An *areal attribute* connects the elements of a population to a geographically identified areal unit. A *quantitative attribute* is a feature of elements that can be measured and expressed numerically, while a *qualitative attribute* is a property that does not belong in any of the other categories.

A quantitative attribute may be *discrete*, when it can be expressed by an integer because there is no gradual transition between the attribute variants, or *continuous,* when it cannot be expressed by an integer because there is a gradual transmission between the attribute variants.

- Describe the students on statistics course with various types of attributes.

1. *Examples of types of attributes*

Type of attribute	Name of attribute
Temporal	Students' dates of birth
Areal	Students' place of residence
Quantitative	Students' height
Qualitative	Students' gender

In regard to a particular attribute, its manifestations in a population are called the variants of the attribute.

If only two variants of an attribute occur in a population, it is called an *alternative attribute*. An example is gender, a qualitative attribute, which is realized in two alternative variants, male and female. However, elements in a population are more frequently distinguished from one another by several different variants of attributes, which allow for a more sophisticated analysis. Variants of quantitative attributes are called *attribute values*, as these are properties of elements that can be expressed numerically.

Mathematically based statistical methods can be applied not only in analyses of quantifiable properties. It is important, however, that we understand the measurability of various attributes, which fundamentally determines the applicable methods. For the sake of statistical analysis, we assign numerical values to non-numerical attribute variants as well.

The numerical values to be assigned to variants of a particular attribute are arranged on a measurement scale.

The significance of measurement scales becomes conspicuous in complex analyses, where we want to analyze the relationship between several different variables on different levels of measurement. Measurement scales relate to one another hierarchically, such that each level of measurement that is higher in the hierarchy possesses all the features of the one below it. Particular attribute types relate to levels of measurement.

Nominal scales, which are typically used in the analysis of qualitative and areal attributes, distinguish between the variants of an attribute. A numerical characterization, which makes large amounts of data more transparent, makes analyses easier, though the values in such cases cannot be used mathematically. An example of a qualitative attribute which can be measured on a nominal scale is the registration number of a car, where the numbers have a distinguishing function, but carry no

mathematical content. *Ordinal scales* also express the order of the values of an attribute. Consumer satisfaction or educational grades can be measured on ordinal scales, for example. Such scales make sophisticated analyses possible beyond a mere comparison of attribute variants, as they also carry information about which value is better than or superior in some sense to another. An interesting example is the level of measurement of winning lottery numbers. It is easy to make the mistake of including such numbers on the wrong scale, either because they are drawn one after the other, or because they are eventually arranged in a sequence. It is important to bear in mind that ordering is not only a matter of mathematics but also a matter of content. It is easy to see that the order of lottery numbers has no statistical significance, because the only thing that matters is that the numbers are different. As drawing a higher or lower number has no additional meaning, lottery numbers are measured on a nominal scale. *Interval scales* allow a meaningful interpretation not only of the order of attribute values but also the differences between them. On such a scale, we not only know if a value is superior to another but we also know the degree to which it is superior. Values, however, cannot be totaled. Temperature is a typical example of an attribute that is measurable on an interval scale. Take two different days, for example, such that the temperature was $+20°C$ on one and $+15°C$ on the other. Measuring temperature values on an interval scale reveals that the temperature dropped by $5°C$. This shows that the difference between the temperature values is meaningfully interpretable, but the values cannot be totaled, nor can any other mathematical operation be carried out with them. *Ratio scales* represent the highest level of measurement. On this level only variants of quantitative attributes can be measured, as any mathematical operation can be carried out on the attribute values represented on such scales. For example, ratio scales can be used to measure traffic, income, which can be totaled, their differences as well as their products can be computed, and their attribute values can be proportioned.

Statistical analysis begins with gathering data about a phenomenon you desire to study. Data gathered may be complete or partial, and data may come from primary or secondary sources. In statistical analysis we work with statistical data, or indices.

Statistical data are numerical features of a population or the number of elements in a population.

Statistical data relate to concrete phenomena in space and time, or some other form, so they are not merely numerical values. For example, if a particular company's annual revenue is EUR 12 000, then it is significant

that this statistical data relates to that specific company. Another example of statistical data is the fact that 6520 children were born in Hungary in 2010. Thus, statistical data are numerical values that relate to a particular population. Statistical data may be absolute in that they come from direct measurements or counts. Statistical data may also be obtained from mathematical operations. Such data are derived values, such as, for example, the data that the average amount of fruit consumption in a country is 62 kg, or the value of population density derived from relating the population in an area to the area.

Statistical data that serve to characterize recurrent phenomena are called statistical indices.

1.4.1 Sources of statistical data

A variety of different indices are used in characterizations of economic and social phenomena. Databases of international organizations are regarded as official data sources, which contain data on different countries or groups of countries and on different topics. Data concerning the same phenomenon may vary depending on which data source they come from. A typical example is data on the GDP of a country. Therefore, it is a good idea to read about questions of data collection and methodology on a particular organization's web site, as well as the content of indices, because it may have an effect on the use of the data.

Eurostat, the statistical office of the European Union, offers data on European countries based on a variety of different systems of indices.

1. Eurostat: http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/

The largest databases on countries of the world belong to OECD (Organization for Economic Co-operation and Development), the United Nations (UN), and The World Bank:

2. OECD database: http://stats.oecd.org/
3. UN database: http://www.un.org/en/databases/
4. World Bank database: http://data.worldbank.org/

Some databases gather data on specific areas. The database of the International Monetary Fund (IMF) contains primarily financial indicators, in addition to numerous economic indicators.

5. Database of IMF: https://www.imf.org/external/data.htm

The International Monetary Fund possesses various databases containing specific financial statistics.

6. International Financial Statistics:
   http://www.econdata.com/databases/imf-and-other-international/ifs/

The International Labour Organization (ILO) is a source of statistics on the international labor market:

7. International Labour Organization database:
   http://www.ilo.org/global/statistics-and-databases/lang--en/index.htm

A widely used source of information on international trade is another special UN database:

8. United Nations Conference on Trade and Development:
   http://unctad.org/en/Pages/Publications/Handbook-of-Statistics.aspx

2. FORMS OF STATISTICAL DATA COLLECTION AND ORGANIZATION

2.1 GOALS AND COMPETENCIES

The first step in statistical work is to gather data. This is followed by organizing the data. Data may come from primary or secondary sources, depending on the nature of the phenomenon of interest. Choice of appropriate statistical methods presupposes the organization of data. The basic unit of statistical data organization is the statistical row. Statistical rows may be combined into statistical tables.

The purpose of this unit is to familiarize students with various ways of statistical data collection and organization. Students will learn about primary and secondary forms of statistical data collection, various techniques of sampling, and their characteristics. Ways of data collection will be presented through practical examples to enable students to select the appropriate sampling technique in the study of a particular phenomenon. It is essential for students to be familiar with the basic forms and criteria of data organization, because this provides the basis for the choice of the appropriate analytic technique.

2.2 TOPICS

We want statistical data to meet three conditions: they must be accurate, quickly available, and inexpensive. What we mean by these requirements is that data need to be sufficiently accurate, they need to be obtainable quickly, so we can used them as soon as we need them, and that we should be able to access them at the lowest possible cost. It is hard for data to meet all three conditions, but we must keep trying. Data may be obtained from either primary or secondary sources. Secondary sources include various company reports or records, and publications issued by offices, such as the statistical office, local governments, or the central bank. Data come from a primary source when researchers themselves gather the data they need. When using secondary sources, one needs to bear in mind that errors may occur in gathering data, which may affect the results of a statistical analysis. We rarely have the occasion to gather information about each element of a phenomenon in real life, therefore we often study only a subpart of a population and use various methods to attempt to attain general conclusions that apply to the entire population.

- **A sample is a subset of elements in a population involved in the study of a phenomenon of interest.**

2.2.1 Sampling techniques

If we have information about all the elements of a phenomenon of interest, then our data are exhaustive. This is possible only for finite populations. It is rare. Census is an example.[1] To gather information about an entire population is costly and time-consuming in most cases. Therefore, we typically gather partial data, which apply only to a subset of a population. In partial data collection, the sampling process needs to be planned carefully. As representativeness is an essential condition in statistical analysis, we focus on methods of representative sampling. If in the analysis of data you focus exclusively on relationships within the body of data, without the intention to make generalizations about the entire population, then your observations are not representative.

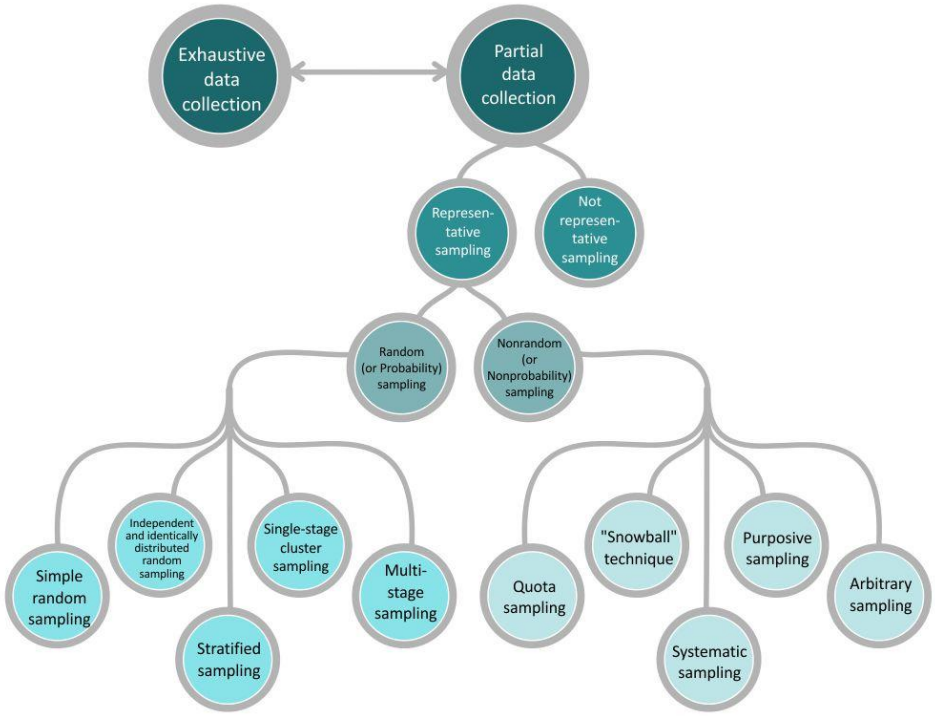


Figure 2 Ways of statistical data collection

[1] When conducting a census, you want to gather information about each member of the p opulation, though this is not always possible, due to difficulties that have to do with the accessibility of the subjects and their willingness to respond.

- **A sample is representative if the elements of the sample are selected along certain principles, and thus it characterizes the base population, which allows us to draw conclusions about the entire population. Any sampling error resulting from the incompleteness of observations of the population can be explicitly identified.**

An important aspect of sampling is the ratio of selection, which is defined as the quotient of the number of elements in the sample and the number of elements in the population. You can decrease the sampling error by increasing the size of the sample. An important factor in the selection of the sampling technique is whether the population is homogenous or heterogeneous. A population is homogenous if all of its elements are uniform from the perspective of the phenomenon of interest. For example, if we gather information about people's coffee consumption habits, the respondents' gender is not particularly relevant, so the population is homogenous from the perspective of the study. A population is heterogeneous if its elements are different from the perspective of the phenomenon of interest. For example, if we were to study the satisfaction of the employees of a company with their salaries, then the status of the employees we intend to ask would matter. Another example could be a survey on college students' entertainment habits. From this perspective, the population could be regarded as homogeneous, since what matters is that all the members are students, regardless of differences in gender or their majors. If, however, we study their learning habits, then the major subject of a student or the year of study would count as relevant differences. So, the distinction between homogenous and heterogeneous populations is determined by the nature of the phenomenon of interest.

Random (or Probability) sampling

A universal feature of random sampling techniques is that the elements of a sample are selected randomly, where each element of a population has the same predetermined probability of being selected as a member of the sample. Sampling a homogeneous population may be carried out by *independent and identically distributed* random sampling (IID) or by *simple random sampling* (SS), depending on the number of elements in the population. Simple random sampling is most commonly used for finite populations, in which each element may be selected as a member of the sample with the same probability. The IID method may be employed for large or infinite populations, under similar conditions. For heterogeneous populations, *stratified sampling* is the preferred technique, in which the population is divided into well-defined layers prior to

sampling, and then elements are selected from individual layers separately. If the list of the elements in a population is not known and we only know the groups of such elements, then we employ the indirect method of *single-stage cluster sampling*. This is a simple sampling technique, which indirectly delivers relevant elements of a population. A similar approach called *multi-stage sampling* may be employed when the clusters are removed farther from the elements of interest.

The minsiter of eduction wants to assess the state of higher education and is interested in students' opinion about it.

Sampling method	Elements included in the sample
Simple random sampling	Students randomly selected
Independent identically distributed sampling	Students randomly selected
Stratified sampling	Students are stratified, e.g., by level of program, and students are selected from each layer
Cluster sampling	It is easier for the ministry to obtain information from the students via the institutions of higher education, as it has information about the population broken down to institutions, but has no list of all and only the students in all institutions, therefore they will select an institution by simple sampling and ask its students
Multi-stage cluster sampling	There is no list containing all students, faculties within higher education institutions have their own student records, so the individuals of interest are reached through multiple layers, clusters are selected by simple sampling, and eventually the students are reached and asked exhaustively or randomly

Nonrandom (or Nonprobability) sampling

Nonrandom sampling is used when the kind of elements in a sample is of particular significance from the perspective future analysis. Therefore, we set up criteria before sampling, which the sampling has to meet. In *quota sampling,* we determine the composition of the sample prior to sampling by specifying the amount of elements that possess a particular attribute variant in the sample, such as, for example, the number of men and women the sample has to contain, or the number of respondents in different age groups.

The *snowball sampling technique* is a very popular sampling technique these days. Questionnaires forwarded through the Internet are typical examples which illustrate the main idea of this technique. Such questionnaires gather members into the sample from the circle of friends of a particular person. This technique, however, may fail to include members in the sample who are otherwise important from the perspective of the analysis.

In *purposive sampling,* the purpose to achieve is for dominant elements to be represented in a higher ratio in the sample, thus making analysis easier. For example, if you want to know what people think about products made for women, men, or children, then you should ask the target groups about them, as they are most likely to be able to judge those products.

In *systematic sampling,* you gather data by starting from a point and then proceed step by step, for example, when you ask one out of every five people you meet on the street. It may be considered a form of random sampling from a research-methodological perspective, as sampling begins at a randomly selected point in time and elements are also randomly selected and entered into a sample. From a statistical perspective, however, it may be considered nonrandom, because it makes use of purposefully determined intervals.

Arbitrary sampling may affect representativeness severely, as the sampler selects elements for a sample subjectively, which may lead to serious issues during research.

The minsiter of eduction wants to assess the state of higher education and is interested in students' opinion about it.

Sampling method	Elements included in the sample
Quota sampling	Asks a specific number of students at each higher education institution
"Snowball" technique	Draws up a questionnaire and then sends it out to a selected higher education institution, where students fill it out, and then the institution and the students forward it to other higher education institutions.
Systematic sampling	One out of every 10 students, e.g., on the higher education records is selected and entered in the sample.
Purposive sampling	Students are selected from specific institutions.
Arbitrary sampling	Students are selected on the basis of an arbitrary criterion, such as academic achievement, for example.

2.2.2 Organizing statistical data

It is important to have your data well organized before starting statistical analysis, because it can make it easier to select the appropriate method. We organize data into statistical rows first, and then we can organize the rows into tables.

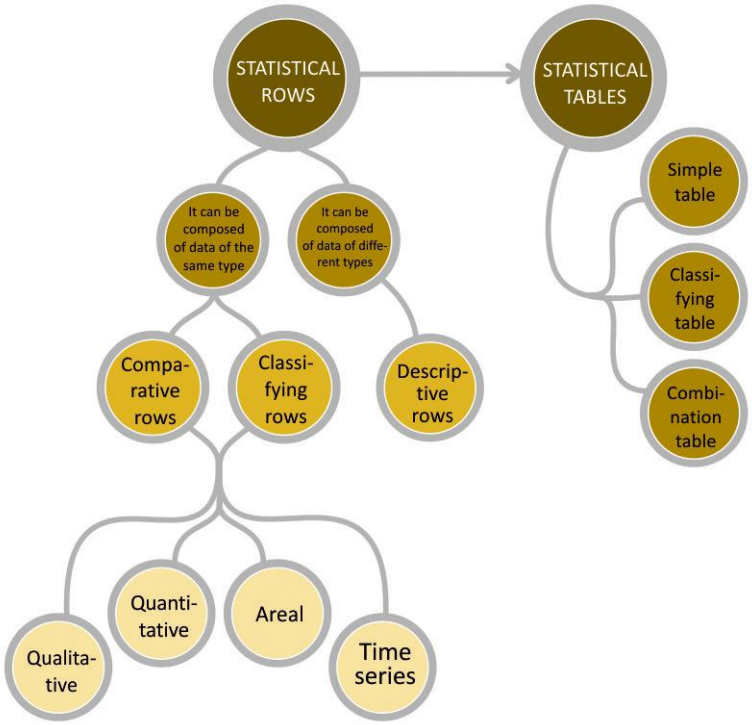


Figure 3 Types of statistical rows and tables

Statistical rows

A statistical row is the basic unit of statistical data organization. It can be composed of data of the same type or of different types.

A statistical row is the arrangement of available data according to a specific attribute.

We can organize different types of data into ***descriptive rows***. Such data will refer to the same population, but characterize it from different perspectives. For example, if we intend to characterize the tourist industry of a country, we may arrange the data in a row like the one below.

2. *Some data on Hungary's tourist industry in 2013*

Year	Visitors from abroad (thousand people)	Total amount of time spent (thousand days)	Average amount of time spent (days/person)
2013	43 665	101 730	2.3

Source: *HCSO* (2014)

The phenomenon of interest here is Hungary's tourist industry. The population is the set of visitors from abroad. One characteristic feature of that population is the total number of its elements, and another one is the total number of days spent. Finally, the data derived from these two is their quotient, the average amount of time spent.

Statistical data of the same kind may be compared with each other or classified. Accordingly, we distinguish between ***comparative rows*** and ***classifying rows***. Depending on the attribute employed, comparison or classification can be areal, quantitative, qualitative, or a time series. Statistical rows composed of the same type of data may differ according as to whether the values can be summed. Classifying rows may contain a sum cell if all attribute variants are included. It is important that the sum have some practical significance, as mathematically, any numbers may be added up, but classifying rows may only contain data where the sum makes some practical sense. If not all attribute variants are included, then we can create a comparative row. The most typical kind of comparative rows is the time series, an example of which is the representation of recent changes in the number of visitors from abroad.

3. *Changes in the number of visitors from abroad in Hungary between 2011 and 2013*

Year	Number of visitors from abroad (thousand people)
2011	41 304
2012	43 565
2013	43 665

Source: *HCSO* (2014)

In this case, it would be impractical to total the years. All we can say is how many visitors arrived during those three years altogether. Studying a particular year may be different. If data are given for quarters, then summing the quarterly data yields the yearly sum, which makes practical sense.

4. *Changes in the number of visitors from abroad in Hungary in 2013 in quarters*

Period	Number of visitors from abroad (thousand people)
2013. I. quarter	8 363
2013. II. quarter	10 640
2013. III. quarter	14 813
2013. IV. quarter	9 849
TOTAL	***43 665***

Source: *HCSO* (2014)

In comparative rows, attribute variants are called classes and the number of elements belonging to particular attribute variants is called their frequency.

The total number of frequencies equals the number of elements observed.
In the quarterly report on the number of visitors from abroad, the population is composed of 43 665 elements, which are distinguished by the year in which they arrived. The total number of elements in each quarter is their frequency, which shows how many elements the particular class is composed of, i.e., how many visitors from abroad arrived in each quarter.

In classifying data, classes are easily derived if we have a small number of attribute variants. Ranking may help to have an overview of the data, as in arranging course grades, for example.

We know the end-term statistics examination grades of 20 students:

4, 3, 2, 5, 5, 4, 1, 3, 4, 4, 5, 2, 3, 5, 5, 3, 4, 3, 4, 1

If we arrange the data in a monotonically increasing sequence, i.e., in a rank, it is easier to use statistical methods.

1, 1, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5

We can define a frequency array on the basis of the attribute variants in the ranking.

5. *Distribution of students according to grades*

Grade	Number of students (persons)
Excellent (5)	5
Good (4)	6
Average (3)	5
Satisfactory (2)	2
Fail (1)	2
TOTAL	*20*

Classifying quantitative characteristics which involve many attribute variants is not an easy task.

A class interval frequency array is a classifying array according to a quantitative attribute in which classes are defined by class intervals.

In class interval classification, class intervals must be defined in such a way that attribute values belong to one and only one class interval.[2] The length and number of class intervals must be defined in a way that they correctly reflect the composition of a population according to its quantitative attributes. An optimal number of class intervals (k) may be estimated by the following formula:

$$2^k > N$$

According to the rule, you have to increase the number of class intervals (k) until the k^{th} power of 2 exceeds the number of elements (N) of the population. If the class intervals are equal, the length of a class interval (h) may be determined by the following formula:

$$\text{length of class interval (h)} = \frac{\text{largest value of data row} - \text{smallest value of data row}}{\text{number of class intervals (k)}}$$

- We know the number of points achieved by 100 students in a statistics examination. *How many class intervals and of what length can we include if we know that the weakest student achieved 16 points and the best one achieved 98 points?*

[2] Violations of this restriction are conspicuously demonstrated by questionnaires in which age intervals are specified as 18-25 and 25-35, for example. This forces a 25-year-old respondent to categorize themselves in one or the other class interval. This leads to serious issues in evaluation.

For a population with 100 elements: $2^7 = 128 > 100 \rightarrow k=7$ that is, the optimal number of class intervals is 7.

The lowest score achieved by the students is 16 and the highest score is 98, therefore, if we know that the optimal number of class intervals is 7, then with equal class interval classification

$$h = \frac{x_{max} - x_{min}}{k} \rightarrow \frac{98 - 16}{7} = 11.7 \sim 12$$

Thus, we optimally include 7 class intervals, each 12 units in length.

In practice, however, you do not have to determine class intervals of equal length. Non-equal class intervals can highlight aspects of the disproportionateness of the population.

Statistical tables

In statistical analysis, we often need more the one attribute, as our goal is to characterize our data from the perspective of as many different attributes as possible.

A statistical table is composed of mutually related statistical rows.

In terms of its function, a statistical table may be a working table, which is used to carry out calculations, or a presentation table, which is used to present the results of our calculations. A statistical table always has a *title*, which indicates the population described by the data, and circumscribes the topic of inquiry in space and time. It is obligatory to provide the *source* of data in presentation tables. It is not always necessary in working tables, where mathematical calculations are in focus. When you edit a statistical table, your primary concern should be *simplicity* and *clarity*. To this end, you need to specify the units of measurement for your data either in the title of the table, or in the row or column heading. Empty cells are not allowed in statistical tables. Therefore, if some data is not available, we cross the cell out, if some data is not known, we use (…), and if the unit of measurement is small in comparison to other data in the table, then we enter 0.0 as the value. A totaling cell has to be included in a table which contains a classifying row.

Statistical tables may be distinguished on the basis of the kinds of statistical rows they contain, as follows:

- A table is *simple*, if it does not contain a classifying row, i.e., there is no totaling cell, as it is composed only of descriptive and/or comparative rows.

6. *Number of visitors from France and Germany in Hungary and the average amount of time spent in 2013*

Country	Number of visitors (thousand people)	Average amount of time spent (days/person)
Germany	3 059	5.1
France	341	5.4

Source: *HCSO* (2014)

– *classifying table* containing a classifying row and a descriptive and/or a comparative row

7. *Changes in the number of visitors from abroad in Hungary according to purpose of visit in 2012 and 2013*

Purpose of visit	Number of visitors	
	2012	2013
Holiday	13 489	13 546
Business	1390	1341
TOTALS	*14 879*	*14 887*

Source: *HCSO* (2014)

– *combination table* composed exclusively of classifying rows, wherefore data can be summed both horizontally and vertically

8. *Number of visitors in Hungary by destination and place of origin*

Destination / Residence	Balaton	Budapest Central Danube Region	Great Plain	Northern Hungary	Transdanubia	Tisza-Lake	*Totals*
Central Hungary	6 445	4 004	3 903	2 363	4 203	270	*21 188*
Transdanubia	5 388	2 547	1 838	749	8 602	194	*19 318*
Northern Hungary	1 159	1 354	1 328	3 203	528	280	*7 852*
Great Plain	1 989	2 070	4 829	1 335	1 004	315	*11 539*
Totals	*14 981*	*9 975*	*11898*	*7 650*	*14 337*	*1 059*	*59 897*

Source: *HCSO* (2014)

The dimension number of a statistical table shows the number of attributes to which a particular value belongs.

When organizing statistical data, explicitness and clarity should be high priority, as well organized data make it easy to select the necessary methods for analysis.

2.3 SUMMARY AND QUESTIONS

2.3.1 Summary

The first step in doing statistical work is to gather and organize data for the study of a particular phenomenon of interest. Information may be obtained from primary and secondary sources. Data may require different methods of analysis. When gathering primary data, it is important to select the appropriate sampling technique. Elements may be selected in the sample either randomly or in a non-random fashion. If our goal is to draw general conclusions from our analysis of the data, criteria of representativeness must be observed in sampling.

Organizing statistical data immediately precedes data analysis. A fundamental feature of statistical rows and tables is that they capture properties of the phenomenon of interest by attributes in a structured way. Rows may be composed of similar or different data. Thus, we distinguish between comparative and classifying rows. An important feature of statistical tables is that they have a title and, for presentation tables, an indication of the source of data. Statistical tables cannot contain empty cells. Different types of rows and tables offer different opportunities for analysis.

2.3.2 Self-test questions

What is the difference between a population and a sample?

What are the characteristics of primary and secondary data sources?

What is a statistically representative sample?

What random (or probability) sampling techniques do you know?

What nonrandom (or nonprobability) sampling techniques do you know?

What is a statistical row and what is a statistical table?

What are the characteristics of descriptive, comparative, and classifying statistical rows?

What are the characteristics of a simple statistical table?

What are the characteristics of a classifying statistical table?

What are the characteristics of a combination statistical table?

2.3.3 Practice tests

❁ Determine the appropriate sampling technique for the phenomenon of interest. *More than one answer is possible.*

	Random					Nonrandom				
	Independent equivalent distribution	simple random	Stratified	Cluster	Multi-stage	Quota	Snowball	Purposive	Systematic	Arbitrary
Coffee consumption habits	X	X					X			
Television viewing habits		X					X		X	
Salary satisfaction of company employees			X			X				
Car purchasing experiences								X		
High school students' plans for further education				X	X					

❁ Decide whether the statements below are true (T) or false (F).

The representativeness of a sample is important for reasons of drawing general conclusions.	T
A classifying table contains only classifying rows.	F
The number of dimensions of a statistical table shows the number of attributes to which a particular value belongs.	T
You can create comparative and classifying rows from different types of data.	F
A combination table contains different kinds of statistical rows.	F

3. RATIOS

3.1 GOALS AND COMPETENCIES

One of the simplest and most frequently used tools of statistical analysis is the ratio. In a ratio, we derive the quotient of two related items of data of the same type or of different types. Students will learn about various ways of employing the statistical apparatus and methods, which will help them understand problems that they may encounter in their everyday life and which can be solved by simple statistical tools.

The purpose of the unit is to familiarize students with the system of ratios and ways in which particular types of ratios can be computed. Students need to deeply understand the logic of ratios, rather than have a superficial familiarity with them. Not very heavy duty mathematical tasks will help and allow them to think about such questions and not only recognize but also solve problems by employing various types of ratios.

3.2 TOPICS

In statistical analysis, we often used data not in absolute form, but as ratios, which allows us to compare data and express differences in magnitude also.

The quotient of two related items of statistical data is called a ratio.

$$\text{Ratio (V)} = \frac{\text{data to compare (A)}}{\text{basis of comparison (B)}}$$

On the basis of the relationship, from any two terms, the third one can be computed. A ratio is used in such computations in coefficient form:

$$\text{data to compare (A)} = \text{basis of comparison (B)} \cdot \text{ratio (V)}$$

$$\text{basis of comparison (B)} = \frac{\text{data to compare (A)}}{\text{ratio (V)}}$$

We distinguish several different types of ratios depending on the type of data compared. Distribution ratios, coordination ratios, dynamic ratios, areal/spatial comparative ratios, and plan accomplishment ratios are used to represent relationships between statistical data of the same type, while intensity ratios express relationships between data of different types. Ratios of data of the same type may be further sub-classified ac-

cording as to whether they are derived from classifying or comparative rows.

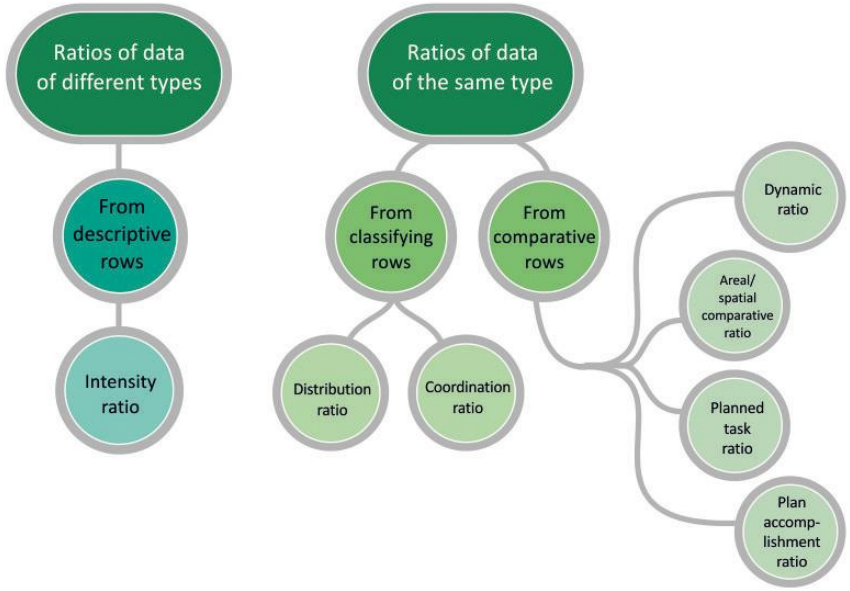


Figure 4 System of ratios

The use of ratios is often motivated by the fact that data which differ either in terms of the unit of measurement or in terms of magnitude are hard to compare. Statistical data converted to ratios often offer much more information than they do in their original form. Therefore, ratios are the most important tools of statistical analysis.

3.2.1 Distribution and coordination ratios

Distribution and coordination ratios are primarily used to represent proportional relationships within a population.

The proportion of a part of a grouped population relative to the entire population is called a distribution ratio.

$$\text{Distribution ratio } (V_m) = \frac{\text{part}}{\text{whole}} \rightarrow V_m = \frac{f_i}{\sum f_i}$$

When calculating the distribution ratio, we divide the data that corresponds to the part of the population by the total of the population, and multiply the derived ratio by 100 and interpret the result in terms of percentage. The

sum of distribution ratios in coefficient form is 1, and in percentage form it is 100%. Distribution ratios are a good way to represent magnitudes and, therefore, the internal structure of a population or a sample.

- **A coordination ratio relates parts of a grouped population to each other.**

$$\text{coordiantion ratio } (V_k) = \frac{\text{part}}{\text{part}} \rightarrow V_k = \frac{f_i}{f_j}$$

Coordination ratios are used less frequently. They serve to express the relationship between parts of a population to each other, i.e., how many elements there are in a sub-population relative to a unit of another. The basis of comparison can be increased, for the sake of easy interpretation, by multiplying the quotient of the two data by an arbitrary value. It will then express the number of elements in a sub-population relative to 10, 100, or 1000 units of another. Coordination ratios are interpreted in coefficient form.

Relationship between distribution and coordination ratios

Distribution and coordination ratios can be computed only on a grouped population. Without the absolute initial data, distribution ratios can be mutually computed from coordination ratios, and conversely, coordination ratios from distribution ratios.

- 30 boys and 50 girls attended a statistics course, attended by a total of 80 students. *Use ratios to analyze the gender distribution of the course and the mutual relative relationship between boys and the.*

  By using DISTRIBUTION ratios, we compare parts of a population to the entire population, in this particular case, the number of boys and girls with the total number of students on the course.

9. *Calculating the distribution of students on a statistics course*

Group of students	Number	
Boys	30	Ratio of boys: $\frac{30}{80} = 0.375 \overset{\cdot 100}{\longrightarrow} \mathbf{37.5}\ \%$
Girls	50	Ratio of girls: $\frac{50}{80} = 0.625 \overset{\cdot 100}{\longrightarrow} \mathbf{62.5}\ \%$
Total	*80*	Together: $0.375 + 0.625 = \mathbf{1} \overset{\cdot 100}{\longrightarrow} \mathbf{100}\ \%$

Based on the distribution ratios of students on the course, boys represent 37.5 % and girls 62.5 %.

We can use COORDINATION ratios to represent the mutual relationship between boys and girls.

$$Number\ of\ girls\ per\ 1\ boy\colon \frac{50}{30} = \mathbf{1.67}\ and\ number\ of\ boys\ per\ 1\ girl\colon \frac{30}{50} = \mathbf{0.6}$$

There are 1.67 girls per one boy in the group, while there are 0.6 boys per one girl. Raising the basis of comparison to 10 for easy interpretation, we get ~17 girls per 10 boys and 6 boys per 10 girls. Raising the basis of comparison to 100, we get 167 girls per 100 boys and 60 boys per 100 girls.

Relationships between distribution ratios and coordination ratios in the example

We know in absence of original data, solely from the coordination ratios, that there are 1.67 girls per 1 boy. From this we can infer that the entire population is composed of 1 boy and 1.67 girls, amounting to a total of 2.67 people. Knowing the simplified number of elements in the entire population, we can compute that

$$the\ proportion\ of\ boys\colon \frac{1}{1+1.67} = \frac{1}{2.67} = 0.375 \xrightarrow{\cdot 100} 37.5\ \%$$

$$the\ proportion\ of\ girsl\ \frac{1.67}{2.67} = 0.625 \xrightarrow{\cdot 100} 62.5\ \%$$

In absence of the original data, solely from the distribution ratios, we know that 37.5 % of the students are boys and 62.5 % are girls. By comparing the two values to each other, in decimal fraction form, we get

$$number\ of\ girls\ per\ 1\ boy\colon \frac{0.625}{0.375} = 1.67$$

$$number\ of\ boys\ per\ 1\ girl\colon \frac{0.375}{0.625} = 0.6$$

3.2.2 Dynamic ratios

Dynamic ratios are the most commonly used analytic tool for temporal comparison, in which such ratios allow us to express the degree, as well as the direction, of change from one period to another in percentage

form. The basis of comparison may be constant or it may vary, which distinguishes between two different ratios.

- **In a temporal comparative row, if comparison is always made to the same period, the ratio is a base ratio.**

$$\text{Base ratio} = \frac{\textbf{Reference}\ \text{period data}}{\textbf{Base}\ \text{period data}}$$

In practice, the first observed value is considered the base period most frequently, but not always. Choice of the base period depends on which period is considered most meaningful as a basis of comparison. Observed data will be compared to the data of that year. It is important to note that you need to be careful with the interpretation of ratios if the basis of comparison is not the period mentioned first.

- **If we are interested in changes from one period to another in successive periods, we use chain ratios, which compare the data of a particular period to the data of the immediately preceding period.**

$$\text{Chain ratio} = \frac{\textbf{Reference}\ \text{period data}}{\text{Data of period immediately}\ \textbf{preceding}\ \text{Reference}}$$

- Suppose we know that the following number of customers have been to a small shop over the past few years, specified as follows: 2000 customers in 2010, 2450 customers in 2011, 2260 customers in 2012, and 2620 customers in 2013. *Use dynamic ratios to describe the changes in the number of customers in this shop over these years.*

- It is a good idea to arrange the data in a table before carrying out the calculations. This will also make it easier to interpret the results. In order that the results are clearly interpretable, the values of the quotients need to be specified to at least four decimal places.

  BASE RATIOS allow us to compare the number of customers in the shop to the data of a selected year. The base period is typically the first year, but any other year may be chosen as the base.

10. *Work table for calculating base ratios, with 2010 as the base period*

Year	Number of customers (people)	Changes in the number of customers compared to 2010 (2010 = 100%)	Degree of change (in %) relative to 2010
2010	2 000	100	—
2011	2 450	$\frac{\text{reference}}{\textbf{base}} = \frac{2011}{\textbf{2010}} \rightarrow \frac{2450}{2000} = 1.225 \overset{\cdot 100}{\longrightarrow} \mathbf{122.5}$	+ 22.5 %
2012	2 260	$\frac{\text{reference}}{\textbf{base}} = \frac{2012}{\textbf{2010}} \rightarrow \frac{2260}{2000} = 1.13 \overset{\cdot 100}{\longrightarrow} \mathbf{113.0}$	+ 13 %
2013	2 620	$\frac{\text{reference}}{\textbf{base}} = \frac{2013}{\textbf{2010}} \rightarrow \frac{2620}{2000} = 1.31 \overset{\cdot 100}{\longrightarrow} \mathbf{131.0}$	+ 31 %

How to fill out the table: The table header must specify the basis of comparison, which is the number of customers in 2010 in this case, considered to be 100 %. The value of the ratio for the period selected as the base is 100 %, and, obviously, there is no change there. Data in succeeding periods are divided by the data of the base period, and multiplying the quotient by 100 yields the form of the ratio. The degree of change compared to the base period is determined by comparing the value of the base ratio to 100.

How to interpret the results: Compared to 2010, the number of customers in the shop increased by 22.5% in 2011, by 13% in 2012, and by 31% in 2013.

How does the picture change if we take the data in 2012 to be the base period?

11. *Work table for calculating base ratios, with 2012 as the base period*

Year	Number of customers (people)	Changes in the number of customers compared to 2012 (2012 = 100%)	Degree of change (in %) relative to 2012
2010	2 000	$\frac{\text{reference}}{\textbf{base}} = \frac{2010}{\textbf{2012}} \rightarrow \frac{2000}{2260} = 0.885 \overset{\cdot 100}{\longrightarrow} \mathbf{88.5}$	– 11.50 %
2011	2 450	$\frac{\text{reference}}{\textbf{base}} = \frac{2011}{\textbf{2012}} \rightarrow \frac{2450}{2260} = 1.0841 \overset{\cdot 100}{\longrightarrow} \mathbf{108.41}$	+ 8.41 %
2012	2 260	100	–
2013	2 620	$\frac{\text{reference}}{\textbf{base}} = \frac{2013}{\textbf{2012}} \rightarrow \frac{2620}{2260} = 1.1593 \overset{\cdot 100}{\longrightarrow} \mathbf{115.93}$	+ 15.93 %

The number of customers in the shop was 11.5% smaller in 2010 than it was in 2012, it was 8.5% higher in 2011 than it was in 2012, and a 15.93% increase may be observed in 2013.

CHAIN RATIOS can be used to represent the changes in the number of customers in the shop compared to the immediately preceding period.

12. *Work table for calculating chain ratios*

Year	Number of customers (people)	Changes in the number of customers compared to the preceding year (Preceding year = 100%)	Degree of change (in %) relative to preceding year
2010	2 000	–	–
2011	2 450	$\frac{\text{reference}}{\textbf{preceding}} = \frac{2011}{\textbf{2010}} \rightarrow \frac{2450}{2000} = 1.225 \overset{\cdot 100}{\longrightarrow} \mathbf{122.5}$	+ 22.5 %
2012	2 260	$\frac{\text{reference}}{\textbf{preceding}} = \frac{2012}{\textbf{2011}} \rightarrow \frac{2260}{2450} = 0.9224 \overset{\cdot 100}{\longrightarrow} \mathbf{92.24}$	– 7.76 %
2013	2 620	$\frac{\text{reference}}{\textbf{preceding}} = \frac{2013}{\textbf{2012}} \rightarrow \frac{2620}{2260} = 1.1593 \overset{\cdot 100}{\longrightarrow} \mathbf{115.93}$	+ 15.93 %

How to fill out the table: The table header must specify the basis of comparison, which is always the data in the preceding period, considered to be 100%. There is no ratio or any change in the first period, as we have no data on any preceding period. Data in any successive period is divided by the data in the immediately preceding period. Multiplying the quotient by 100, we get the form of the ratio. The degree of change relative to the preceding period is determined by comparing the chain ratio to 100.

How to interpret the results: Compared to 2010, the number of customers increased by 22.5% in 2011, it decreased by 7.76% in 2012 as compared to 2011, and there was another increase by 15.93% in 2013, relative to 2012.

3.2.3 Areal/Spatial comparative ratios

An important aspect of areal comparative ratios is that they involve the comparison of statistical data which concern a geographical area. The basis of comparison and the data to compare can be freely replaced with each other, as determined by the direction of the purpose of the comparison.

An areal/spatial ratio compares some statistical data of two geographical units with each other.

$$\text{Areal comparative ratio} = \frac{\text{areal data to compare (A)}}{\text{areal data chosen as basis of comparison (B)}}$$

Areal comparative ratios may be interpreted either in coefficient or in percentage form.

Suppose a company has two sites in two different cities, call them A and B. The site in city A produces 10 250 products and the site in city B produces 8 200 products. *Compare the production of the two sites by calculating the appropriate kind of ratio.*

$$\frac{production\ of\ site\ in\ city\ A}{production\ of\ site\ in\ city\ B} = \frac{10\,250}{8\,200} = \mathbf{1.25}$$

The production of the site in city A exceeds the production of the site in city B 1.25 times, i.e. by 25%.

The basis of comparison and the data to compare are mutually interchangeable, depending on whether we are interested in how much one site's production exceeds the other, or, conversely, how much the production of one site is less than that of the other.

$$\frac{production\ of\ site\ in\ city\ B}{production\ of\ site\ in\ city\ A} = \frac{8\ 200}{10\ 250} = \mathbf{0.8}$$

The production of the site in city B is 80% (4/5) of the production of the site in city A, i.e., it is behind the latter by 20%.

3.2.4 Planned task and plan accomplishment ratios

Ratios calculated from rows which compare planned data and factual data can be used for planning tasks and to quantify plan accomplishment.

- **A planned task ratio compares the planned data to earlier factual data.**

$$\text{Planned task ratio} = \frac{\text{plan}}{\text{fact}}$$

- **A plan accomplishment ratio compares the factual data to planned data.**

$$\text{Plan accomplishment ratio} = \frac{\text{fact}}{\text{plan}}$$

- Suppose a company sold 18 250 products last year. They planned to sell 19 000 products this year. Eventually, they sold 18 880 products. *Evaluate the production of the company by working out the appropriate ratios.*

  We can use the PLANNED TASK RATIO to numerically express how the plan compares to the previous factual accomplishment.

$$\frac{\textbf{Planned}\ \text{sales for this year}}{\textbf{Factual}\ \text{sales last year}} = \frac{19\ 000}{18\ 250} = \mathbf{1.0411} \rightarrow +4.11\%$$

  The company intended to increase its sales by 4.11% compared to last year.

  We can use the PLAN ACCOMPLISHEMNT RATIO to numerically express the extent to which a plan has been accomplished.

$$\frac{\textbf{Factual}\ \text{sales this year}}{\textbf{Planned}\ \text{sales this year}} = \frac{18880}{19000} = \mathbf{0.9937} \rightarrow -0.63\%$$

  The factual sales this year fall behind the plan by 0.63%.

- When we know the factual data of the preceding period and of the reference period, we can compare the two to each other and work out a dynamic ratio.

$$\frac{\textbf{Factual}\ \text{sales this year}}{\textbf{Factual}\ \text{sales last year}} = \frac{18880}{18250} = \mathbf{1.0345} \rightarrow +3.45\ \%$$

The company has been able to increase its sales by 3.45% compared to last year.

3.2.5 Intensity ratios

We often compare different types of data to each other in practice. We can derive intensity ratios from data of different units of measurement or the same unit of measurement. Ratios of the former kind are more popular and more transparent.

- **Intensity ratios are derived from different types of data. They show how the number of elements in a population in relation to a unit in a different population.**

For reasons of easy interpretation, it often makes sense to increase the basis of comparison and determine the value projected onto 10, 100, or 1000 units of the particular population. The quotient of the two data will thus be multiplied by one of these values.

We distinguish between *normal* and *reversed intensity ratios*. The reversed intensity ratio is the mathematical reciprocal of the normal intensity ratio, which is determined by exchanging the data to compare with the basis of comparison.

- Suppose that 1000 instructors teach and 4200 students study at a higher education institution. *Choose the appropriate ratio to characterize the relationship between instructors and students at the institution.*

  We can use INTESITY RATIOS to express the number of students per instructor, and conversely, the number of instructors per student.

  $$Number\ of\ students\ per\ instructor: \frac{total\ of\ students}{total\ of\ instructors} = \frac{4200}{1000} = \mathbf{4.2}$$

  $$Number\ of\ instructors\ per\ student: \frac{total\ of\ instructors}{total\ of\ students} = \frac{1000}{4200} = \mathbf{0.2381}$$

  In this institution, there are 4.2 students per instructor, while there are 0.2381 instructors per student. By increasing the basis of comparison, we can say that there are 420 students per 100 instructors and there are ~ 24 instructors per 100 students. If we on-

ly knew the intensity ratios, but not the absolute data, we could say there are over four times as many students studying at the institution as instructors.

We distinguish between *raw* and *purified intensity ratios* in terms of how they relate to the basis of comparison. A purified intensity ratio relates to a subpart of the basis of comparison, which is closer to the data to compare. When computing a purified intensity ratio, the same data to compare is compared to a subpart of the basis of comparison, therefore, its value will certainly be greater than that of the raw intensity ratio. It should be clear that, mathematically, the division of a number with a smaller number yields a greater quotient than the division of the same number with a greater divisor.

$$\text{Raw intensity ratio} = \mathbf{\frac{A}{B}} \qquad \text{Purified intensity ratio} = \mathbf{\frac{A}{b}}$$

$$\text{where } \mathbf{\frac{b}{B}} = \text{the distribution ratio computed relative to the basis of comparison}$$

From this, the following relationship follows between a raw and a purified intensity ratio and the distribution ratio:

$$\text{Raw intensity ratio} = \text{Purified intensity ratio} \cdot \text{Distribution ratio}$$

Mathematically, the relationship is transparent: $\frac{A}{B} = \frac{A}{b} \cdot \frac{b}{B}$

We can compute the third ratio if we know any two terms of the multiplication.

- Suppose a total of 4000 students study and 1000 instructors teach at a higher education institution, 400 out of the latter teach statistics.
  Use the appropriate ratios to characterize the relationship between the total of instructors, the statistics instructors, and the students at this institution.

  $Number\ of\ students\ per\ instructor$: $\frac{total\ students}{total\ instructors} = \frac{4000}{1000} = \mathbf{4}$

  $Number\ of\ students\ per\ statistics\ instructor$:

  $$\frac{total\ students}{total\ statistics\ instructors} = \frac{4000}{400} = \mathbf{10}$$

There are 4 students per instructor at the institution. However, if we restrict the basis of comparison to statistics instructors, then we get 10 students per statistics instructor.

This example transparently illustrates the relationship between the ratios.

$$Rate\ of\ statistics\ instructors\ within\ total\ instructors\ (distribution\ ratio)$$

$$\frac{statistics\ instructors}{total\ instructors} = \frac{400}{1000} = 0{,}4$$

$$V_{i\ RAW} = V_{i\ PURIFIED} \cdot V_M \ \rightarrow\ \mathbf{4 = 10 \cdot 0.4}$$

3.3 SUMMARY AND QUESTIONS

3.3.1 Summary

Ratios are simple tools of statistical analysis. We can compare either data of the same type or data of different types to each other. Therefore, ratios can be derived from classifying, comparative, or descriptive rows. We illustrated the application of various types in practical examples. We can often encounter dynamic ratios, which represent temporal comparison, distribution and coordination ratios, which characterize the internal structure of a population or a sample, areal ratios, which are used for areal comparisons, and intensity ratios, which involve the comparison of different types of data, in everyday life. The self-test questions and problems help students assess the depth of their understanding of the basics of this simple analytic method.

3.3.2 Self-test questions

What is a ratio?

What are the similarities and differences between the properties of distribution and coordination ratios?

What type of ratio can be used for temporal comparison?

What is the difference between base and chain ratios?

When can you employ areal comparative ratios?

How do you define a plan task ratio?

How do you define a plan accomplishment ratio?

What kind of ratio can be used to compare different types of data?

What is the difference between a normal and a reversed intensity ratio?

What relationship holds between the raw and purified intensity ratio and the distribution ratio in terms of how they relate to the basis of comparison?

3.3.3 Practice tests

❁ Determine the types of ratios.

	Distribution	Coordination	Dynamic	Areal comparative	Plan task	Plan accomplishment	Intensity
EURO's rate of exchange 1.35 EUR/USD							X
A company's revenues increased by 30% compared to last year			X				
30% of a store's sales revenues come from selling trousers	X						
Per capita average annual chocolate consumption in 2010 was 28 kg							X
Thanks to exceptional achievements last year, the company expects a 10% increase in revenues next year					X		
Per capita GDP in the USA is nearly five times that of Hungary				X			
According to visitors' data of a museum there were two female visitors per one male visitor		X					
A company suffered a 25% set-back compared to the plan, due to the crisis						X	
Amount of imported goods in 2013 increased by 20% compared to 2010			X				
There is a 50% discount on some products at the beginning of the year	X						

4. PRACTICAL USE OF RATIOS: LABOR-MARKET AND FINANCIAL STATISTICS

4.1 GOALS AND COMPETENCIES

Ratios are often used in the study of social and economic phenomena. The most frequently used ratios are dynamic, distribution, and intensity ratios, which appear in demographic and labor-market statistics, and financial calculations.

The purpose of this unit is to familiarize students with the practical use of ratios, especially the methodology of labor-market analysis and indices in financial areas, and various forms in which intensity ratios appear. Students will learn to identify the major labor-market ratios and know how to calculate them. They will learn about financial indices, which can be defined as ratios and are most relevant for the assessment of a company. A general goal is for students to be able to identify ratios in social and economic contexts.

4.2 TOPICS

Ratios appear in all areas of our lives. We use primarily distribution ratios in labor-market statistics and dynamic ratios to quantify temporal changes in particular categories of the labor market. We most frequently compare different types of data in areas of finance and in calculating demographic ratios.

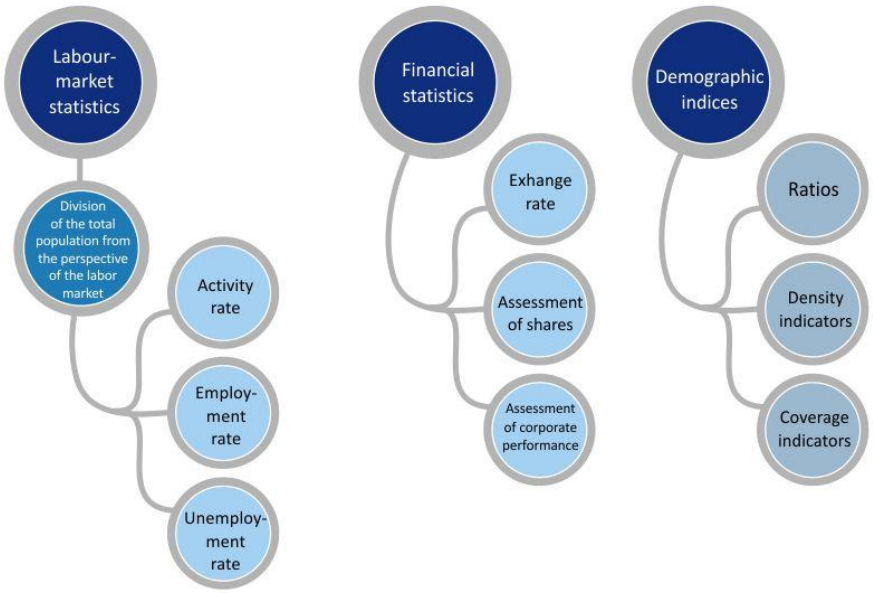


Figure 5 Major areas of the use of ratios

4.2.1 Labor-market statistics

From a macroeconomic perspective,[3] a country's total population is divided into working-age and nonworking-age groups. The working-age population includes people who are able and willing to work, determined primarily by their age and health.

- Age-based accounts of the working-age population vary across countries. National, international and European statistics are not uniform. According to OECD statistics and statistics in Europe, people in the age-group between 15 and 64 belong here. In statistics published by the ILO, however, as well as statistics in Hungary, people between 15 and 74 years of age are classified in this category. Some statistics specify 16 years of age as the lower limit, while some others set it as low as 14. Therefore, we must carefully consider the content of particular categories when interpreting labor-market indices that are found in different databases.

The working-age population can be further subdivided into two subcategories, economically active and economically inactive groups. Economically inactive people do not participate in doing socially organized

[3] You can learn more about labor-market relationships in the Economy II course.

work. For example, housewives, and also lottery millionaires, belong here. The group of economically active people is composed of people who have a job and the unemployed. The unemployed population can be further subdivided into out-of-necessity and voluntary subgroups.

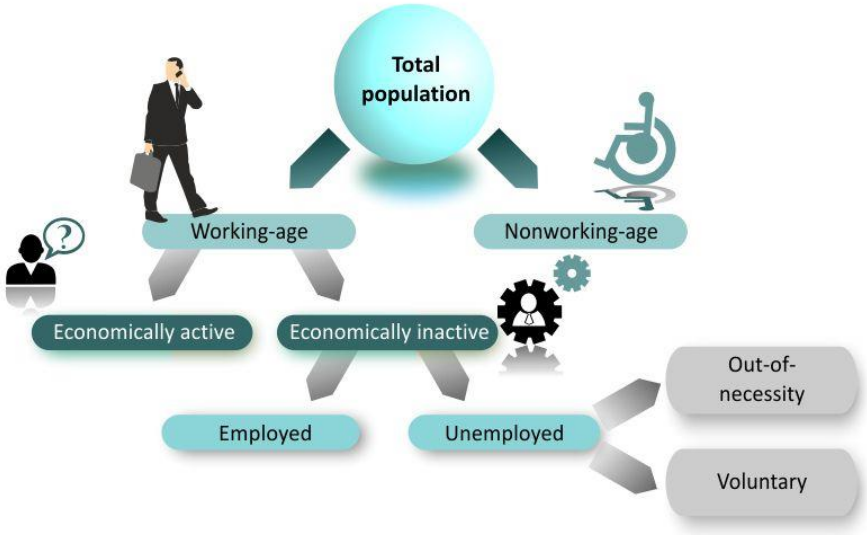


Figure 6 Division of the total population from the perspective of the labor market

The most important labor-market indices, such as the rate of activity, the rate of employment, and the rate of unemployment, are calculated on the basis of this division.

Rate of activity

The rate of activity represents the rate of economically active people within the working-age population.

$$\text{Activity rate} = \frac{\text{Economically active}}{\text{Working} - \text{age population}}$$

The activity rate in the United States and in Japan exceeds 70%, while the European average barely approximates this value. In a number of countries, the rate hardly exceeds 60%. Low activity rate may be a symptom of serious labor-market problems, as the economically active population represents the supply side of the labor market. It is not just the current value of the activity rate that may be an important labor-market

indicator, but also its change in time, as this signals the direction of changes in the labor market.

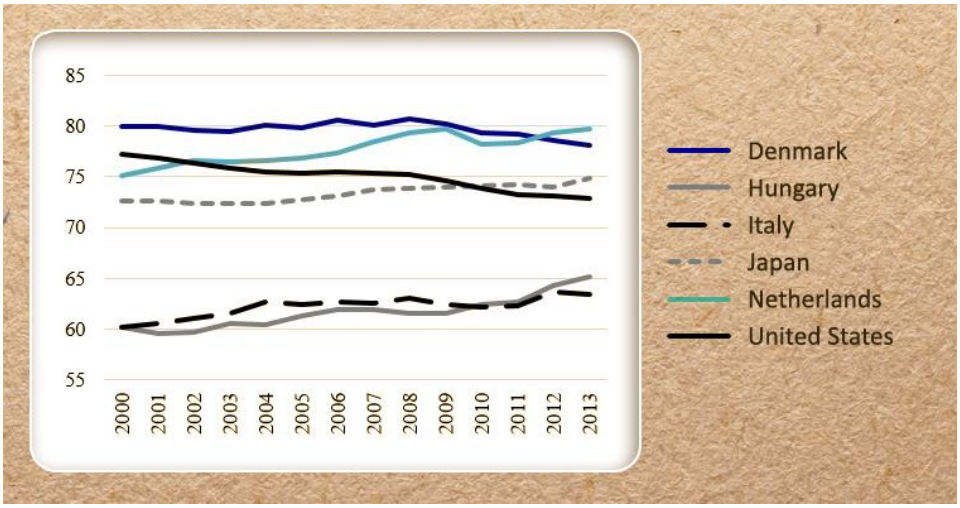


Figure 7 Changes in the activity rate in some countries between 2000 and 2013

Source: *OECD* (2014)

As the diagram shows, the rate of activity, as well as the direction of its change, varied across these countries during the reference period. The rate rose in some, while it fell in some other countries. The rate of activity in the United States declines steadily, while it rises in Japan and the Netherlands. The activity rate does give us a picture of the state of the labor market, but the observed variations suggest that they may be in part due to variations „in the background" in the labor-market categories that appear in the numerator and the denominator. In the practical use of ratios, you need to bear in mind that variations in a kind of ratio, which is derived as the quotient of two items of data, may be due either to variation in the data to compare or in the basis of comparison.

Rates of employment and unemployment

The employment rate represents the rate of employed within the economically active population.

$$\text{Employmnet rate} = \frac{\text{Employed}}{\text{Economically active}}$$

The unemployment rate represents the rate of registered unemployed within the economically active population.

$$\text{Unemployment rate} = \frac{\text{Registered unemployed}}{\text{Economically active}}$$

The two indices may be defined as distribution ratios, where one subpopulation is represented by the employed and the other by the unemployed, such that these two groups jointly constitute the economically active population, in accordance with the partition of the labor market. As far as employed people are concerned, we can only consider those who have a registered job. Similarly, only registered unemployed can be considered. Also, in practice, the equation that involves the distribution ratios does not in fact yield 100%, because there are data that never get reported. For example, people who do not report themselves as unemployed in employment offices cannot be considered, nor can employed people who are illegally employed. Despite such problems of recording data, the relationship between the two rates which derives from a shared basis of comparison is clear in the case of changes over extended periods of time: if the employment rate rises, the unemployment rate falls.

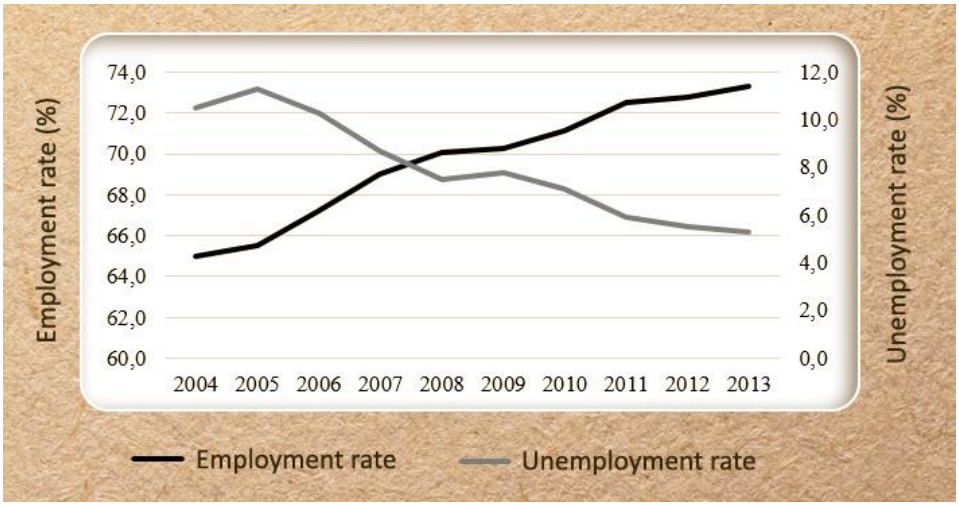


Figure 8 Changes in the employment and unemployment rates in Germany between 2004 and 2013

Source: *EUROSTAT* (2014)

4.2.2 Financial statistics

Ratios are commonly used in finance,[4] as a range of indices can be derived as quotients of two items of data. In this section, we will focus on the statistical characteristics of financial indices that can be derived as ratios. We will not be concerned with a detailed description of specific features of finance or with questions of the practical applicability of indices.

The most important index that can be defined as a ratio in international finance is the **exchange rate**, which compares the values of two different instruments of payment. Official exchange rates are published by international banks. Changes in exchange rates have a considerable effect on the volume of international trade. You can learn about the exchange rate of the Euro compared to other currencies on the website of the European Central Bank:

9. Web site of the European Central Bank: http://www.ecb.europa.eu/stats/exchange/eurofxref/html/index.en.html

Most ratios, however, are used in the assessment of companies. A range of indices, derived as quotients of two related items of data, may help you take investment decisions or inform about shares.

Use of ratios in assessing shares

A number of different indices, defined as ratios, can be mentioned in connection with shares. One of the most important ones is **EPS** (Earnings per Share), which compares the earnings after tax to the number of shares.

$$\mathrm{EPS} = \frac{\text{ernings after tax}}{\text{number of shares}}$$

The **P/E** (Price/Earnings) index, another kind of ratio, frequently occurs in the context of investment decision making and shares. This index compares the price of shares to earnings per share. The higher the value of the index is the better, from the perspective of investments.

$$\mathrm{P/E} = \frac{\text{share price}}{\text{earnings per share}} = \frac{\mathrm{P_0}}{\mathrm{EPS_1}}$$

You can read more about the index on the following site:

10. P/E: http://www.investopedia.com/terms/p/price-earningsratio.asp

[4] You can read more about financial statistics in: Copeland, T. – Koller, T. - Murrin, J. (2000): *Valuation. Measuring and managing the value of companies*. 3rd edition. McKinsey & Company Inc.

Use of financial indices in assessing corporate performance[5]

We often use financial indices, derived as ratios in the statistical sense, in assessing the performance of companies. These indices are easy to compute and interpret, although they have their own limits in finance.

Indices that are used to assess the performance of a company include ROE, ROA, ROS, and ROI. **ROE** (Return on Equity), which measures the profitability of equity, may be defined by comparing the earnings after tax to the equity.

$$\text{ROE} = \frac{\text{Earnings after tax}}{\text{Equity}}$$

You can read more about this index on the technical portal below:

11. ROE: http://www.investopedia.com/terms/r/returnonequity.asp

ROA (Return on Asset) is the index of returns in relation to assets. Essentially, it gives you a picture about the profitability of a company's assets. The index compares the earnings after tax to the total assets.

$$\text{ROA} = \frac{\text{Earnings after tax}}{\text{Total assets}}$$

You can read more about this index on the technical portal below:

12. ROA: http://www.investopedia.com/terms/r/returnonassets.asp

ROS (Return on Sales), the index of the operational efficiency relative to revenues, compares the earnings after tax to revenues.

$$\text{ROS} = \frac{\text{Ernings after tax}}{\text{operation related net revenue and other income}}$$

You can read more about this index on the technical portal below:

13. ROS: http://www.investopedia.com/terms/r/ros.asp

ROI (Return on Investment) is an index that is used for the assessment of investments. It expresses the returns on investment capital. Other ratios include turnover rate indices, as indicators of efficiency, and indices of capital structure, indebtedness, and liquidity.

You can read more about financial indicators on the technical portal below:

[5] This chapter draws on Hollóné Kacsó Erzsébet (2011): *Vállalatértékelés*. Mutatók, mod ellek, üzlet- és vagyonértékelés.

14. Financial indicators:
    http://kfknowledgebank.kaplan.co.uk/KFKB/Wiki%20Pages/Financial%20Performance%20Indicators%20%28FPIs%29.aspx

4.2.3 Demographic indices

Most of the indices that relate to a country's population are intensity ratios. The best known **density index** is the indicator of population density, which is the number of people per unit of area, most frequently quoted per square kilometer, people/km^2. As its unit of measurement shows, it is a quotient of different types of data.

$$\text{Population density} = \frac{\text{Country's total population}}{\text{Country's area területe}} \rightarrow \frac{\text{people}}{\text{km}^2}$$

Supply and services **coverage indicators**, such as healthcare or cultural services coverage, are also intensity ratios. The latter compares the number of cultural institutions or events to the number of residents.

$$\text{Health care coverage} = \frac{\text{total doctors}}{\text{population}} \rightarrow \frac{\text{doctors}}{\text{people}}$$

The feature that these indices share is that the basis of comparison in each is the human population or a specific part of it. The data to compare and the basis of comparison are often switched around in these indices, offering examples of normal, as well as reversed intensity ratios. A coverage index may be determined for a particular unit of area, such as the number of shops per village or town.

Ratios (in the narrow mathematical sense) play an important role in demographic descriptions, when, for example, they relate the number of births or deaths to the population to derive an index of natural reproduction or natural decline. Raw and purified intensity ratios occur frequently in these contexts in practice. For instance, the fertility rate is an example of a purified intensity ratio, which represents the number of live births per women of maternal age.
You can find a range of different demographic indices in the database of the World Bank:

15. World Bank demographic indices:
    http://econ.worldbank.org/WBSITE/EXTERNAL/DATASTATISTICS/0,,contentMDK:20451597~pagePK:64133150~piPK:64133175~theSitePK:239419,00.html

Indicators of the national economy also include indices that can be defined as intensity ratios. The most common example is the per capita

GDP, or the GDP per worker, is derived by narrowing down the basis of comparison, and represents work productivity.

4.3 SUMMARY AND QUESTIONS

4.3.1 Summary

Ratios are a commonly used means of studying economic and social phenomena. Labor-market statistics mostly employ distribution and dynamic ratios, while in the context of finance and demography, intensity ratios are more common.

Major labor-market indicators, such as the activity ratio and the rates of employment and unemployment, can be calculated on the basis of partitioning the total population. The most common types of ratios in the context of finance include the exchange rate, P/E, which is relevant for the assessment of shares, ROE, ROA, and ROS, which play an important role in making investment decisions. Intensity ratios used in demography include population density, coverage indices, and birth and death rates, among many others. GDP, either per capita, or per worker, is also a quotient of connected data, an intensity ratio.

4.3.2 Self-test questions

Which are the major areas of the use of ratios?

Which are the most commonly used types of ratios?

Which ratios are used in the description of the labor market?

How do you calculate the activity rate?

How do you calculate the employment rate?

How do you calculate the unemployment rate?

What is common in the calculation of the employment rate and the unemployment rate?

Give some examples of ratios used in finance.

What ratios can be used to characterize corporate performance?

Give some examples of ratios used in demography.

4.3.3 Practice tests

☸ Decide whether the statements below are true (T) or false (F).

The total human population can be divided into economically active and inactive groups from the perspective of the labor market.

TRUE **FALSE**

The employment rate represents the ratio of employed within the total population.

TRUE **FALSE**

A person is economically inactive if they do not take part in doing socially organized work.

TRUE FALSE

When calculating the unemployment rate, we can only consider registered unemployed.

TRUE FALSE

The activity rate represents the ratio of economically active within the working-age population.

TRUE FALSE

When calculating population density, we compare the population of a country to the area of the country.

TRUE FALSE

P/E is an indicator which compares the earnings after tax to equity.

TRUE **FALSE**

In determining the healthcare coverage index, we compare the number of residents to the number of doctors.

TRUE **FALSE**

The ROA index compares the earnings after tax to the total assets.

TRUE FALSE

The fertility rate is a raw intensity ratio, which represents the number of live births per women of maternal age.

TRUE **FALSE**

5. DESCRIPTIVE STATISTICS

5.1 GOALS AND COMPETENCIES

The apparatus of descriptive statistics includes basic distribution descriptors that can be used to characterize a population or a sample according to a quantitative attribute. Indices can be divided into three main groups. Means carry information about the characteristic values of a data-set, measures of dispersion represent differences between data, and shape indices carry information about the shape of a histogram.

The purpose of this unit is to familiarize students with the apparatus of descriptive statistics. Once students have acquired the knowledge of the properties of indicators and ways to calculate and interpret them, they will be able to statistically characterize data sets according to some quantitative attribute. It is particularly important that students acquire a general understanding of the system of indices, including specific advantages and disadvantages of particular indicators, in order that they can draw appropriate inferences about a particular phenomenon of interest.

5.2 TOPICS

Descriptive statistical tools are used for a concise characterization of a population or a sample according to a quantitative attribute. Basic distribution descriptors are among the simple tools of statistics.

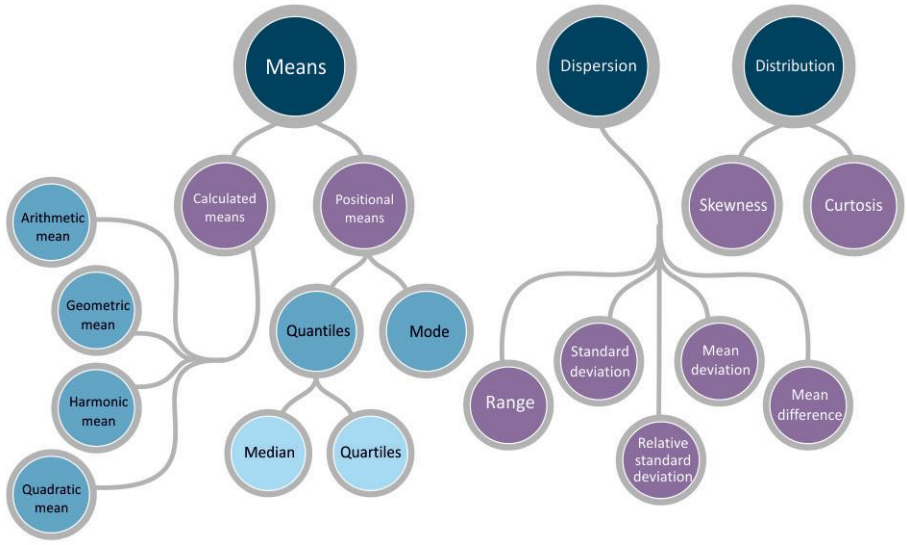


Figure 9 Descriptive statistic tools

In calculating basic distribution descriptors, a distinction has to be made between individual values and values calculated from class interval frequency arrays. With the exception of averages, index calculations will be discussed with reference to individual values.

5.2.1 Means

A common feature of means as values characteristic of a population or a sample is that they ALWAYS fall between the smallest and the highest values of a data set. They are calculated by simple algebraic means, their interpretation is unequivocal, though you need to pay attention to their properties, which affect the generalizability of the inferences you can draw from them.

Calculated means: averages

The concept of an average is commonly used in everyday life, for example, in talking about average salaries, a student's grade point average, average prices, or in expressions like "above average" or "below average." Such expressions normally carry the concept of a simple arithmetic mean. However, we distinguish between several different kinds of means in statistics, which are used in different ways to express different meanings. Depending on the nature of the data, we distinguish between arithmetic, geometric, harmonic, and quadratic means.

Arithmetic mean

The arithmetic mean is the simplest and most commonly used type of mean. The way it is calculated may depend on whether we have individual values or grouped values, i.e., class interval frequency arrays.

- **For individual values, the arithmetic mean is calculated by first summing the values and then dividing the sum by the number of values.**

A student got the following grades in a semester: 2, 3, 5, 5, 4

$$Her\ Gradepoint\ average: \bar{x} = \frac{2+3+5+5+4}{5} = \mathbf{3.8}$$

Increasing the number of elements makes the operation of calculating the arithmetic mean increasingly long. Therefore, it makes sense to group the data. If there are few attribute values in the data set and they occur repeatedly, it is useful to construct a frequency array, which allows us to compress the data by specifying the number elements that correspond to an attribute value, i.e., their frequency. Thus, we calculate a (frequency) weighted arithmetic mean.

- **In the case of a frequency array, we calculate the weighted arithmetic mean by multiplying the values to be averaged by the value of frequency of their class, and divide the result by the number of elements, which in this case equals the sum of frequencies.**

$$\bar{x} = \frac{f_1 \cdot x_1 + f_2 \cdot x_2 + \cdots + f_n \cdot x_n}{f_1 + f_2 + \cdots + f_n} = \frac{\sum f \cdot x}{\sum f}$$

- The $\sum$ symbol means ‘sum’, and refers to the entire expression that follows the symbol. So, in the case of a weighted average, the class-specific frequencies are multiplied in pairs by their own values, and then the values of the products are summed.

A student got the following grades in a semester:
2, 3, 5, 5, 4, 4, 3, 5, 2, 5, 5, 4, 4, 5, 3, 4, 5, 5, 4, 4

Determine her grade point average.

For calculating the arithmetic mean, it is useful to group our data by considering how many grades the student got of each type of grade.

13. *Work table for constructing and interpreting a frequency array*

Grade	Number of student's grades (pcs)	This frequency array means that the student got
5	8	← 8 pcs of grade 5
4	7	← 7 pcs of grade 4
3	3	← 3 pcs of grade 3
2	2	← 2 pcs of grade 2
1	0	← 0 pcs of grade 1
Total	*20*	← total 20 pcs of grades in the semester

The grade point average is a mean of grades, not a mean of frequencies.

$$\bar{x} = \frac{f_1 \cdot x_1 + f_2 \cdot x_2 + \cdots f_n \cdot x_n}{f_1 + f_2 + \cdots f_n} \rightarrow \frac{8 \cdot 5 + 7 \cdot 4 + 3 \cdot 3 + 2 \cdot 2 + 0 \cdot 1}{8 + 7 + 3 + 2 + 0} = \mathbf{4.05}$$

With an increase in the number of data, calculating the arithmetic mean becomes increasingly complex. If you have a large number of data, it is useful to arrange them in a class interval frequency array. Classes must be defined in such a way that each element can be unambiguously assigned to a class. It is important to bear in mind that the boundaries of real class intervals are marked by the upper bounds, while lower bounds are introduced for the purpose of making distinctions, so they are treated as fictitious boundaries.[6] Values higher than the upper limit fall into the next interval. Therefore, when setting the values of the lower limits, you need to consider the magnitude of the data. All calculations will be based on considering the upper bounds of successive class intervals.

- **In the case of a class interval frequency array, we calculate the weighted arithmetic mean by taking the data arranged in class intervals and multiplying the class mean calculated as a simple arithmetic mean of the upper limit of a class interval and the upper limit of the immediately preceding class interval by the frequency value of the class, and divide the product by the sum of frequencies.**

[6] When forming class intervals, we eventually designate *separating values*, which will be t he upper bounds. Thus, for example, in considering examination paper scores, the valu es 20, 40, 60, and 80 will partition the data set. For reasons of unambiguous classificati on, we specify the lower boundaries of the intervals, too, by taking the magnitude of the data into consideration. In working with examination scores, we only use integers. Ther efore, if the score of a test is either 20 or 21, then 21 belongs in the next class interval, as we set the number which immediately follows 20 as the lower limit.

$$\bar{x} = \frac{f_1 \cdot x_1 + f_2 \cdot x_2 + \cdots f_n \cdot x_n}{f_1 + f_2 + \cdots f_n} = \frac{\sum f \cdot x}{\sum f}$$

❁ We know the test scores of a statistics class arranged in intervals. *Determine the average of the tests.*

14. *Work table for interpreting data in a class interval frequency array*

Score	Number of tests (pcs)		This class interval frequency array means that
0 – 20	8	←	8 students scored between 0 and 20 points
21 – 40	12	←	12 students scored between 21 and 40 points
41 – 60	18	←	18 students scored between 41 and 60 points
61 – 80	12	←	12 students scored between 61 and 80 points
81 – 100	10	←	10 students scored between 81 and 100 points
Total	*60*	←	a total of 60 students completed the test

For the calculation of the weighted arithmetic mean, we need to determine the class mean, which is calculated as a simple arithmetic mean of the upper bound of a class interval and the upper bound of the immediately preceding class interval.

15. *Work table for calculating class means*

Score	CLASS MEAN	Number of tests (pcs)
0 - 20	$\frac{0+20}{2} = \mathbf{10}$	8
21 - 40	$\frac{20+40}{2} = \mathbf{30}$	12
41 – 60	$\frac{40+60}{2} = \mathbf{50}$	18
61 – 80	$\frac{60+80}{2} = \mathbf{70}$	12
81 - 100	$\frac{80+100}{2} = \mathbf{90}$	10
Total		*60*

$$\bar{x} = \frac{f_1 \cdot x_1 + f_2 \cdot x_2 + \cdots f_n \cdot x_n}{f_1 + f_2 + \cdots f_n} \rightarrow \frac{8 \cdot 10 + 12 \cdot 30 + 18 \cdot 50 + 12 \cdot 70 + 10 \cdot 90}{8 + 12 + 18 + 12 + 10} = 51{,}33$$

The average score of the 60 statistics students is 51.33 on the test.

Properties of the arithmetic mean

It is easy to calculate the arithmetic mean either from individual values, or from frequency arrays, and it can be employed whenever the data can be summed. There is only one mean for a particular array of data. This makes the analysis of values easy. Different arrays of data can be compared on the basis of their arithmetic means. For example, assuming people viewing two different films at a cinema, and assuming that their opinions about the films are measured on a 1 to 5 scale, the means of their opinions will represent their preference as between the two films. Similarly, the means of test scores in two different groups of students will give their instructor a picture of which group performed better on the test. Because you need a complete set of data for the calculation of an arithmetic mean, you can draw inferences in regard to the entire data set. The greatest disadvantage of the arithmetic mean, however, is that it is highly sensitive to outstandingly low or outstandingly high values, which may affect its usefulness by introducing a certain amount of uncertainty into the inferences drawn from it.

Outstandingly low or outstandingly high values in a data set are called outliers.

The task of handling outliers raises a number of questions. From a statistical perspective, it makes sense to exclude them from the data set, but this may be easily criticized from a practical aspect. We can get a complete picture about a data set only if we consider all the data available. A typical example of the distortion of arithmetic means is the calculation of average salaries, which everyone finds a bit strange, since your actual salaries are always lower than that, and so are everyone else's you know. This distortion is due to some, though not many, exceptionally high salaries in the data set. However, if we were to exclude them from the calculation, it would raise the question of how the calculated value could be interpreted as a generalization. Another typical example is European Union GDP data from the area of macroeconomic indicators. Income data from Luxemburg are exceptionally high within the European community, which distorts any GDP calculations, including calculations of the average income in the Union. One wonders if one can get a true picture of the income situation in the European Union, if Luxemburg is excluded from the calculation. For these reasons, there is no universally accepted statistical method concerning the treatment of outliers. The question of whether to exclude or include them in an analysis is decided on a case by case basis.

Geometric mean

A geometric mean can be calculated when the values to be averaged cannot be summed, but their multiplication is interpretable.

We derive the geometric mean from individual values by multiplying the elements together and extracting the n^{th} root of the product, where n is the number of elements in the data set.

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} \rightarrow \bar{x}_g = \sqrt[n]{\pi\, x_n}$$

If an element occurs more than once, we use the weighted geometric mean form:

$$\bar{x}_g = \sqrt[n]{x_1^{f_1} \cdot x_2^{f_2} \cdot \ldots \cdot x_n^{f_n}} \rightarrow \bar{x}_g = \sqrt[n]{\pi\, x_n^{f_n}}$$

The sign π means 'multiply', and it refers to the entire expression that follows it. That is to say that we multiply the values together. If an element occurs more than once, then first we apply exponentiation to the value of such an element, and then multiply the exponential values together.

Suppose a hotel's turnover of visitors increased by 1% from 2010 to 2011, by 2% from 2011 to 2012, and by 5% from 2012 to 2013. *Determine the average change in the turnover of visitors between 2010 and 2013*

$$\bar{x}_g = \sqrt[3]{1.01 \cdot 1.02 \cdot 1.05} = \mathbf{1.0265}$$

On the basis of the geometric mean, it turns out that the average increase in the hotel's turnover of visitors during the reference period was 2.65%.

Given the changes in the hotel's turnover of visitors over the past few years, determine the average degree of change.

16. *Average change in the hotel's turnover of visitors*

Year	Change relative to the preceding year
2008	-
2009	1.05
2010	1.10
2011	0.95
2012	1.05
2013	1.10

$$\bar{x}_g = \sqrt[5]{1.05^2 \cdot 1.1^2 \cdot 0.95} = 1.0485$$

The hotel's turnover of visitors between 2008 and 2013 increased on average by 4.85%.

The geometric mean is typically used in averaging dynamic ratios. Because of the properties of the ratios, we compute the mean not from de-

grees of change in percentage in this case, but from decimal fraction forms.

Harmonic mean

We use the harmonic mean when neither the sum, nor the product of values is interpretable, but the sum of their reciprocals is. The harmonic mean is generally used with intensity ratios.

We compute the harmonic mean from individual values by dividing the number of elements by the sum of the reciprocals of the values.

$$\bar{x}_h = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum \frac{1}{x_n}}$$

If an element occurs more than once, then we use the weighted harmonic mean form.

$$\bar{x}_h = \frac{n}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \cdots + \frac{f_n}{x_n}} = \frac{n}{\sum \frac{f_n}{x_n}}$$

Three printers are used in an office. One prints color pages at a speed of 12 pages/minute, another at 6 pages/minute, and the third at 4 pages/minute. *Determine the average performance of the printers, i.e., the number of color pages printed on average in the office.*

In this case, we cannot calculate either the arithmetic, or the geometric mean, as the performance of each printer is different. Therefore, what we need to consider first is the amount of time it takes for them to print one page.

$$\frac{\text{pages}}{\text{minute}} \rightarrow \frac{\text{pages}}{\text{minute}} \text{ i. e., the total performance is: } \frac{1}{12} + \frac{1}{6} + \frac{1}{4} = \frac{6}{12} = \frac{1}{2}$$

It turns out from the calculation that one of the printers would print 1/12 of a page, the other 1/6 of a page, and the third 1/4 of a page in a minute. Since we have three printers, and the question is not how long it takes to print a page, but how many pages they print in a minute, we use the concept of the harmonic mean thus:

$$\bar{x}_h = \frac{3}{\frac{1}{12} + \frac{1}{6} + \frac{1}{4}} = \frac{3}{\frac{1}{2}} = 6$$

Using the harmonic mean in the calculation, we have learned that the three printers, each different in performance, taken together print 6 color pages a minute.

Quadratic mean

The quadratic mean is rarely used in practice. It is mostly employed with other types of indicators, primarily to determine the measures of statistical dispersion. The use of the quadratic mean is justified by the use of values different in sign.

- **We calculate the quadratic mean from individual values thus: first we square the elements, then sum the results, then divide the sum by the number of elements, and finally extract the square root of the result.**

$$\bar{x}_q = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}} = \sqrt{\frac{\sum x_n^2}{n}}$$

weighted mean form: $\bar{x}_q = \sqrt{\frac{f_1 \cdot x_1^2 + f_2 \cdot x_2^2 + \cdots + f_n \cdot x_n^2}{f_1 + f_2 + \cdots f_n}} = \sqrt{\frac{\sum f_n \cdot x_n^2}{n}}$

Positional means

Positional means are such characteristic values in a data set which characterize the data set in virtue of their position.

Mode

Mode is an appropriate way to characterize a population when one or more elements in the data set occur more than once. Its advantage derives from its applicability to qualitative and areal attributes, in addition to attributes of quantity.

- **Mode is the most frequently occurring element in a data set, the typical value. In the case of individual values, the mode is determined observationally.**

We cannot calculate a mode from any data set, because there is no mode if every element is different. Several different elements may recur several times in a data set. In such a case, as no single mode can be determined, we regard each recurring element as a mode, and such a data set is called multimodal or polymodal.

- Suppose a student got the following grades in a semester:
  2, 3, 5, 5, 4, 4, 3, 5, 2, 5, 5, 4, 4, 5, 3, 4, 5, 5, 4, 4

  Determine the mode of the data set.

  The grade that occurs most frequently is “excellent (5).”

Quantiles

Quantiles are special data values that mark the boundaries in an ordered dataset. The best known quantile is the *median*, which is the midpoint, trisection points are *tertiles* (or *terciles*), and quartering points are called *quartiles*. All quantiles are calculated uniformly. We will discuss two of them in more detail, the median and the quartile, where we detail the method of calculation as well.

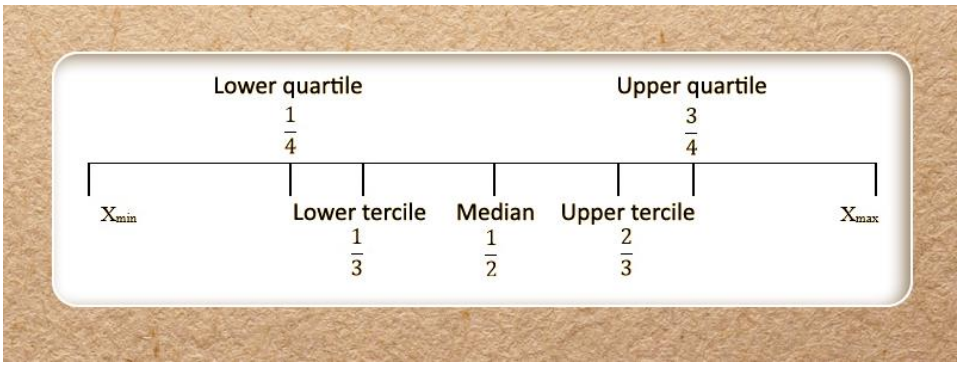


Figure 10 Position of special quantiles in a data set

Median

Median is the element in the middle of an array of data, the midpoint. It is easy to determine the median, as it has a clear position, there is only one. Its advantage is that it can be used both for quantitative and for qualitative attributes which can be measured at least on an ordinal scale. It is sensitive to outstanding values, but still it characterizes a data set better than a mean. The median is interpreted as the value, relative to which, half of the elements of the data set are greater, and the other half smaller.

A condition for calculating the median is that the data be arranged in ascending order. For individual values, the value of the median is the $\frac{n+1}{2}$th element of the ordered data set.

For data sets of individual values, it is important to know whether the number of elements is an even or an odd number. For a data set containing and odd number of elements, the median is literally the element in the middle of the data set, while for an even number of elements it falls in between the two elements in the middle, as their simple arithmetic mean.

- Suppose a student got the following grades in a semester:
  2, 3, 5, 5, 4, 4, 3, 5, 2, 5, 5, 4, 4, 5, 3, 4, 5, 5, 4, 4

  Determine the median of this data set.

  First, we need to arrange the data in ascending sequence:
  2, 2, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5

  We look for the median in the $(n+1)/2^{th}$ position, which in the case of this data set, composed of 20 elements, is (20+1)/2 = 10.5. As the data set is composed of an even number of elements, the median falls in between two elements, which in this particular example are the 10^{th} and the 11^{th} elements, that is 4 and 4, the arithmetic mean of which is also 4 → the value of the median is 4. *Half of the grades the student got in the semester are lower, half are higher than 4.*

Quartiles

Quartiles are quartering points of the data set. Essentially, in between the quartiles, there is a median in position 2/4, i.e., 1/2. The lower quartile is located at ¼ of the data set, and the upper quartile at ¾.

- **The lower quartile is the value *x*, for which, ¼ of the elements of the data set are lower than *x*, and ¾ are higher than *x*. The upper quartile is the value *y*, for which ¾ of the elements of the data set are lower than *y*, and ¼ are higher than *y*.**

A condition for the calculation of quartiles, too, is that the data be arranged in ascending order. For individual values, the lower quartile is the $\frac{n+1}{4}$-th element of the ordered data set, and $\frac{3(n+1)}{4}$ –th element is the upper quartile.

- *Using the example above, determine the lower and upper quartiles in the data set.*

  Data: 2, 2, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5

  We look for the lower quartile in position (n+1)/4, which in the case of this data set, composed of 20 elements, is (20+1)/4 = 5.25. In such a data set composed of an even number of elements, the

lower quartile also falls in between two values, which, in this case, are the 5th and the 6th elements, that is 3 and 4. The value of the lower quartile is calculated by adding ¼ of the difference between the two values to the lower value, i.e., 3 + (4-3)/4→ the value of the lower quartile is **3.25.**

¼ of the grades the student got in the semester are worse than 3.25 (~3), ¾ of the grades are better than 3.25 (~3).

We look for the upper quartile in position 3(n+1)/4, which in the case of this data set, composed of 20 elements, is 3(20+1)/4 = 15.75. In such a data set composed of an even number of elements, the upper quartile also falls in between two values, which, in this case, are the 15th and the 16th elements, that is, 5→ the value of the upper quartile[7] is **5**.

¾ of the grades the student got in the semester are worse than excellent (5), and ¼ of the grades in this particular example are obviously excellent (5).

5.2.2 Measures of statistical dispersion

In the apparatus of statistical analysis, indices that quantify differences between data are of special significance. We get a better picture about the data set by calculating the range, standard deviation, relative standard deviation, and the mean deviation or mean difference.

Range

It is easy to determine the range, as it denotes the interval within which the values of the data set vary.

Range is the difference between the largest and the smallest values of a data set.

Thus, in order to determine the range, we need to know the maximum and minimum values of the data set. The index carries no information about how the data are arranged within these limits. The range can be confidently determined only for individual values.

$$R = X_{max} - X_{min}$$

For a class interval frequency array, the range could only be calculated as the difference between the upper limit of the last class in-

7 The upper quartile is calculated by adding ¾ of the difference between the two values t o the lower value.

terval and the lower limit of the first class interval. However, this would not lead to a real result, as in many cases, both the first and the last class intervals are open, and we closed them only in order for the basic distribution features to be computable. Therefore, we do not compute these range values.

In practice, the interquartile range delivers more information about the distribution of the elements.

- **The interquartile range represents the position of the internal 50% of the elements, and is defined as the difference between the upper and lower quartiles.**

$$IQR = Q_3 - Q_1$$

Standard deviation and relative standard deviation

Standard deviation and relative standard deviation are the two most important measures of dispersion. The way they are calculated, as well the way they are denoted, is different for populations and for samples.

- **Standard deviation represents the average squared deviation of elements from the arithmetic mean in the unit of measurement of the base data, while relative standard deviation represents it in percentage.**

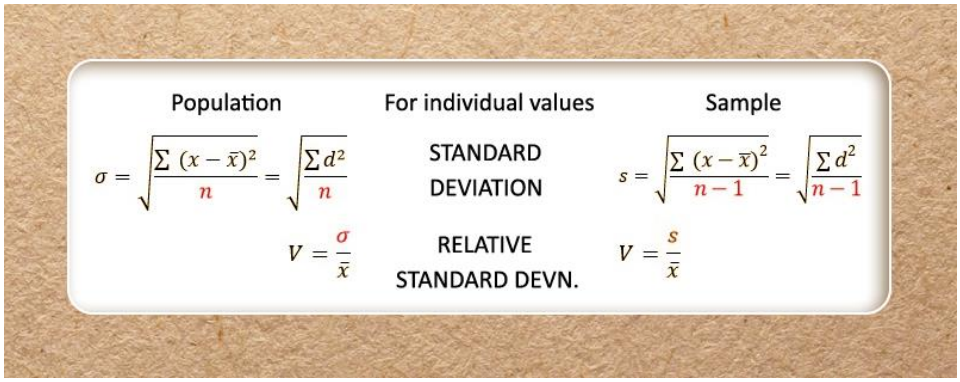


Figure 11 Calculation of the standard deviation and relative standard deviation for a population and a sample

While standard deviation is used for measuring dispersion in terms of the units of measurement of the base data, relative standard deviation is used for comparing data sets of different units of measurement. When calculating the standard deviation, first we take the deviation of elements from the arithmetic

mean, then we square it, because what is important is the degree of deviation, not whether it is positive or negative compared to the mean. Then we sum the squares of the pair-by-pair deviation values, and then divide the sum by the number of elements, and extract the square root of the result. As is clear from the method of calculation, the *value of the standard deviation is always POSITIVE*. In the case of relative standard deviation, we compare the mean deviation of elements from the mean to the mean. As the deviation from the mean cannot exceed the mean itself, the value of standard deviation is always smaller than the value of the mean. Therefore, the value of the *relative standard deviation*, which is the quotient of the standard deviation and the mean, *ALWAYS falls between 0 and 1*.

☸ Suppose a student got the following grades in a semester:
2, 3, 5, 5, 4, 4, 3, 5, 2, 5, 5, 4, 4, 5, 3, 4, 5, 5, 4, 4

Determine the values for the range, the interquartile range, the standard deviation, and the relative standard deviation for this data set.

$R = X_{max} - X_{min} = 5\text{-}2 = 3 \rightarrow$ data vary between 2 and 5 within an interval of 3 units in length.

$IQR = Q_3 - Q_1 = 5 - 3.25 = 1.75 \rightarrow$ the middle 50% of the data, between 3.25 and 5, vary within an interval of 1.75 in length.

$$\sigma = \sqrt{\frac{\sum f \cdot (x - \bar{x})^2}{n}}$$

$$\sigma = \sqrt{\frac{8 \cdot (5 - 4.05)^2 + 7 \cdot (4 - 4.05)^2 + 3 \cdot (3 - 4.05)^2 + 2 \cdot (2 - 4.05)^2}{20}} = \mathbf{0.9734}$$

$$V = \frac{\sigma}{\bar{x}} = \frac{0.9734}{4.05} = 0.2404 \rightarrow \mathbf{24.04}\%$$

The grades deviate from the mean by 0.9734 grades on average, i.e. by 24.04%.

Mean deviation and mean difference

There are various ways to quantify the differences in a data set. The simplest way to quantify the difference from the mean is to average the absolute deviations from the mean.

⌓ **Mean deviation numerically expresses the mean absolute deviation of the elements from the arithmetic mean.**

Values may be compared not only to a fix value, such as the arithmetic mean, but also to each other. Mean difference determines the deviation of data from each other.

- **Mean difference is the arithmetic mean of the absolute differences of the data from each other.**

We use this difference in the calculation of the GINI index, which is used to quantify concentration.

- **Concentration is the condensation or clustering of data around an element of a population.**

Measuring concentration by the GINI index is most common in practice in the quantification of income inequalities. The GINI coefficient expresses the distribution of the population of the world and its income. From a statistical perspective, it is much more efficient to study concentration by the Lorenz curve, which compares changes in relative frequencies to changes in relative value sums. The Lorenz curve is represented in a square with sides of one unit of length, whose diagonal represents the absence of concentration, as, along that line, relative frequency and the relative value sum are identical. The farther away from the diagonal the curve is, which connects values of relative frequency (g') and relative value sum (z') calculated from the data set, the larger the concentration.

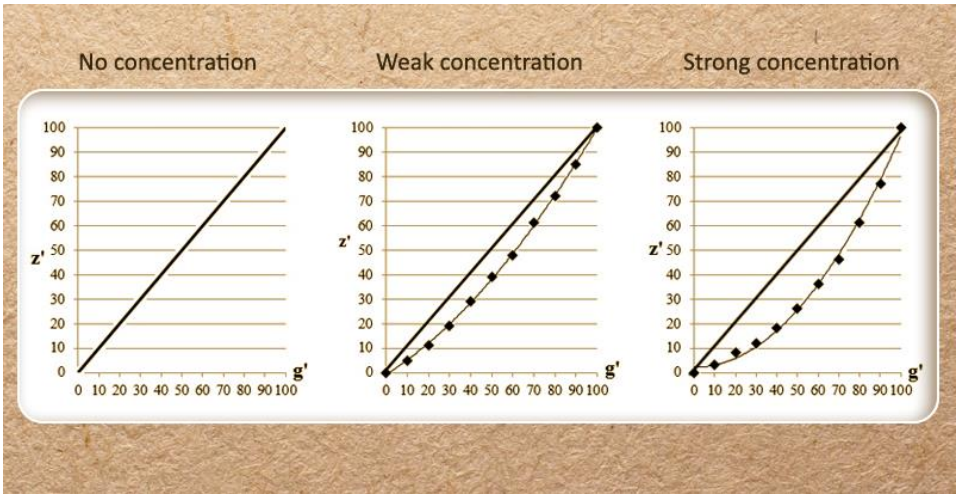


Figure 12 Study of concentration by Lorenz curve

5.2.3 Shape indices

Shape indices allow us to draw inferences about the shape of the frequency curve. We can quantify the deviation of the data set from a symmetric distribution, in terms of skewness or asymmetry, or in terms of curtosis.

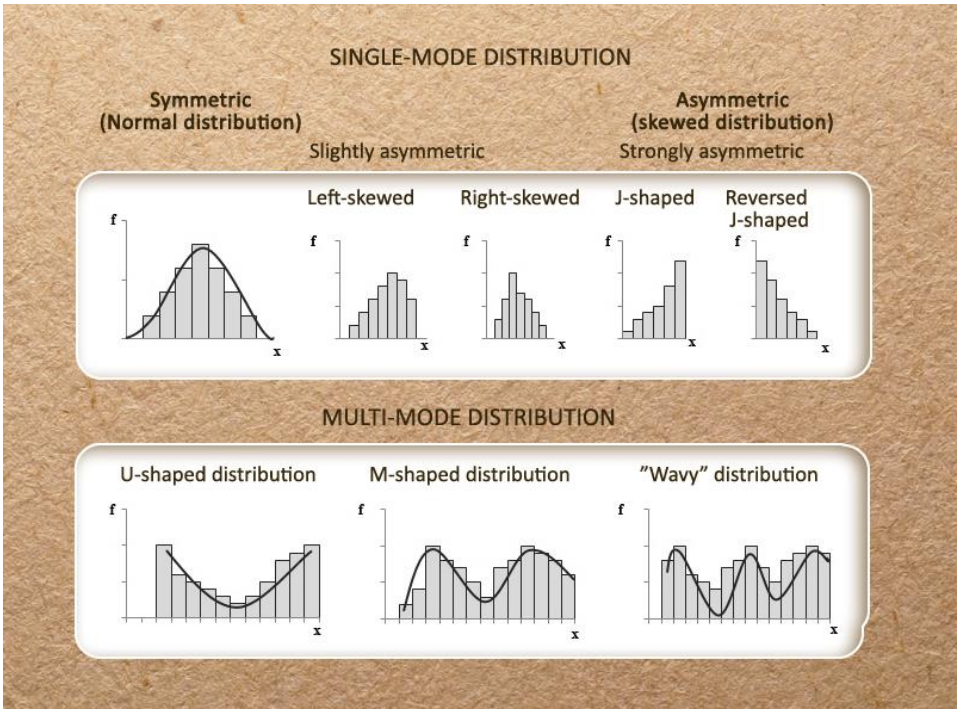


Figure 13 Types of empirical distribution

Measures of asymmetry (skewness)

The most commonly used measures for the quantification of asymmetry or skewness are Person's *A* coefficient and *F* coefficient. The *A coefficient* uses the arithmetic mean, mode, and standard deviation, while the *F coefficient* uses the lower and upper quartiles and the median, to determine the degree of asymmetry.

$$A = \frac{\bar{x} - Mo}{\sigma} \qquad F = \frac{(Q_3 - Me) - (Me - Q_1)}{(Q_3 - Me) + (Me - Q_1)}$$

The *A* coefficient has no absolute limit; the direction of the asymmetry is expressed by its sign.

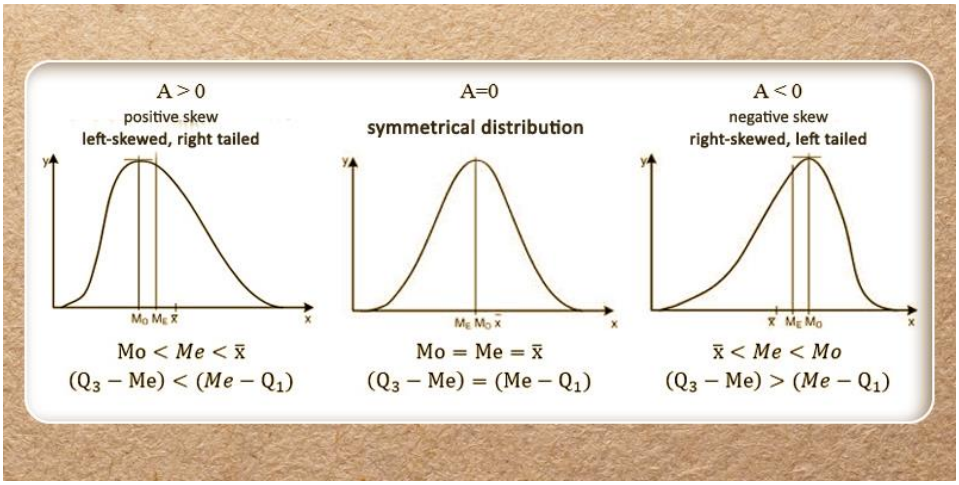


Figure 14 Properties of symmetric and asymmetric distributions

In symmetric distribution, the arithmetic mean, mode, and median coincide. Elements are symmetrically distributed in relation to that point, which is graphically represented by the normal bell curve. Whether the asymmetry is left-skewed or right-skewed, or whether it is left-tailed or right-tailed, depends of the position of the arithmetic mean and the mode:

- if, when compared to the mode, the arithmetic mean is greater, then the asymmetry is skewed to the left
- if, when compared to the mode, the arithmetic mean is smaller, then the asymmetry is skewed to the right
- if, when compared to the arithmetic mean, the mode is smaller, then the asymmetry is right-tailed
- if, when compared to the arithmetic mean, the mode is greater, then the asymmetry is left-tailed

The value of the *F* coefficient falls between -1 and 1, i.e., in absolute value, between 0 and 1. The *F* coefficient also represents the strength of the asymmetry: the closer it is to 0, the weaker is the asymmetry, and approaching the maximum values, asymmetry gets stronger and stronger.

Curtosis

This shape index represents magnitude differences of the data, compared to symmetric distribution, and causes the distribution curve to be peaked or flat. The coefficient of curtosis represents deviation from the normal: if its value is positive, then the curve is peaked; if its value is negative, it is flat.

5.3 SUMMARY AND QUESTIONS

5.3.1 Summary

Descriptive statistical tools are commonly used in statistical analysis. It is very important in primary data collection that the data set is concisely characterized and offers plenty of information.

Means share the property that they always fall between the lowest and the largest elements in a data set. Calculated and positional means are equally easy to determine, though they have a disadvantage: they can be used confidently only if some specific conditions are met. It is important to note that the way we calculate averages is determined by the type of data we have. In practice, we often compute not only a mean, but also the standard deviation, because the two jointly give us a complex picture about the data set. For further statistical analysis of your data, you need a more detailed quantification of dispersion and the form of distribution, which allows you to determine asymmetry and curtosis, which, in practice, normally takes the form of graphical representations.

5.3.2 Self-test questions

How can you classify the tools of descriptive statistics?

What computed means do you know?

What are the advantages and disadvantages of calculating the arithmetic mean?

What are quantiles?

What is mode, and what are the advantages and disadvantages of its calculation?

What do measures of dispersion numerically express?

What is the difference between standard deviation and relative standard deviation?

How can you measure and illustrate concentration?

How do the mean and the mode relate to each other in left-skewed asymmetry?

How do the mean and the mode relate to each other in right-skewed asymmetry?

5.3.3 Practice tests

❁ 100 customers were observed on a particular day doing shopping in a hypermarket. Half of the customers spent over € 25 and half spent less than that. *Which basic distribution descriptor is interpreted?*

Means

	Mean
	Mode
X	Median
	Lower quartile
	Upper quartile

Dispersion and shape indices

	Mean
	Mode
	Median
	Lower quartile
	Upper quartile

Customers spent € 24 on average.

Means

X	Mean
	Mode
	Median
	Lower quartile
	Upper quartile

Dispersion and shape indices

	Range
	Standard deviation
	Relative standard devn.
	A index
	F index

The amount of money spent by individual customers deviates from the average amount of money spent by € 4 on average.

Means

	Mean
	Mode
	Median
	Lower quartile
	Upper quartile

Dispersion and shape indices

	Range
X	Standard deviation
	Relative standard devn.
	A index
	F index

Most customers spent € 20.

Means

	Mean
X	Mode
	Median
	Lower quartile
	Upper quartile

Dispersion and shape indices

	Range
	Standard deviation
	Relative standard devn.
	A index
	F index

¼ of the customers spent less than € 15 and ¾ spent more than that.

Means

	Mean
	Mode
	Medián
X	Lower quartile
	Upper quartile

Dispersion and shape indices

	Range
	Standard deviation
	Relative standard devn.
	A index
	F index

❁ Decide whether the statements below are true (T) or false (F).

Mode is sensitive to outstandingly low or outstandingly high values in the data set.	F
Quantiles are special data values that mark the boundaries in a data set.	T
Standard deviation expresses the average deviation between data in terms of the unit of measurement of the base data, while relative standard deviation expresses the same in percentage.	F
The Lorenz curve can be used for the study of concentration.	T
The *F* coefficient can be used for the numerical expression of curtosis.	F

6. COMPARISON OF COMPLEX RATIOS (GRAND MEANS) BY STANDARDIZATION

6.1 GOALS AND COMPETENCIES

Populations (or samples) of interest selected for statistical analysis are very often not homogeneous. This makes the use of simple analytic tools difficult. We can handle heterogeneity by dividing a heterogeneous population into homogeneous subpopulations on the basis of an attribute and then we carry out our analyses not only on the entire population but also on its subpopulations. What we most commonly do on heterogeneous populations is carry out comparisons, which may be done on the basis of temporal, areal, or qualitative features. Factors responsible for differences between populations or for changes in a particular population may be studied by the method of standardization.

The purpose of this unit is to familiarize students with a statistical method that enables them to use the simple statistical methods in the analysis of heterogeneous populations. Students will learn about ways to decompose differences and the method of index calculation based on standardization and its use. Students will learn to interpret the components and indices that cause a difference or a change and understand their interrelationships.

6.2 TOPICS

Heterogeneous populations may be analyzed in a variety of different ways by simple statistical methods, but this requires that we do something about their heterogeneity. One way to handle heterogeneous populations is to break them up into homogenous subpopulations. This involves creating groups within the entire population on the basis of some distinctive attribute (such as, for example, dividing the class of statistics students into groups according to their major subjects). If you want to use simple statistical methods in the study of a population which is heterogeneous in regard to a feature of interest, then you must also examine the homogeneous subpopulations.

The subpopulations of a population become comparable and analyzable on the basis of qualitative, areal, or temporal features. A qualitative or areal comparison can be conducted by decomposing the difference, whereby we decompose the difference in the overall index, the grand mean, or complex ratio, which refers to the two populations jointly. In the study of temporal changes within a population, we use the method of

index calculation based on standardization, in which we decompose the quotient of the grand means of the recent (later) and the preceding (earlier) period.

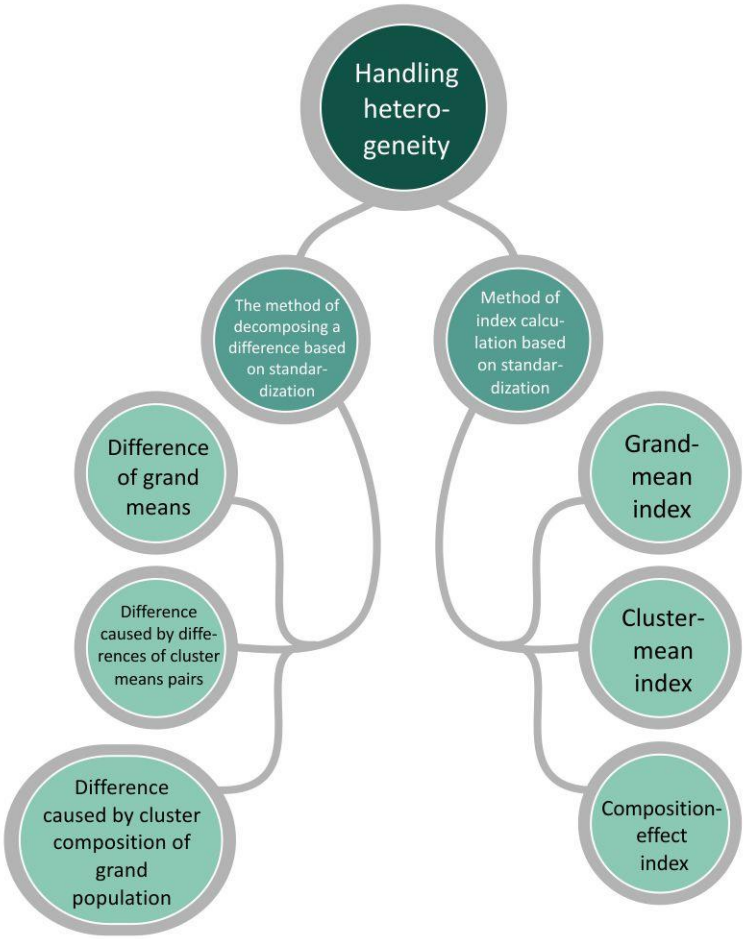


Figure 15 Methods of comparing complex ratios (grand means)

6.2.1 Handling heterogeneity: clustering

A heterogeneous population is not adequately characterized solely by an average. You need the averages of the subpopulations as well.

- **The mean defined on the entire population is the grand mean. A mean defined on a subpopulation is a cluster mean. The grand mean is the arithmetic mean of cluster means weighted by the ab-**

solute or relative frequency of the subpopulations. The grand mean can also be calculated in the form of a weighted harmonic mean, depending on the data available.

Average values, as determined by the form in which they are calculated, are intensity ratios, which can be derived as quotients of two different types of data, in the case of a clustered population, from data in the non-clustering rows.

Ratios that refer to the particular clusters are called cluster ratios, and the ratio that refers to the entire population is called a complex ratio.

The employees of a company are divided into blue-collar and white-collar workers. We can determine their average wages separately, if we know the total of wages for all the employees and we know the number of employees in each subcategory. We can carry out the same type of calculation for all employees in the knowledge of the data.

17. *Average wages of blue-collar and white-collar workers*

Subcategory of employees	Total wages (Ft)	Number of employees (prsns)	AVERAGE WAGE (Ft/prsn)	
	A	**B**	$\mathbf{V = \frac{A}{B}}$	
Blue-collar	27 500 000	250	**110 000**	← **cluster ratio / cluster mean**
White-collar	21 750 000	150	**145 000**	← **cluster ratio / cluster mean**
Totals	*49 250 000*	*400*	***123 125***	← ***complex ratio / grand mean***

The complex ratio ($\overline{V}$) and grand mean ($\overline{X}$) can be defined in various ways, using different symbols.

$$\overline{V} = \frac{\sum A_i}{\sum B_i} = \frac{\sum(\mathbf{B_i} \cdot \mathbf{V_i})}{\sum \mathbf{B_i}} = \frac{\sum A_i}{\sum \frac{A_i}{V_i}} \qquad \overline{X} = \frac{\sum A_i}{\sum B_i} = \frac{\sum(\mathbf{B_i} \cdot \overline{\mathbf{x}}_\mathbf{i})}{\sum \mathbf{B_i}} = \frac{\sum A_i}{\sum \frac{A_i}{\overline{x}_i}}$$

The average wages of all the employees can be calculated as follows, with the use of the formulae above:

$$\frac{\sum A_i}{\sum B_i} = \frac{49\,250\,000}{400} = 123\,125$$
$$\frac{\sum (B_i \cdot V_i)}{\sum B_i} = \frac{250 \cdot 110\,000 + 150 \cdot 145\,000}{400} = 123\,125$$
$$\frac{\sum A_i}{\sum \frac{A_i}{V_i}} = \frac{49\,250\,000}{\frac{27\,500\,000}{110\,000} + \frac{21\,750\,000}{145\,000}} = 123\,125$$

The value of the grand mean is affected by two factors. On the one hand, it is affected by the values of the cluster means, and, on the other hand, by the number of elements in the subpopulations (weight). The grand mean must fall between the smallest and the largest cluster means, as this is an average, and its value is affected by the distribution of the subpopulations within the entire population.

The grand means of two identically clustered populations can be compared. The difference is in part caused by the difference between cluster means of subpopulations within the entire population, and in part by a difference in the composition of the entire population in terms of its subpopulations. Comparison can be carried out on the basis of qualitative, areal, or temporal attributes. For the decomposition of a difference in qualitative or areal comparisons, we must specify in advance which population's data will be taken as the data to compare and which will be the basis of comparison, which must also be indicated in the subscript. In temporal comparison, however, the order of comparison is fixed, as we compare the data of the later period (reference) to the data of the earlier period (base). Data that refer to the base period are indexed with a 0 subscript, while data that refer to the reference period are indexed with the subscript 1. Thus, we have a formula for each index calculation.

6.2.2 The method of decomposing a difference based on standardization

The method of decomposing a difference is employed in the qualitative of areal comparison of heterogeneous populations. The difference between the grand means of the two populations is in part caused by the difference between paired cluster means and in part by the difference in the composition of grand populations in terms of their subpopulations.

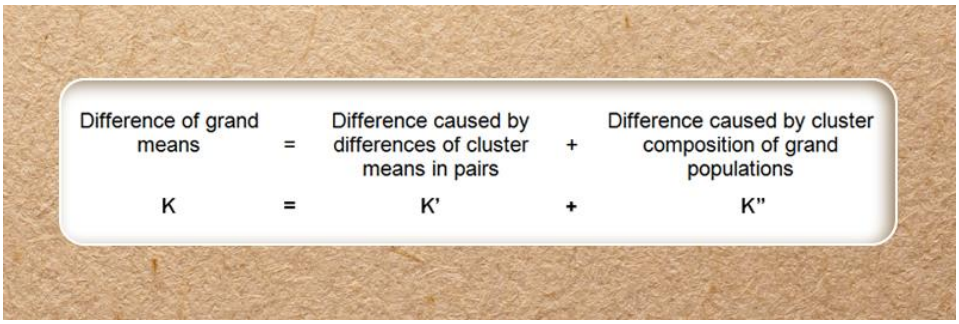


Figure 16 Factors of difference decomposition and their relationship

In standardization, we decompose the difference between grand means into an effect caused by two factors in such a way that we examine only the difference in terms of one factor at a time, regarding the other unchanged. In other words, if we intend to identify the effect of the difference in the cluster means, the composition remains constant (standard composition), while if we are interested in the effect of the composition in terms of subpopulations, then cluster means remain constant (standard cluster mean). Standard values are determined thus: for K', the composition of the first grand population (data to compare) is standard ($B_S = B_1$), while for K", the cluster mean of the second grand population (basis of comparison) is unchanged ($V_S = V_2$).

In decomposing a difference, first we determine the difference of the grand means (K), and then, separately, we quantify the effects of differences of cluster means (K') and of the differences of composition (K"). The main point of the method is that if we want to determine the effect of the difference in individual cluster means, then the composition of the clusters must be kept unchanged (standard), and if we are interested in the effect of the difference in composition, then we keep the cluster ratios unchanged. When determining K', the composition of the minuend is standard, while in the calculation of K" the cluster ratios of the subtrahend remain unchanged.

$$\mathbf{K}_{1-2} = \overline{V_1} - \overline{V_2} = \frac{\sum B_1 V_1}{\sum B_1} - \frac{\sum B_2 V_2}{\sum B_2}$$

$$\mathbf{K'}_{1-2} = \frac{\sum \mathbf{B_S} V_1}{\sum \mathbf{B_S}} - \frac{\sum \mathbf{B_S} V_2}{\sum \mathbf{B_S}}, where\ B_s = B_1$$

$$\mathbf{K''}_{1-2} = \frac{\sum B_1 \mathbf{V_S}}{\sum B_1} - \frac{\sum B_2 \mathbf{V_S}}{\sum B_2}, where\ V_S = V_2$$

It is useful to indicate the direction of the subtraction in a subscript.

In some cases, only one of the factors accounts for the difference. In such cases, the value of the difference of identical factors is 0.

If K' = 0, then K = K" OR if K" = 0, then K = K'

❁ In a statistics class composed of 76 students, some students major in Technical management (TM), some others in Human resources management (HR), and some in Tourism and catering (TC). All students completed an examination test. They completed two different variants of the test, variant A and variant B. *Variant A* was completed by 40 students, composed of 20 TM students, 5 HR students, and 15 TC students. Variant B was completed by 36 students, composed of 18 TM students, 6 HR students, and 12 TC students. Regarding *Variant A*, we also know that the average score of TM students was 36.2 points, the average score of HR students was 40.4 points, and the average score of TC students was 35.8 points. Regarding Variant B, we know that the average score of TM students was 38.4 points, the average score of HR students was 32.6 points, and the average score of TC students was 39.2 points. *Determine the average score of students who completed Variants A and B.*

$$\overline{V_A} = \frac{\sum B_A \cdot V_A}{\sum B_A} = \frac{20 \cdot 36.2 + 5 \cdot 40.4 + 15 \cdot 35.8}{40} = 36.575 \sim \mathbf{36.6}$$

$$\overline{V_B} = \frac{\sum B_B \cdot V_B}{\sum B_B} = \frac{18 \cdot 38.4 + 6 \cdot 32.6 + 12 \cdot 39.2}{36} = \mathbf{37.7}$$

18. *Average test scores of students in the Statistics class*

Major subject	VARIANT A		VARIANT B	
	No. of students (prsns)	Average score	No. of students (prsns)	Average score
	B_A	V_A	B_B	V_B
TM	20	36.2	18	38.4
HR	5	40.4	6	32.6
TC	15	35.8	12	39.2
Totals	40	**36.6**	36	**37.7**

Use the method of difference decomposition to analyze the difference in the average results of the two variants of the test. Find a

way to quantify the degree to which different factors contribute to the differences in question.

$$\boldsymbol{K_{B-A}} = \overline{\boldsymbol{V_B}} - \overline{\boldsymbol{V_A}} = 37.7 - 36.6 = \mathbf{1.1}$$

The average score of *Variant B* is 1.1 points higher than that of Variant A. This is due to two factors:

- different average scores achieved by different student subgroups (by major subject) on the test (**K'**: effect of the difference in cluster means – standard composition)
- distribution of students in terms of their major subjects in regard to the two test variants (**K"**: composition-effect – standard cluster means)

In the calculation of K', the direction of the subtraction is decisive for the choice of composition as standard; we take the composition of the minuend (B in this case) to be unchanged.

$$\boldsymbol{K'_{B-A}} = \frac{\sum \boldsymbol{B_B V_B}}{\sum \boldsymbol{B_B}} - \frac{\sum \boldsymbol{B_B V_A}}{\sum \boldsymbol{B_B}}$$

$$\boldsymbol{K'_{B-A}} = 37.7 - \frac{18 \cdot 36.2 + 6 \cdot 40.4 + 12 \cdot 35.8}{36} = 37.7 - 36.8 = \mathbf{0.9}$$

Interpretation: for test variants A and B, due to the difference in cluster means (corresponding to student subgroups by major), the average score of students on variant B would be 0.9 points higher; i.e., if the only difference was that different clusters of students achieved different scores (assuming identical distribution by major on each test variant = standard composition), then the average score of variant B would be higher only by 0.9 points.

In the calculation of K'', when the cluster mean is chosen as standard, the direction of subtraction is equally decisive; we take the cluster mean of the subtrahend (A in this case) as unchanged. It is important that the standard factor for the calculation of K' and K" be chosen from different populations.

$$\boldsymbol{K''_{B-A}} = \frac{\sum \boldsymbol{B_B V_A}}{\sum \boldsymbol{B_B}} - \frac{\sum \boldsymbol{B_A V_A}}{\sum \boldsymbol{B_A}} = 36.8 - 36.6 = \mathbf{0.2}$$

Interpretation: for test variants A and B, due to the difference in student distribution by major subject, the average score of students on variant B would be 0.2 points higher; i.e., if only the distribution of different subgroups of students was different for each test variant (assuming identical cluster means for each subgroup

on each test variant = standard cluster mean), then then the average score of variant B would be higher only by 0.2 points.

Once we are done with the difference decomposition, if we did everything correctly, what we have on each side should be identical with the other, i.e., the sum of the factors must be equal to the difference between the grand means. Let us check.

Check: K = K' + K'' → in this particular case 1.1 = 0.9 + 0.2 ✓

6.2.3 Method of index calculation based on standardization

In standardization-based index calculation, first we quantify the relative change between the two periods. Then we consider the change in the cluster ratios (I'), which we calculate assuming unchanged composition, with the later period remaining standard. For the determination of change in composition (I"), we take the cluster means unchanged, choosing the ratios of the earlier period as standard.

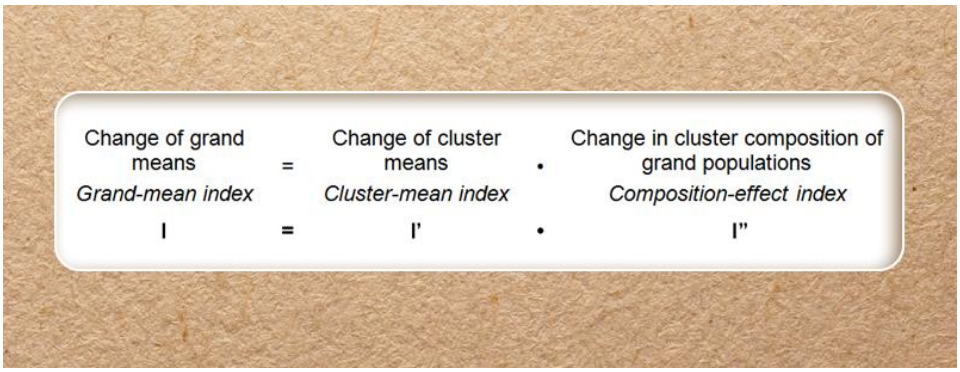


Figure 17 Factors of standardization-based index calculation and their relationship

- We have a freshmen's class of students majoring in Technical management (TM), Human resources (HR) and Tourism and catering (TC). We have the following information on their semester-final grade point average for the two semesters of the 2012/2013 academic year. 47 students enrolled for the first semester of the 2013/2014 academic year, out of which, 16 were Technical management majors, 5 were Human resources majors, and 26 were Tourism and catering majors. Due to various factors of study management and logistics, 52 students enrolled for the second semester of the 2013/2014 academic year. 14 of them were Technical

management majors, 8 were Human resources majors, and 30 were Tourism and catering majors. We also know, in relation to the first semester of the 2013/2014 academic year, that the grade point average (GPA) of TM students was 3.8, the GPA of HR students was 4.4, and that the GPA of TC students was 3.6. We also know concerning the second semester of the 2013/2014 academic year that the GPA of TM students was 3.82, the GPA of HR students was 4.0, and that the GPA of TC students was 3.5

Determine the GPAs for each semester.

$$\overline{V_0} = \frac{\sum B_0 \cdot V_0}{\sum B_0} = \frac{16 \cdot 3.8 + 5 \cdot 4.4 + 26 \cdot 3.6}{47} = \mathbf{3.75}$$

$$\overline{V_1} = \frac{\sum B_1 \cdot V_1}{\sum B_1} = \frac{14 \cdot 3.82 + 8 \cdot 4.0 + 30 \cdot 3.5}{52} = \mathbf{3.66}$$

The GPA of the entire class in the first semester of the 2013/2014 academic year was 3.75, and in the second semester it was 3.66.

19. *GPAs of the class in the 2013/2014 academic year*

Major subject	2013/2014 1st semester		2013/2014 2nd	
	No. of students (prsns)	Grade Point Average	No. of students (prsns)	Grade Point Average
	B_0	V_0	B_1	V_1
TM	16	3.8	14	3.82
HR	5	4.4	8	4.0
TC	26	3.6	30	3.5
Totals	47	**3.75**	52	**3.66**

In calculating the index, we always compare the reference period to the base period, since changes occur as time progresses forward, which is why there is no need to additionally indicate the direction of the change.

Use the method of standardization-based index calculation to analyze changes in the Grade Point Average of students. Find a way to quantify what factors contributed to the change and the degree to which they did.

$$I = \frac{\overline{V_1}}{\overline{V_0}} = \frac{\frac{\sum B_1 V_1}{\sum B_1}}{\frac{\sum B_0 V_0}{\sum B_0}} = \frac{3.66}{3.75} = \mathbf{0.976} \xrightarrow{\cdot 100} \mathbf{97.6} \rightarrow \mathbf{-2.4\%}$$

There was a drop in GPAs in the 2nd semester by 2.4 %. This was due to two factors:

- different GPAs of different groups of students (by major) (**I'**: effect of change in cluster means – standard composition in reference period)
- changes in the composition of students (in terms of their majors) in each semester (**I''**: effect of change in composition – standard base-period cluster means)

When changes occur in time, we use standard reference-period composition for the calculation of the index, i.e., in calculating I', we take the distribution of groups in the reference period to be unchanged.

$$\boldsymbol{I'} = \frac{\frac{\sum \boldsymbol{B_S} V_1}{\sum \boldsymbol{B_S}}}{\frac{\sum \boldsymbol{B_S} V_0}{\sum \boldsymbol{B_S}}} = \frac{\frac{\sum \boldsymbol{B_1} V_1}{\sum \boldsymbol{B_1}}}{\frac{\sum \boldsymbol{B_1} V_0}{\sum \boldsymbol{B_1}}} = \frac{3.66}{\frac{14 \cdot 3.8 + 8 \cdot 4.4 + 30 \cdot 3.6}{52}} = \frac{3.66}{3.78} = \mathbf{0.968} \xrightarrow{\cdot 100} \mathbf{96.8} \rightarrow -3.2\%$$

Interpretation: Considering the entire class of students in the 2nd semester of the academic year 2013/2014, GPAs dropped by 3.2%, due to changes in the GPAs of particular groups of students. That is to say that if all that changed was the GPAs of different groups of students (assuming identical student group distribution in each semester = standard composition), then GPAs would have dropped overall by 3.2%.

When changes occur in time, we use standard base-period cluster means in the calculation of the index, i.e., we regard the cluster means of the base period as unchanged.

$$\boldsymbol{I''} = \frac{\frac{\sum B_1 \boldsymbol{V_S}}{\sum B_1}}{\frac{\sum B_0 \boldsymbol{V_S}}{\sum B_0}} = \frac{\frac{\sum B_1 \boldsymbol{V_0}}{\sum B_1}}{\frac{\sum B_0 \boldsymbol{V_0}}{\sum B_0}} = \frac{3.78}{3.75} = \mathbf{1.008} \rightarrow +\mathbf{0.8}\%$$

Interpretation: Considering the entire class of students in the 2nd semester of the academic year 2013/2014, GPAs would have improved by 0.8% due to changes in the group distribution of students (in terms of their major subjects). That is to say that if all that changed was the number students in the subgroups (by major), i.e., the composition of the populations (assuming identical GPAs in the subgroups by major in each semester), then GPAs would have improved overall by 0.8%.

If we did everything correctly, when we are done with calculating the index, what we should have is that each side is identical with

the other, i.e., the power of the particular factors should equal the value of the grand mean index. Let us check.

Check: $I = I' \times I'' \rightarrow$ in this case $0.976 = 0.968 \times 1.008$ ✓

6.3 SUMMARY AND QUESTIONS

6.3.1 Summary

Heterogeneous populations may be analyzed in a variety of different ways by simple statistical methods, but this requires that we do something about their heterogeneity. We can construct homogeneous subpopulations within a population by clustering, which allows us to carry out temporal and qualitative or areal comparisons. The method of standardization allows us to quantify the effect of factors that account for a difference or change. In decomposing a difference, we decompose the absolute difference in the grand ratios of the populations into factors. To find a way to quantify temporal factors that account for change, we employ the method of index calculation based on standardization.

6.3.2 Self-test questions

When can you use the method of decomposing a difference?

When can you use the method of index calculation based on standardization?

Into what factors can you decompose the numerical difference in grand means?

Into what factors can you decompose the grand mean index that represents changes in grand means?

What is the standard factor in calculating the difference due to the difference in pairs of cluster means?

What is the standard factor in calculating the difference due to the difference in the composition of grand populations in terms of their subpopulations?

How do we decide the direction of comparison in index calculation?

The composition of which period is taken as unchanged in determining the cluster mean index?

The cluster means of which period are taken as unchanged in calculating the index of composition effect?

What is the relationship between the factors of difference decomposition and the factors of index calculation?

6.3.3 Practice tests

☸ Choose the correct answer for each question below.

An overview of the wages of the employees of a company shows that men make HUF 30 000 more on average than women. We know that due to the difference in the distribution of professional qualifications among men and women, men would make only HUF 24 000 more than women.

Due to the difference in average wages between men and women by professional qualification

a) men would make HUF 6000 less than women
b) men would make HUF 6000 more than women
c) women would make HUF 6000 more than men

An overview of the wages of the employees of a company shows that men make HUF 30 000 more on average than women. We know that due to the difference in average wages by professional qualification between men and women, men would make only HUF 6 000 more than women. Due the difference in the distribution of men and women by professional qualification,

a) men would make HUF 24 000 less than women
b) men would make HUF 24 000 more than women
c) women would make HUF 24 000 more than men

An overview of data on two resort areas shows that the average amount of time spent by visitors has dropped by 10% compared to last year. On average, the time spent by visitors from Austria, Germany, and Poland has increased by 2%, but the yearly distribution of nationalities has changed so much that it has caused an average drop of 11.76% in overall time spent by visitors.

a) the value of the grand mean index is 110.0
b) the value of the cluster mean index is 102.0
c) the value of the composition-effect index is 111.76

An overview of data on two resort areas shows that the average amount of time spent by visitors has dropped by 10% compared to last year. On average, the time spent by visitors from Austria, Germany, and Poland has increased by 2%, but the yearly distribution of nationalities has changed so much that it has caused an average drop of 11.76% in overall time spent by visitors.

a) the value of the grand mean index is 90.0

b) the value of the cluster mean index is 98.0
c) the value of the composition-effect index is 111.76

An overview of data on two resort areas shows that the average amount of time spent by visitors has dropped by 10% compared to last year. On average, the time spent by visitors from Austria, Germany, and Poland has increased by 2%, but the yearly distribution of nationalities has changed so much that it has caused an average drop of 11.76% in overall time spent by visitors.

a) the value of the grand mean index is 110.0
b) the value of the cluster mean index is 98.0
c) the value of the composition-effect index is 88.24

☸ Decide whether the statements below are true (T) or false (F).

When calculating a temporally comparative index, the direction of the comparison can be chosen freely.	**F**
The grand mean index is the power of the cluster mean and the comparative index.	**T**
The way to determine the effect of the difference in cluster means is to take the composition of the grand population by subpopulations unchanged.	**T**
If the composition of two populations compared is identical, then the effect of the difference in composition by subpopulations is 1.	**F**
If the value of cluster means is unchanged from one period to another, then the value of the grand mean index is identical with value of the cluster mean index.	**F**

7. VALUE-BASED INDEX COMPUTATION

7.1 GOALS AND COMPETENCIES

Ratios may occur in various forms in the context of economy. Indices are mostly dynamic ratios, which are derived as quotients of data from a later period and data from an earlier period, in reference to a particular population. Areal indices can be derived on analogy with areal comparative ratios. The method of value-based index computation can be used for the temporal and areal comparison of prices and quantities, and values, which can be defined as the product of the other two. By computing indices we can quantify changes in price and quantity numerically in relative form. Furthermore, by introducing the concept of value to statistical analysis, we will be able to carry out temporal and areal comparison even in the case of heterogeneous product groups. The category of value is of exceptional significance in economic computations, because it allows us to sum various data, even when they are expressed in different units of measurement or unit prices.

The purpose of this unit is to familiarize students with the method of value-based index computation, which centers on the temporal and areal comparative study of price, quantity, and value. Students will learn to be able to define and interpret the concept of value and to be able to identify its realizations in practice. They will understand the difference, as well as the relationship, between elementary and aggregate indices, and acquire the logic of their computation.

7.2 TOPICS

The method of value-based index computation is used for the temporal and areal comparison of data on price and quantity (volume) and value, defined as the product of the other two.

- **Ratios derived from a temporal or areal comparison of price, quantity, and value data are called index numbers.**

Comparison may be carried out with respect to a specific type of product or a heterogeneous group of products composed of products of different kinds, often associated with different units of measurement, which, therefore, cannot be summed directly. In the case of temporal comparison, we quantify the changes in value, price, and volume between two periods of time. In the case of areal comparison, we compare prices and volumes between two geographical units.

In order to understand the forms of computation, we need to clarify the following concepts and formulas. Value is defined as the product of the unit price of some good and its quantity or volume.

$$\text{value} = \text{price} \cdot \text{quantity} \rightarrow \mathbf{v} = \mathbf{p} \cdot \mathbf{q}$$

The concept of value covers the following notions, among other things: sales receipts, turnover, gross production or sales.

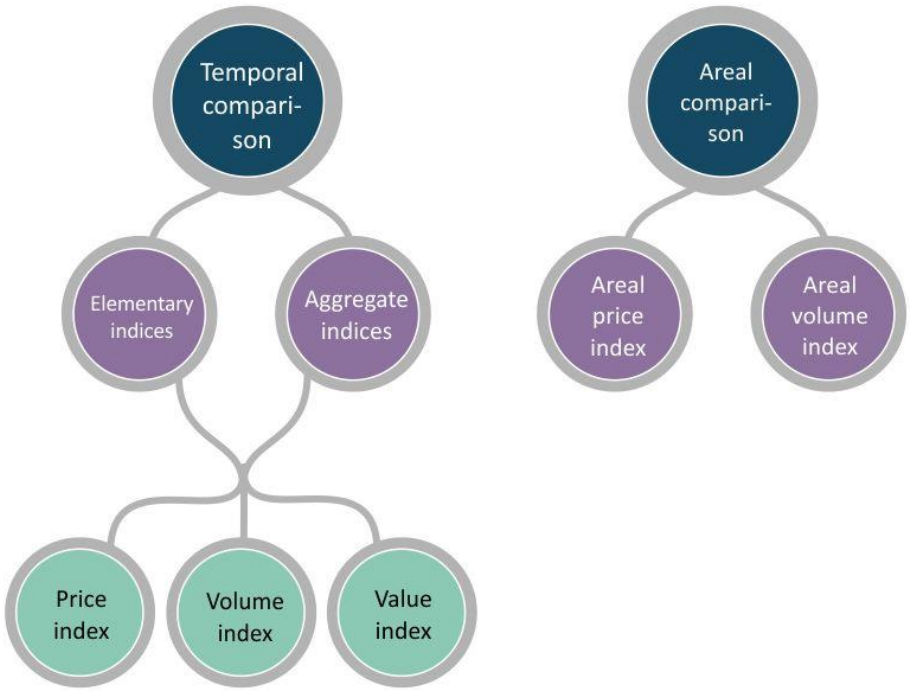


Figure 18 System of relationships in value-based index computation

7.2.1 Temporal comparison

The method of value-based index computation is most commonly used for temporal comparison. We indicate the earlier period (base) with the subscript 0, and the later period with (reference) with the subscript 1. The method can be easily understood through an example.

- *Suppose we have bought several different products in a shop, such as, for ample, some bread, milk, fruit, etc. Assuming we buy the same products, we can compare the prices of particular products, the amount we bought, and the amount of money we paid for*

them. So, we can quantify the changes in reference to particular products, and, with reference to the entire shopping basket, we can also express the changes in the amount of money paid, the degree (in percentage) to which prices have risen or fallen, or how much amounts have increased or decreased, either separately, or in conjunction with the prices.

Indices that quantify changes in prices, volume, or value are subcategorized as elementary or aggregate, depending on what group of products they relate to.

- **For a particular product (or a homogeneous group of products), changes in quantity, price, and value are quantified by elementary indices. For a heterogeneous group of different types of products, changes in quantity, price, and value are quantified by aggregate indices.**

Elementary indices

For a particular good, a change in price, quantity, and value may be expressed by computing an elementary index.

- **An elementary price index expresses the change in the price of a product between two periods of time.**

$$\textbf{elementary price index} = \frac{\textbf{price data of later period}}{\textbf{price data of earlier period}} \rightarrow i_p = \frac{p_1}{p_0}$$

- **An elementary volume index expresses the change in the quantity of a product between two periods of time.**

$$\textbf{elementary volume index} = \frac{\textbf{quantity data of later period}}{\textbf{quantity data of earlier period}} \rightarrow i_q = \frac{q_1}{q_0}$$

- **An elementary value index expresses the change in the value of a product between two periods of time.**

$$\textbf{elementary value index} = \frac{\textbf{value data of later period}}{\textbf{value data of earlier period}} \rightarrow i_v = \frac{v_1}{v_0}$$

Relationships among elementary indices

The elementary value index is the product of the elementary price index and the elementary volume index. On the basis of this relationship, given any two elementary indices, the third one can be computed.

$$\mathbf{i_v = i_p \cdot i_q} \rightarrow \; i_p = \frac{i_v}{i_q} \text{ or } i_q = \frac{i_v}{i_p}$$

❁ Suppose a family bought 20 kg of bread and 40 l of milk in March, while in April, they bought 22 kg of bread and 38 l of milk. The price of 1 kg of bread in March was HUF 300 and milk cost HUF 200 a liter, while in April the price of bread went up to HUF 320 and the price of milk went down to HUF 180.

Summarize the data in a table, and determine the amount of money the family spent on bread and milk in each month together and separately.

	MARCH			APRIL		
	Quantity consumed	Unit price	Amount of money spent on products VALUE	Quantity consumed	Unit price	Amount of money spent on products VALUE
	q_0	p_0	$v_0 = q_0 \cdot p_0$	q_1	p_1	$v_1 = q_1 \cdot p_1$
Bread (kg)	20	300	$20 \cdot 300 = \mathbf{6000}$	22	320	$22 \cdot 320 = \mathbf{7040}$
Milk (1)	40	200	$40 \cdot 200 = \mathbf{8000}$	38	180	$38 \cdot 180 = \mathbf{6840}$
Together			**14 000**			**13 880**

Figure 19 Amount of money spent on purchasing the products in each month

Calculating the values allows us to make the following statements:

- in March, the family spent HUF 6000 on bread and HUF 8000 on milk, so they spent a total of HUF 14 000 on the two products together
- in April, the family spent HUF 7040 on bread and HUF 6840 on milk, so they spent a total of HUF 13 880 on the two products together

As the values in the table show, the family consumed more bread in April, and its price rose too, so they had to spend more on purchasing bread. In contrast, both the amount of milk consumed and its price went down, therefore, the amount of money spent on it decreased too. Altogether, the family spent less on purchasing these products in April than in March.

How did the amounts consumed of the individual products, their unit prices, and the amount of money spent on purchasing them change in April, relative to March?

20. *Changes in amounts consumed, unit prices, and amount of money spent on purchasing the products from March to April*

	Change in amount consumed	**Change in unit price**	**Change in amount of money spent on purchasing products**
	$\boldsymbol{i_q = \frac{q_1}{q_0}}$	$\boldsymbol{i_p = \frac{p_1}{p_0}}$	$\boldsymbol{i_v = \frac{v_1}{v_0}}$
Bread	$\frac{22}{20} = 1.1 \overset{\cdot 100}{\longrightarrow} \mathbf{110.0}$	$\frac{320}{300} = 1.0667 \overset{\cdot 100}{\longrightarrow} \mathbf{106.67}$	$\frac{7040}{6000} = 1.1733 \overset{\cdot 100}{\longrightarrow} \mathbf{117.33}$
Milk	$\frac{38}{40} = 0.95 \overset{\cdot 100}{\longrightarrow} \mathbf{95.0}$	$\frac{180}{200} = 0.9 \overset{\cdot 100}{\longrightarrow} \mathbf{90.0}$	$\frac{6840}{8000} = 0.855 \overset{\cdot 100}{\longrightarrow} \mathbf{85.5}$

The indices so computed may be interpreted as follows:

- the family consumed 10% more bread and 5% less milk in April than they did in March
- the price of bread was 6.67% higher, while the price of milk was 10% lower in April than it was in March
- the family spent 17.33% more on purchasing bread and 14.5% less on milk in April than they did in March

Aggregate indices

We compute aggregate indices to express temporal changes in price, quantity, and value of a heterogeneous group of goods, typically associated with different units of measurement.

For multiple goods, the average aggregate change in value is expressed by the aggregate value index.

$$I_v = \frac{\sum v_1}{\sum v_0} = \frac{\sum p_1 \cdot q_1}{\sum p_0 \cdot q_0}$$

How did the amount of money spent on purchasing the two products taken together change between the two periods?

$$I_v = \frac{\sum v_1}{\sum v_0} = \frac{13\,880}{14\,000} = 0.9914 \xrightarrow{\cdot 100} \mathbf{99.14} \rightarrow -0.86\%$$

The family spent HUF 120 (13880-14000), i.e. 0.86% (99.14-100) , less on purchasing the two products together in April.

The increase in value is caused by two factors:

- the average aggregate change in prices, expressed by the aggregate price index
- the average aggregate change in the amount consumed, expressed by the aggregate volume index

If we desire to quantify the aggregate change in prices, then we must take the quantity of products unchanged. And the aggregate change in quantity can be quantified by keeping the prices of products unchanged.

The factor kept unchanged in these cases is called *weight*, which may be data either of the base period or of the reference period. The most commonly used hybrid formula for the elimination of differences caused by the different weightings is the Fisher index, which can be represented as the geometric mean of the two different weighted indices.

A base weighted price and volume index is called Laspeyres index, and a reference weighted index is called Paasche index. The geometric mean of the base weighted index and the reference weighted index is the Fisher index.

Aggregate price indices		*Aggregate volume indices*
$I_p^0 = \frac{\sum q_0 \cdot \boldsymbol{p_1}}{\sum q_0 \cdot \boldsymbol{p_0}}$	Base-weighted (Laspeyres)	$I_q^0 = \frac{\sum \boldsymbol{q_1} \cdot p_0}{\sum \boldsymbol{q_0} \cdot p_0}$
$I_p^1 = \frac{\sum q_1 \cdot \boldsymbol{p_1}}{\sum q_1 \cdot \boldsymbol{p_0}}$	Reference-weighted (Paasche)	$I_q^1 = \frac{\sum \boldsymbol{q_1} \cdot p_1}{\sum \boldsymbol{q_0} \cdot p_1}$
$I_p^F = \sqrt{I_p^0 \cdot I_p^1}$	Fisher-average	$I_q^F = \sqrt{I_q^0 \cdot I_q^1}$

Figure 20 Formulae to compute aggregate price and volume indices

The computation involves calculating two fictitious aggregates: in $q_0 \cdot p_1$, we multiply the quantity of an earlier period by the price of the

reference period, while in $q_1 \cdot p_0$ we multiply the quantity of the reference period by the price data of the earlier period.

	q_0	p_0	$v_0 = q_0 \cdot p_0$	q_1	p_1	$v_1 = q_1 \cdot p_1$	$q_0 \cdot p_1$	$q_1 \cdot p_0$
Bread (kg)	20	300	$20 \cdot 300 = \mathbf{6000}$	22	320	$22 \cdot 320 = \mathbf{7040}$	$20 \cdot 320 = \mathbf{6400}$	$22 \cdot 300 = \mathbf{6600}$
Milk (1)	40	200	$40 \cdot 200 = \mathbf{8000}$	38	180	$38 \cdot 180 = \mathbf{6840}$	$40 \cdot 180 = \mathbf{7200}$	$38 \cdot 200 = \mathbf{7600}$
Together			**14 000**			**13 880**	**13 600**	**14 200**

Figure 21 Illustration of computing fictitious aggregates

Find a way to quantify the changes in the amount of bread and milk consumed and their price.

Aggregate change in the amount consumed
(aggregate volume indices)

$$I_q^0 = \frac{\sum \boldsymbol{q_1} \cdot p_0}{\sum \boldsymbol{q_0} \cdot p_0} = \frac{14\,200}{14\,000} = 1.0143 \xrightarrow{\cdot 100} \mathbf{101.43} \rightarrow +1.43\%$$

$$I_q^1 = \frac{\sum \boldsymbol{q_1} \cdot p_1}{\sum \boldsymbol{q_0} \cdot p_1} = \frac{13\,880}{13\,600} = 1.0206 \xrightarrow{\cdot 100} \mathbf{102.06} \rightarrow +2.06\%$$

$$I_q^F = \sqrt{I_q^0 \cdot I_q^1} = \sqrt{1.0143 \cdot 1.0206} = 1.0175 \xrightarrow{\cdot 100} \mathbf{101.75} \rightarrow +1.75\%$$

Across March and April, the aggregate average amount of products consumed increased by 1.43% when weighted by the prices of the base period, by 2.06% when weighted by the prices of the reference period, and by 1.75% when weighted by the prices of both periods.

The aggregate change in unit prices (aggregate price indices)

$$I_p^0 = \frac{\sum q_0 \cdot \boldsymbol{p_1}}{\sum q_0 \cdot \boldsymbol{p_0}} = \frac{13\,600}{14\,000} = 0.9714 \xrightarrow{\cdot 100} \mathbf{97.14} \rightarrow -2.86\%$$

$$I_p^1 = \frac{\sum q_1 \cdot \boldsymbol{p_1}}{\sum q_1 \cdot \boldsymbol{p_0}} = \frac{13\,880}{14\,200} = 0.9775 \xrightarrow{\cdot 100} \mathbf{97.75} \rightarrow -2.25\%$$

$$I_p^F = \sqrt{I_p^0 \cdot I_p^1} = \sqrt{0.9714 \cdot 0.9775} = 0.9745 \xrightarrow{\cdot 100} \mathbf{97.45} \rightarrow -2.55\%$$

Across March and April, on average, the aggregate unit prices of products fell by 2.86% when weighted by the volumes of the base period, by 2.25% when weighted by the volumes of the reference period, and 2.55% when weighted by the volumes of both periods.

We can choose between the two different methods of weighting on the basis of the data that we have. If both indices may be computed, it is a good idea to use the hybrid formula also.

Relationships between aggregate indices

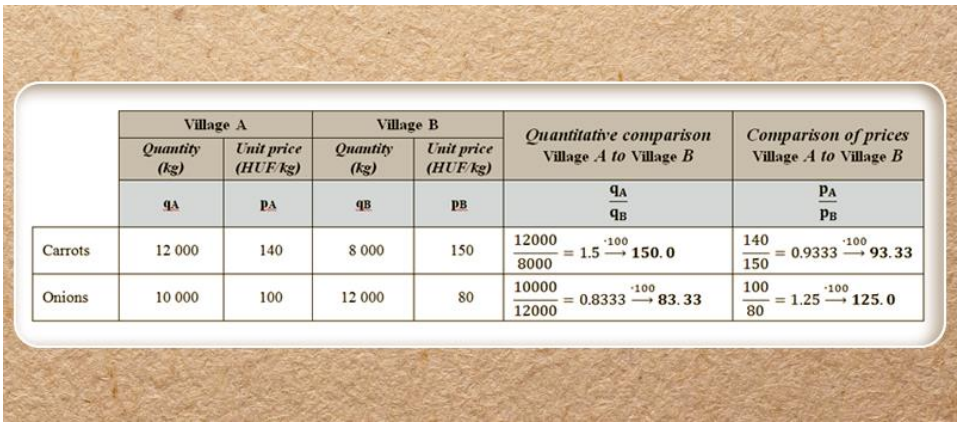
For aggregate indices, the value of the aggregate value index is identical to the product of the inversely weighted or Fisher aggregate volume and price indices.

$$\mathrm{I_v} = \mathrm{I_q^0} \cdot \mathrm{I_p^1} = \mathrm{I_q^1} \cdot \mathrm{I_p^0} = \mathrm{I_q^F} \cdot \mathrm{I_p^F}$$

7.2.2 Areal comparison

In areal comparison, the direction of the comparison may be chosen freely. We indicate the area to compare and the basis of comparison in a subscript index. The value index is not interpreted for areal comparison, but price and volume indices play an important role.

- Agricultural products are supplied to a city's market place from two nearby villages, call them A and B. We know both the amount of products delivered and their unit prices. *Compare the prices of the products and the amount in which they are offered for sale in regard to the two villages.*



	Village A		Village B		*Quantitative comparison* Village *A to* Village *B*	*Comparison of prices* Village *A to* Village *B*
	Quantity (kg)	*Unit price (HUF/kg)*	*Quantity (kg)*	*Unit price (HUF/kg)*		
	q_A	p_A	q_B	p_B	$\frac{q_A}{q_B}$	$\frac{p_A}{p_B}$
Carrots	12 000	140	8 000	150	$\frac{12000}{8000} = 1.5 \overset{\cdot 100}{\longrightarrow} \mathbf{150.0}$	$\frac{140}{150} = 0.9333 \overset{\cdot 100}{\longrightarrow} \mathbf{93.33}$
Onions	10 000	100	12 000	80	$\frac{10000}{12000} = 0.8333 \overset{\cdot 100}{\longrightarrow} \mathbf{83.33}$	$\frac{100}{80} = 1.25 \overset{\cdot 100}{\longrightarrow} \mathbf{125.0}$

Figure 22 Comparison of prices and amounts offered for sale in the two villages

In regard to the two villages, the following conclusions can be drawn on the basis of the computed values in the table above:

- *Village A* supplies 50% more carrots and 16.7% (83.3-100) less onions to the market place
- The price of carrots is 6.67% (93.33-100) lower and the price of onions is 25% higher in *village A*

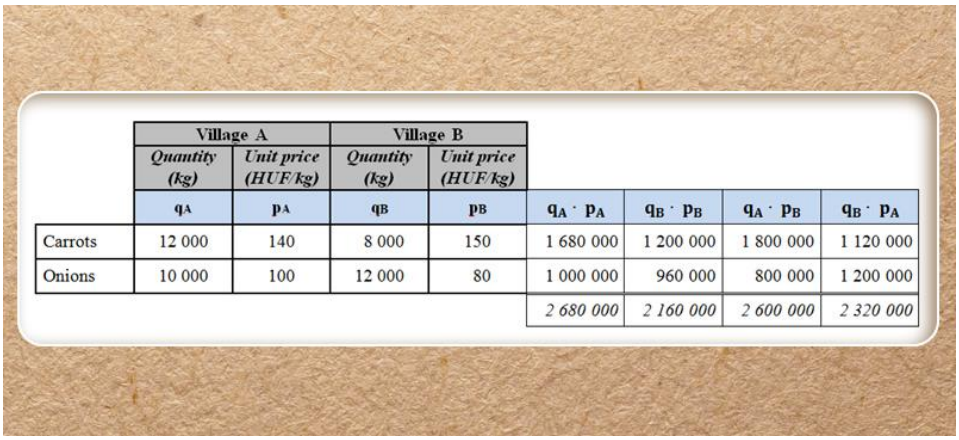
The comparison could be conducted in the opposite direction too (by comparing B to A), but in that case, the relation would be studied from the opposite direction, and so the result would be different in magnitude, due to the change in the basis of comparison.

We can compare not just individual products, but also several different products. This is done by aggregate areal index computation.

We construct the formulae for the computation of the areal price and volume index on the analogy of temporal comparison. It is useful to indicate the direction of the comparison in the index. We can use either the quantities or the prices in villages A and B as weights. The hybrid formula can be used in this case too. It will *suffice to interpret only the Fisher indices* in the study, because they carry the information that really matters.

Let us compare the aggregate amounts and prices of the two products in the two villages by comparing village A to village B.

First, we need to compute values of the aggregates necessary for the indices.



	Village A		Village B					
	Quantity (kg)	*Unit price (HUF/kg)*	*Quantity (kg)*	*Unit price (HUF/kg)*				
	q_A	p_A	q_B	p_B	$q_A \cdot p_A$	$q_B \cdot p_B$	$q_A \cdot p_B$	$q_B \cdot p_A$
Carrots	12 000	140	8 000	150	1 680 000	1 200 000	1 800 000	1 120 000
Onions	10 000	100	12 000	80	1 000 000	960 000	800 000	1 200 000
					2 680 000	*2 160 000*	*2 600 000*	*2 320 000*

Figure 23 Values of aggregates necessary for the computation of areal indices

Comparison of aggregate amounts: areal volume index

$$I^A_{q\left(\frac{A}{B}\right)} = \frac{\sum q_A \cdot p_A}{\sum q_B \cdot p_A} = \frac{2\,680\,000}{2\,320\,000} = 1.1552 \overset{\cdot 100}{\longrightarrow} \mathbf{115.52}$$

$$I^B_{q\left(\frac{A}{B}\right)} = \frac{\sum q_A \cdot p_B}{\sum q_B \cdot p_B} = \frac{2\,600\,000}{2\,160\,000} = 1.2037 \overset{\cdot 100}{\longrightarrow} \mathbf{120.37}$$

$$I^F_{q\left(\frac{A}{B}\right)} = \sqrt{I^A_{q\left(\frac{A}{B}\right)} \cdot I^B_{q\left(\frac{A}{B}\right)}} = \sqrt{1.1552 \cdot 1.2037} = 1.1792 \overset{\cdot 100}{\longrightarrow} \mathbf{117.92}$$

(On the basis of the Fisher index) 17.92% more products can be purchased on average in *village A*.

Comparison of unit prices: areal price index

$$I^A_{p\left(\frac{A}{B}\right)} = \frac{\sum q_A \cdot p_A}{\sum q_A \cdot p_B} = \frac{2\,680\,000}{2\,600\,000} = 1.0308 \overset{\cdot 100}{\longrightarrow} \mathbf{103.08}$$

$$I^B_{p\left(\frac{A}{B}\right)} = \frac{\sum q_B \cdot p_A}{\sum q_B \cdot p_B} = \frac{2\,320\,000}{2\,160\,000} = 1.0741 \overset{\cdot 100}{\longrightarrow} \mathbf{107.41}$$

$$I^F_{p\left(\frac{A}{B}\right)} = \sqrt{I^A_{p\left(\frac{A}{B}\right)} \cdot I^B_{p\left(\frac{A}{B}\right)}} = \sqrt{1.0308 \cdot 1.0741} = 1.0522 \overset{\cdot 100}{\longrightarrow} \mathbf{105.22}$$

The products under investigation can be purchased at a price 5.22% higher on average in *village A* (on the basis of the Fisher index).

7.3 SUMMARY AND QUESTIONS

7.3.1 Summary

The method of value based index computation is employed in various fields in the context of economy. Even ordinary people use this statistical method, though probably implicitly, when they do their shopping in different locations or at different times.

Central to this method of index computation, used for summing data in different units of measurement, is the concept of value, which is defined as the product of price and quantity. We compute elementary indices in the comparison of the prices, quantities, and values of a particular type of product, and, for the representation of aggregate changes in a heterogeneous group of products, we compute aggregate indices. Indices computed on the analogy of ratios can be used for temporal and areal comparisons too.

7.3.2 Self-test questions

How do you define value in statistical terms?

What is the difference between an elementary and an aggregate index?

What does an elementary price index show?

What does an elementary volume index show?

What does an elementary value index show?

How do different elementary indices relate to one another?

What does an aggregate value index show?

Why can you compute the aggregate price and volume index in two different ways?

What do you use the Fisher hybrid formula for, and how is it computed?

How do aggregate indices relate to one another?

7.3.3 Practice tests

☸ Choose the correct answer.

Which of these does not statistically belong in the category of value?

– turnover

- selling price
- sales receipts
- gross production

If the value of the base weighted price index is 107.82, and the value of the chain weighted price index is 109.12, then the value of the Fisher index is:

- 117.65
- 106. 48
- 108.47

The value of the elementary price index is 100, and the value of the elementary volume index is 110.
The value of the elementary value index is:

- 110
- 90.1
- 100

A greengrocer's income from the aggregate sale of all of its products has fallen by 10% compared to last year.

- the elementary value index is 110.0
- **the aggregate value index is 90.0**
- the aggregate volume index is 90.0

In regard to a family's expenditure on food-stuffs, we know that the price index of fruits is 110 and that of vegetables is 90. From this we know that

- **the price of fruits rose by 10%**
- the price of vegetables rose by 10%
- changes in the price of fruits and vegetables balance each other out, the aggregate price index is 100

☸ Decide whether the statements below are true (T) or false (F).

Fisher indices are computed as the geometric means of base and reference weighted indices.	**T**
The elementary price index is the quotient of the elementary volume index and the elementary value index.	**F**
For aggregate indices, the following relationship holds: the aggregate value	**T**

index is the product of the inversely weighted price and volume indices.	
Value is derived as the product of price and quantity.	**T**
In index computation, we compare the data of the earlier period to the data of the later period.	**F**

8. PRACTICAL APPLICATIONS OF INDEX COMPUTATION: MEASURING CORPORATE PERFORMANCE, INDICATORS OF NATIONAL ECONOMY, AND EXTERNAL TRADE STATISTICS

8.1 GOALS AND COMPETENCIES

Indices are commonly employed in economic analyses. Index computation based on standardization is chiefly used in corporate statistics, while the method of value based index computation is employed in accounting for the performance of the national economy and in external trade statistics. The method of index computation based on standardization is an efficient method for the comparison of the performance of companies across periods of time, their specific use of materials, their costs, or the average wages of employees. We use the method of value based index computation to quantify changes in economic performance, price level, and external trade turnover.

The purpose of this unit is to familiarize students with practical applications of index computation. Students will learn about the major areas of the application of indices, and they will acquire the methods of computing the primary indicators of national economy. They will learn to navigate through the world of international statistics, and will acquire the methodological foundations necessary for the assessment of corporate performance.

8.2 TOPICS

Indices commonly occur in micro- as well as macro-level economic analyses. Quantification of relative changes plays an essential role in both corporate practice and at the level of national economy. National or corporate performance may be examined either at the level of particular factors or in an aggregate fashion by the methods of standardization based and value based index computation.

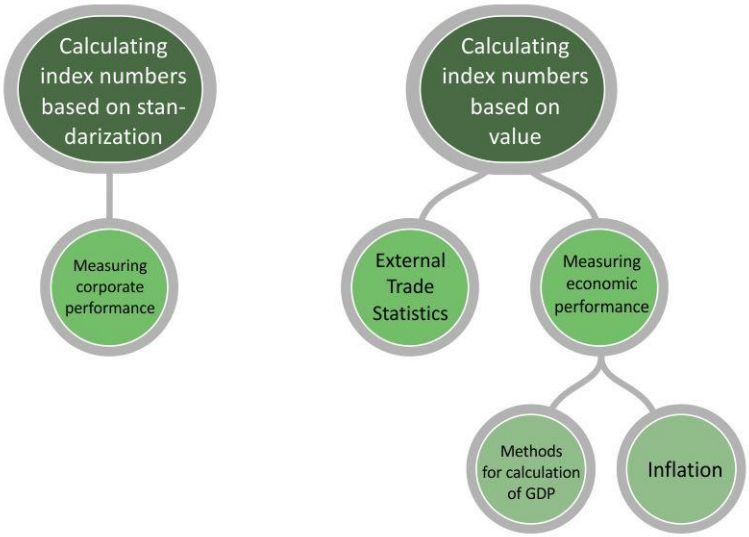


Figure 24 Areas of the practical application of indices

8.2.1 Quantifying corporate performance

Corporate performance may be quantified with both methods of index computation. We employ standardization in the examination of productivity, cost, and the specific use of materials, while we use the method of value, price, and volume index computation in the analysis of a company's sales receipts, turnover, gross production, or sales.

Productivity is the quotient of the total products produced and the number of employees involved in production. Changes in corporate-level productivity – grand mean from the perspective of standardization – are considerably affected by the production of individual units (cluster mean) within the company, i.e., their unit-level productivity. The method of standardization can be used to demonstrate changes in corporate-level productivity over time, and to distinguish between the degree to which changes in unit-level productivity contribute to changes in corporate-level productivity across periods of time and the degree to which differences in the composition of employees are the contributing factor.

Cost, from a statistical perspective,[8] is the quotient of the total expenses and the amount of products produced. It shows the rate of cost per unit of production. The quantification of cost is particularly important

[8] The calculation of cost from an accounting perspective is more complex than this.

in a company which has changed its production technology and production resources, since innovations and modernizations are expected to reduce expenses.

The **specific use of materials** is the quotient of the total materials used and the amount of products produced, which shows the amount of materials that is necessary for the production of a (unit of) product. The direction of change in the specific use of materials is an important indicator for a company, which has a considerable effect on the structure of its expenses. In order for us to be able to employ the method of standardization, a distinction should be made in the specific use of materials at the level of products or product groups – cluster mean –, as changes at this level have an effect on the company as a whole. The composition-effect index represents changes in the structure of production.

In practice, the method of index computation based on standardization can be used by companies for the quantification of changes in average wages, or, in the tourist industry, for the analysis of the average duration of tourists' stay. The same method can be used in agriculture, for the analysis of changes in the average yield, which is derived as the quotient of the crop and the size of the harvested area.

8.2.2 Indices in quantifying the performance of the national economy

We employ the method of value based index computation for measuring the performance of the national economy. It allows us to quantify changes in price level, volume, and value. The two most important indicators of the macroeconomic performance of a country are GDP and inflation.

The most commonly used index for the measurement of the performance of the national economy is the Gross Domestic Product, abbreviated as **GDP.**

GDP is the monetary value of consumer goods and services produced within a nation's borders in a particular year.

GDP is the total value of all products and services produced within the borders of a country, defined as the product of the prices and quantities of products. Depending on the way it is computed, we distinguish between nominal and real GDPs. Nominal, or current-price, GDP is computed by multiplying the quantities in a particular year by the prices in that year. In contrast, real GDP, or constant-price GPD, or Purchasing Power Parity (PPP) GDP, is computed by multiplying the quantities of a

particular year by the prices of previous years, thereby eliminating the effect of changes in prices. With respect to the year whose prices are adopted for the computation of real GDP, the values of the nominal and real GDPs are identical.

We can use indices to quantify changes in macroeconomic performance. A change in nominal GDP can be expressed as a value index, as we compare values to each other, which are defined as products of prices and quantities of the same year. For two different products, call them A and B, and for two different periods, 0 and 1, the value index may be computed as follows:

$$\text{Change of nominal GDP} = \frac{p_A^1 \cdot q_A^1 + p_B^1 \cdot q_B^1}{p_A^0 \cdot q_A^0 + p_B^0 \cdot q_B^0}$$

The Purchasing Power Parity (PPP) GDP is more commonly used in international comparisons, because this index eliminates the effect of inflation, and thus it better reflects changes in performance. When computing real GDP, the same price occurs in the numerator and in the denominator, therefore it only represents aggregate changes in quantity, which is why it is also called the volume index of GDP.

$$\text{Change of real GDP} = \frac{p_A^0 \cdot q_A^1 + p_B^0 \cdot q_B^1}{p_A^0 \cdot q_A^0 + p_B^0 \cdot q_B^0}$$

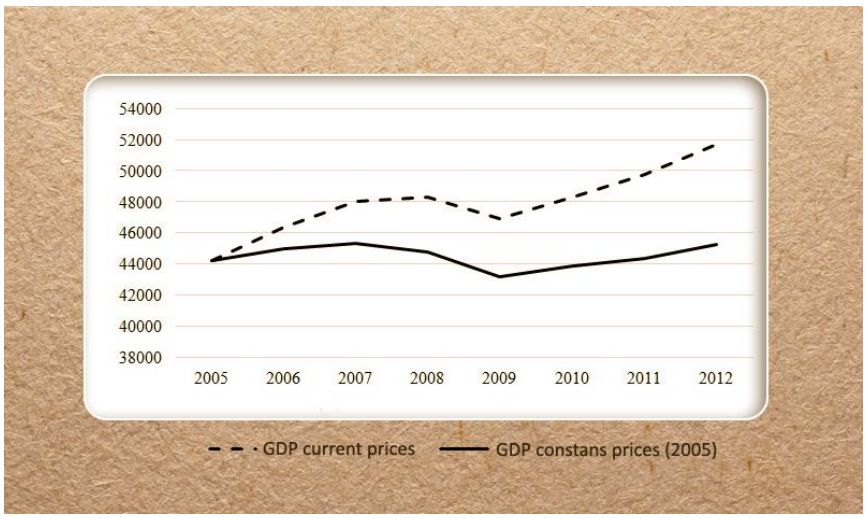


Figure 25 Changes in current price and constant (2005) price US GDP data across 2005 and 2012

Source: *OECD* (2014)

Figure 25, which compares changes in current price GDP with changes in constant price GDP, clearly shows the difference which results from the difference in computation. For the computation of constant price GDP, data from 2005 are taken as constant, therefore the values from the two different computations are identical for that year. Current price values are higher than constant price values, because the former also reflect changes in prices. You can read more about the relationship between the GDP and inflation on the following technical web site.

16. The relationship between GDP and inflation:
    http://www.investopedia.com/articles/06/gdpinflation.asp

8.2.3 Inflation

Inflation is the decrease in the purchasing power of money, a price index that represents the aggregate change in prices, which may be computed in two different ways. The GDP deflator quantifies aggregate price changes by taking the reference year quantities as constant. A more commonly used manner of computing inflation is the Consumer Price Index (CPI), which measures aggregate changes in prices by taking the prices of a previous period as constant. For more details about the computation of the GDP deflator and the consumer price index, listen to the following audio recording.

For two different products, the GDP deflator can be computed thus:

$$\mathrm{GDPD} = \frac{p_A^1 \cdot q_A^1 + p_B^1 \cdot q_B^1}{p_A^0 \cdot q_A^1 + p_B^0 \cdot q_B^1}$$

For two different products, the consumer price index can be computed thus:

$$\mathrm{CPI} = \frac{p_A^1 \cdot q_A^0 + p_B^1 \cdot q_B^0}{p_A^0 \cdot q_A^0 + p_B^0 \cdot q_B^0}$$

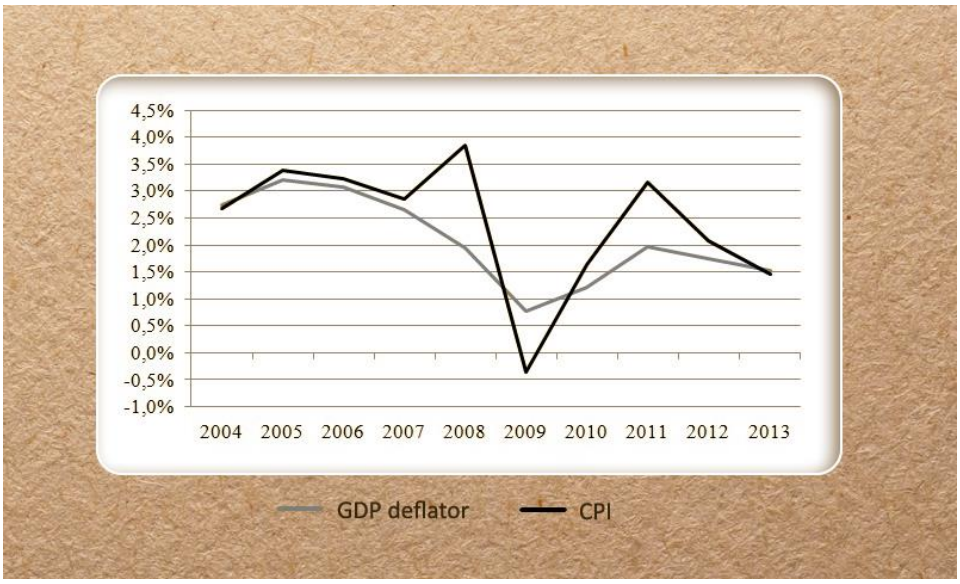


Figure 26 Comparison of GDP deflator and consumer price index for measuring inflation on data on the USA across 2004 and 2013

Source: World Bank (2014)

Figure 26 nicely illustrates the difference between the two different methods of measuring inflation. The consumer price index is much more widely used than the GDP deflator. This is because aggregate changes in prices are better demonstrated when we take the quantities of previous years as constant, since the quantities of the reference year indirectly reflect the effect of price changes across the time periods. The factor of prices and the factor of quantities, used to derive values, are inversely related to each other: rises in prices cause a decrease in quantities. If prices rise across two periods, then the quantities purchased of the products affected are likely to decrease. The effect of changes in prices is measured by the consumer price index by taking the quantities of a period preceding the price changes to be constant. From a statistical perspective, the kind of weighting that the GDP deflator involves is just as correct, but common practice tends to follow the logic of the consumer price index.

Inflation and economic performance are tightly interconnected. Regardless of the method of computation, as rises in prices occur more frequently, the presence of inflation in national economies is more likely. However, we must also consider other processes involved in the aggre-

gate change in prices. You can read more about these on the following technical web site:

17. Inflation, deflation, and stagflation: http://www.investopedia.com/exam-guide/cfp/economics-time-value/cfp6.asp

The computation of aggregate indices is greatly affected by the choice of weights, as we have seen in the case of GDP and inflation. It was in order to avoid such distortive effects of weighting that the Fisher formula was proposed. It eliminates weighting related distortions by determining the geometric mean of differently weighted indices.

8.2.4 External trade statistics

Value based index computation is of particular significance in external trade statistics. We can gain a good picture of the nature and intensity of international relations from quantifying changes in international product turnover, external trade prices and volumes. For more information on international trade and statistics, you can visit the web site of the World Trade Organization (WTO)

18. World Trade Organization: http://www.wto.org/english/res_e/statis_e/statis_e.htm

An important ratio employed in external trade is **terms of trade,** which is the quotient of the price index of exported and imported goods. The index represents the percentage which expresses how much more or how much less exported goods had to be sold in exchange for a unit of import in the accounting period compared to the base period.[9] If the value of the index is greater than 1, then the terms of trade improve, because the volume of export grows, while if it is less than 1, then the terms of trade decrease.

The methodology of turnover and price statistics in the context of external trade is based on value based index computation, though it is still special because of the international aspects it involves. The following **video** is a taped interview conducted with a specialist, who talks about special aspects of international price statistics.

[9] Központi Statisztikai Hivatal [Central Statistical Office] (2014): Áralakulás [Price change] . URL: https://www.ksh.hu/thm/1/indi1_2_3.html. Accessed: August 19, 2014.

8.3 SUMMARY AND QUESTIONS

8.3.1 Summary

Indices are widely used in measurements of the performance of companies and national economies. We employ the method of standardization in the examination of changes, and the factors contributing to changes, in productivity, cost, and the specific use of materials. We can use value based index computation to quantify changes in macroeconomic output and aggregate changes in prices. External trade statistics are also based on these methods.

8.3.2 Self-test questions

In what specific areas can indices be used?

Can you mention some examples of the application of standardization based index computation?

Can you mention some examples of the application of value based index computation?

What is GDP, and what types GDP do we distinguish?

What is purchasing power parity GDP?

How can you measure changes in nominal GDP?

How can you measure changes in real GDP?

What is inflation?

How can you measure inflation?

What is meant by terms of trade?

8.3.3 Practice tests

☸ Choose the correct answer.

Productivity shows

- the number of workers per unit of products
- the number of products per worker
- the average amount of products produced

The value index of GDP is

- the change in nominal GDP
- the change in real GDP
- the quotient of the nominal and real GDP

In computing the consumer price index used to measure inflation, the weight is

- quantities of the reference year
- quantities of the previous year
- prices of the previous year

Terms of trade is

- the quotient of the price indices of exported and imported goods
- the quotient of the priced indices of imported and exported goods
- the quotient of export and import

The specific use of materials is

- the quotient of the total materials used and the products produced
- the product of the total materials used and the products produced
- the quotient of the products produced and the total materials used

Nominal GDP is

- the product of quantities and prices of a particular year
- the product of the prices of a particular year and a quantity typical of previous years
- the product of the quantities of a particular year and the prices typical of previous years

Real GDP is

- the product of the quantities and prices of a particular year
- the product of the prices of a particular year and a quantity typical of previous years
- the product of the quantities of a particular year and the prices typical of previous years

In computing the GDP deflator for the measurement of inflation, the weight is

- quantities of the reference year
- quantities of the previous year
- prices of the reference year

The volume index of GDP is

- the change in nominal GDP
- the change in real GDP
- the quotient of the nominal and real GDP

The index used in external trade price statistics is the ……index

- Fisher
- Pearson
- Yule

9. EXAMINATION OF RELATIONSHIPS BETWEEN ATTRIBUTES I: ASSOCIATION AND ANALYSIS OF VARIANCE

9.1 GOALS AND COMPETENCIES

When conducting statistical analyses, we are often interested not only in a description of a particular state of affairs, but we also want to understand the factors that account for it. That means exploring the relationships among the attributes of a particular phenomenon of interest. Types of relationship studies are distinguished in terms of the types of attributes involved.

First, we shall be concerned with two types of relationships. Association is a relationship that holds between two non-quantitative attributes. Analysis of variance quantifies a relationship between a quantitative and a non-quantitative attribute.

The purpose of this unit is to familiarize students with the basics of the examination of relationships between attributes. In particular, the unit offers a detailed discussion of the methods of association and analysis of variance. Students will learn to recognize which technique of relationship analysis is appropriate for a particular set of attributes, and also to quantify those relationships by the relevant indices. It is essential to the acquisition of methods that students can correctly interpret the results gained.

9.2 TOPICS

Two extreme cases in the examination of relationships are independence and a function-like relationship. In independence, there is no detectable relationship, while in a function-like relationship, you can infer from one attribute to another, which means that a particular variant of an attribute always co-occurs with a variant of another. In between the two extremes, there is a stochastic relationship between attributes.

A probability relationship between two or more attributes is called a stochastic relationship. From a variant of an attribute we probabilistically infer the variant of another attribute.

We use indices to discover possible relationships between attributes. The absolute value of an index will always fall between 0 and 1, where 0 stands for independence and 1 stands for a function-like relationship. When the value of an index falls somewhere between 0 and 1, it is a sign of a stochastic relationship. An index value approximating 0 is a sign of

weak relationship, and a value approximating 1 shows a strong relationship.

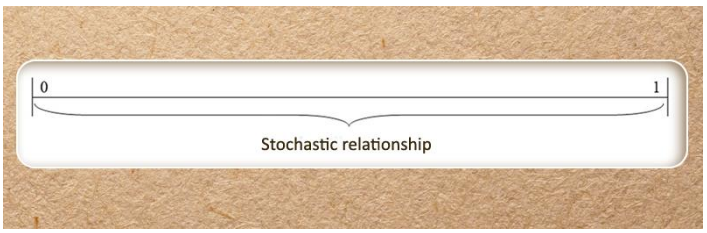


Figure 27 Basic cases in the examination of relationships

Which method we use to discover relationships between attributes is determined by the types of attributes examined.

- ***association***: the relationship between two non-quantitative attributes (qualitative of areal) – e.g., whether or not a student takes an examination successfully (passes or fails) or attending lectures (attended or did not attend)
- ***analysis of variance***: relationship between a non-quantitative (qualitative or areal) and a quantitative attribute – e.g., variants of a test (A or B) and scores achieved

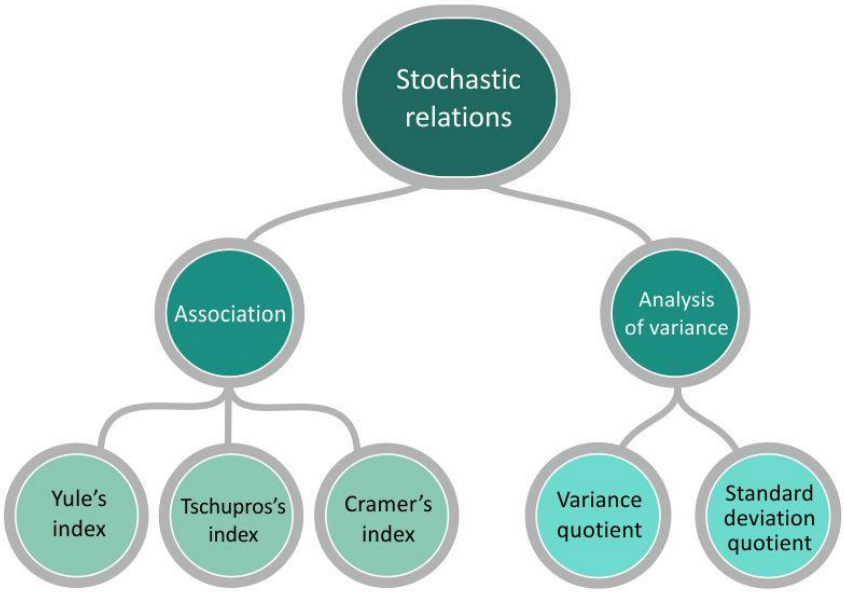


Figure 28 Indices of association and analysis of variance

9.2.1 Association

In examining the relationship between two non-quantitative attributes, we can determine the presence of a relationship and quantify its strength. We can use distribution or coordination ratios, but the most commonly used tools are association coefficients — Yule's, Tschuprov's, and Cramer's.

Yule's index

Yule's index can be computed only for alternative attributes, i.e. attributes with only two variants. The sign of the index can be either positive or negative, which shows the interrelationship between the attribute variants.

$$Y = \frac{f_{11} \cdot f_{00} - f_{10} \cdot f_{01}}{f_{11} \cdot f_{00} + f_{10} \cdot f_{01}}$$

Its properties are: $0 \leq |Y| \leq 1$

If the value of the index is 0, then the two attributes are mutually independent; if it is 1, then the relationship between them is function-like. Any value between the two extremes indicates a stochastic relationship.

- *Use the Yule index to examine the relationship between students' success in an examination and their class attendance habits.*

	Attended lectures	Did not attend lectures	Totals
Successful	50	10	60
Failed	15	25	40
Totals	65	35	100

	Attended lectures (1)	Did not attend lectures (0)	Totals
Successful (1)	f_{11}	f_{10}	$f_{1.}$
Failed (0)	f_{01}	f_{00}	$f_{0.}$
Totals	$f_{.1}$	$f_{.0}$	n

Figure 29 Distribution in students' group in terms of exam success and class attendance

$$Y = \frac{f_{11} \cdot f_{00} - f_{10} \cdot f_{01}}{f_{11} \cdot f_{00} + f_{10} \cdot f_{01}} = \frac{50 \cdot 25 - 10 \cdot 15}{50 \cdot 25 + 10 \cdot 15} = \frac{1250 - 150}{1250 + 150} = \mathbf{0.7857}$$

We discovered a strong relationship between success in an examination and class attendance. The sign of the index is positive, which means that students who had attended the lectures were likely to take their examination successfully, while those who hadn't weren't.

Tschuprov's and Cramer's index

For attributes with several different variants, we use Tschuprov's and Cramer's indices to discover possible relationships between non-quantitative attributes. Assuming that the number of variants of one attribute is s and the number of variants of the other attribute is t, we select them so they meet the condition that s < t. When computing the indices, we assume that if the attributes were independent, then their distribution in the group would be proportional. The number of elements in groups according to particular attributes is called marginal frequency. f* is the value of expected frequency, which is computed by dividing the product of marginal frequencies by the number of elements.

For the calculation of Tschuprov's index, first we need to compute χ^2 (Chi square), for which, for each subpopulation, we square the difference of the value of the observed frequency and the value of the expected frequency, and then we divide the result by the corresponding expected frequency, and then we sum the values for all subpopulations.

$$\chi^2 = \sum \frac{(f - f^*)^2}{f^*}$$

The value of χ^2, the number of elements, and the number of attribute variants are necessary for the computation of Tschuprov's index.

$$\boldsymbol{T} = \sqrt{\frac{\chi^2}{n \cdot \sqrt{s-1} \cdot \sqrt{t-1}}}$$

Cramer's index can be computed either with the help of Tschuprov's index, or by using χ^2, the number of elements, and the number of attribute variants of the attribute that has a smaller number of them.

$$\boldsymbol{C} = \frac{T}{T_{max}} = \frac{T}{\sqrt[4]{\frac{s-1}{t-1}}} \qquad \boldsymbol{C} = \sqrt{\frac{\chi^2}{n \cdot (s-1)}}$$

- A study has been conducted in a higher education institution to determine any relationship between the success rate in a statistics examination and students' active participation in class. Consider the data obtained from this study below.

21. *Distribution of success in statistics exam and students' active participation in class*

Success in statistics exam	Active in class	Less active in class	Did not attend	Totals
Successful	47	30	7	84
Failed	3	25	8	36
Totals	50	55	15	120

Use the appropriate ratio to determine the strength of a stochastic relationship between the success rate in the statistics exam and their active participation in class.

In order to determine the strength of the relationship, we must compute Tschuprov's index and Cramer's index, because class activity is not an alternative attribute. The first step is to determine the values of expected frequencies of attribute variants that belong together, and then we need to compute the value of χ^2.

Success in statistics exam	Active in class		Less active in class		Did not attend		Totals
	f	f*	f	f*	f	f*	
Successful	47	$\frac{84 \cdot 50}{120} = 35$	30	$\frac{84 \cdot 55}{120} = 38.5$	7	$\frac{84 \cdot 15}{120} = 10.5$	84
Failed	3	$\frac{36 \cdot 50}{120} = 15$	25	$\frac{36 \cdot 55}{120} = 16.5$	8	$\frac{84 \cdot 15}{120} = 4.5$	36
Totals	50	50	55	55	15	15	120

Figure 30 Computation of expected frequencies

Success rate in exam and class activity	Observed frequency f	Expected frequency f*	(f – f*)	(f – f*)²	(f – f*)² / f*
Successful – active in class	47	35	12	144	4.11
Successful – less active in cl.	30	38.5	-8.5	72.25	1.88
Successful – did not attend	7	10.5	-3.5	12.25	1.17
Failed – active in class	3	15	-12	144	9.60
Failed – less active in class	25	16.5	8.5	72.25	4.38
Failed – did not attend	8	4.5	3.5	12.25	2.72
Totals	120	120			$\chi^2 = 23.86$

Figure 31 Computation of χ^2

$$T = \sqrt{\frac{\chi^2}{n \cdot \sqrt{s-1} \cdot \sqrt{t-1}}} = \sqrt{\frac{23.86}{120 \cdot \sqrt{2-1} \cdot \sqrt{3-1}}} = \mathbf{0.375}$$

$$C = \frac{T}{T_{max}} = \frac{0.375}{\sqrt[4]{\frac{2-1}{3-1}}} = \frac{0.375}{\sqrt[4]{\frac{1}{2}}} = \frac{0.375}{0.84} = \mathbf{0.4464}$$

There is a *weaker than moderate* relationship between the success rate in the statistics exam and their activity in class.

9.2.2 Analysis of variance

Analysis of variance is a relationship between a quantitative and a non-quantitative (areal or qualitative) attribute. In an examination of this relationship, we can measure the strength of the relationship between the two attributes, and we can also determine the degree in percent to which the non-quantitative attribute contributes to changes in the quantitative attribute, its dispersion.

The first step in the examination of the relationship is to cluster the population according to the non-quantitative attribute. Further, we also need to know the number of elements in the entire population as well as in the clusters, or the magnitude of the distribution of the number of ele-

ments across clusters. For the quantitative attribute, we compute the mean values for the clusters and for the entire population.

- **The mean of a cluster is called a cluster mean, while the mean computed for the entire population is called the grand mean. The grand mean may be defined as the arithmetic mean of cluster means weighted by the number elements in the clusters.**

In our analysis of the relationships, we focus on deviation from the means. The deviation of a value from a mean interpreted over an entire population is the total deviation, which can be decomposed into internal and external deviation. Internal deviation is the deviation of a value from a mean value within its own cluster, while external deviation is the deviation of cluster means from the grand mean. Average deviation from a mean is expressed by standard deviation. Along lines just discussed, we distinguish between total standard deviation, on the one hand, and between internal and external standard deviation, on the other.

For the computation of the indices of such complex relationships, analysis of variance is required, in which we need to determine the values of the internal, external, and total variance.

Deviation		Variance
$S_K = \sum n \cdot (\bar{x} - \bar{X})^2$	EXTERNAL	$\sigma_K^2 = \frac{\sum n \cdot (\bar{x} - \bar{X})^2}{N} = \frac{S_K}{N}$
$S_B = \sum n \cdot \sigma^2$	INTERNAL	$\sigma_B^2 = \frac{\sum n \cdot \sigma^2}{N} = \frac{S_B}{N}$
$S = S_B + S_K$	TOTAL	$\sigma^2 = \sigma_K^2 + \sigma_B^2 = \frac{S}{N}$

Figure 32 Ways to compute internal, external, and total deviation, and variance

Variance quotient – H^2 index

The variance quotient shows to what percentage the non-quantitative attribute determines the dispersion of the quantitative attribute.

$$H^2 = \frac{\sigma_K^2}{\sigma^2} = \frac{\text{external variance}}{\text{total variance}}$$

Standard deviation quotient – H index

The standard deviation quotient shows the strength of a relationship between a quantitative and a non-quantitative attribute. The index is the square root of the variance quotient.

$$H = \sqrt{H^2}$$

Properties of the ratios of the analysis of variance

In stochastic relationships both the value of the variance quotient and the value of the standard deviation quotient fall between 0 and 1. Of the two extreme values, 0 indicates independence, while 1 indicates a function-like relationship.

- In a higher education institution, a study has been conducted on the relationship between the results of a statistics examination test (max. 100 points) and the variants (A and B) of the test the students completed. We know how many students completed the test. The instructor calculated the mean results for each test variant, and also the degree to which students' scores deviated from the mean for each test variant. The results are summarized in the table below.

22. *Information on test variants*

Test variant	Number of students	Mean result (points)	Standard deviation of results (points)
	n_i	$\bar{x}_i$	σ_i
A	80	75	9
B	60	82	6
Total	140		

The following observations can be made on the basis of the table:

- out of a total of 140 students, 80 students completed *variant A* of the test*;* their mean score was 75 points; the average deviation from that was 9 points.
- out of a total of 140 students, 60 students completed *variant B* of the test; their mean score was 82 points; the average deviation from that was 6 points.

Characterize the relationship between the test variants and the test results.

For the characterization of the relationship, first we need to determine the grand mean, which is the arithmetic mean of students' results weighted by the number of students:

$$\overline{X} = \frac{80 \cdot 75 + 60 \cdot 82}{140} = 78$$

In the group of 140 students, examinees scored 78 points on average.

23. *Mean results in the group per test variant and altogether*

Test variant	Number of students	Mean result (points)	Standard deviation of results (points)
	n_i	$\bar{x}_i$	σ_i
A	80	75	9
B	60	82	6
Totals	140	**78**	

The next step in the examination of the relationship is to determine the external, internal, and total variance.

External variance

$$\sigma_K^2 = \frac{\sum n \cdot (\bar{x} - \overline{X})^2}{n} = \frac{80 \cdot (75 - 78)^2 + 60 \cdot (82 - 78)^2}{140} = \mathbf{12}$$

$$\rightarrow \sigma_K = \sqrt{12} = 3.46 \sim 3 \text{ points}$$

Average scores per test variant deviate from the mean result of the entire group by 3 points.

Internal variance

$$\sigma_B^2 = \frac{\sum n \cdot \sigma^2}{n} = \frac{80 \cdot 9^2 + 60 \cdot 6^2}{140} = \mathbf{61.71} \rightarrow \sigma_B = \sqrt{61.71} = 7.86 \sim 8 \text{ points}$$

Students' scores on each test variant deviate from the mean of the test variant by an average of 8 points.

Total variance

$$\sigma^2 = \sigma_K^2 + \sigma_B^2 = 12 + 61.71 = \mathbf{73.71}$$

$$\rightarrow \sigma = \sqrt{73.71} = 8.59 \sim 9 \text{ points}$$

Students' scores deviate from the total mean of the group by an average of 9 points.

Ratios of the analysis of variance

$$H^2 = \frac{\sigma_K^2}{\sigma^2} = \frac{12}{73.71} = 0.1628 \rightarrow \mathbf{16.28\%} \rightarrow H = \sqrt{H^2} = \sqrt{0.16.28} = \mathbf{0.4034}$$

We detected a weak relationship between students' results and test variants. The type of test variant determined students' results to an extent of 16.28%.

9.3 SUMMARY AND QUESTIONS

9.3.1 Summary

When conducting statistical analyses, we are often interested in understanding the relationships between different factors that a study of a particular phenomenon of interest involves. Depending on the type of attributes, stochastic, i.e. probabilistic relationships may be examined in different ways. Association is the method of quantifying the relationship between two non-quantitative attributes. We can determine the presence of a relationship and quantify its strength with the appropriate association coefficients, Yule's, Tschuprov's, and Cramer's. In measuring the complex relationship between a quantitative and a non-quantitative attribute, additional possibilities for analysis open up. In addition to the presence and strength of a relationship, we can also measure the degree to which the non-quantitative attribute determines changes in the quantitative attribute. The standard deviation quotient and the variance quotient deliver more information about the nature of the relationship between the attributes. These methods of examination, which are applicable in the examination of the relationship between two attributes, allow us to study statistical relationships that are connected with non-quantitative attributes.

9.3.2 Self-test questions

What do we mean by a stochastic relationship?

What is meant by the independence of two attributes and by a function-like relationship between them?

Association quantifies a relationship between what kinds of attributes?

Which are the ratios of association?

What is meant by the expected frequency?

Analysis of variance quantifies a relationship between what kinds of attributes?

What is the grand mean, and how is it defined?

Which are the ratios of analysis of variance?

What is shown by the quotient of the variance?

What is quantified by the quotient of standard deviation?

9.3.3 Practice tests

❁ Choose the correct answer.

Which index can you use to quantify the relationship between students' major subjects and their scores on a test?

Yule's Tschuprov's **Standard deviation quotient**

An examination of the relationship between employees' qualifications and wages delivered the following value for the quotient of variance: 0.674. What does it mean?

a. there is a stronger than moderate relationship between the two attributes
b. qualification determines changes in wages to a degree of 67.4%
c. wages by qualification of employees deviate from the grand mean by an average of 0.674

The value of the H index between type of train and number of delays is 0.26. What inference can you draw from that?

a. type of train determines the number of delays to a degree of 0.26%
b. there is a weak relationship between the type of train and the number of delays
c. type of train determines the number of delays to a degree of 26%

The value of the Tschuprov index between the position and gender of employees is 0.28. How strong is the relationship?

Weak Weaker than moderate Strong

Which index can you use to quantify the relationship between students' success rate in an exam (pass or fail) and their place of residence (capital/small town/village)?

Yule **Tschuprov** Standard deviation quotient

An examination of the relationship between employees' qualifications and wages delivered the following value for the Standard deviation quotient: 0.82. What does it mean?

a. there is a strong relationship between the two attributes
b. qualification determines changes in wages to a degree of 82%

c. wages determine qualification to a degree of 82%

The value of the Cramer index between two attributes is 0.64. Which could be the two attributes?

a. visitor's place of origin and residence
b. visitor's place of origin and type of accommodation
c. visitor's place of residence and the amount of money they spent

The value of the quotient of the standard deviation between type of train and number of delays is 0.64. What inference can you draw from this?

a. type of train affects the number of delays to a degree of 40.96%
b. type of train has no effect on the number of delays
c. type of train affects the number of delays to a degree of 64%

If, in a class of students, only those passed the statistics exam who had attended the lectures, then

a. the exam success rate and attendance at lectures are mutually independent
b. the relationship between exam success rate and attendance at lectures can be quantified by the standard deviation quotient
c. there is a function-like relationship between exam success rate and attendance at lectures

An examination of the relationship between employees' qualifications and wages delivered the following value for the H^2 index: 0.325. What does that mean?

a. there is a weak relationship between the two attributes
b. qualification determines changes in wages to a degree of 32.5%
c. qualification affects changes in wages to a degree of 0.325%

10. EXAMINATION OF RELATIONSHIPS BETWEEN ATTRIBUTES II: COMPUTING CORRELATION AND REGRESSION

10.1 GOALS AND COMPETENCIES

In statistical analysis we are often interested in the factors that contribute to changes in a phenomenon of interest. Most social and economic phenomena can be quantified, which allows us to examine the relationships between attributes. In the examination of quantitative attributes, we can not only quantify the presence and strength of a relationship, but we can also determine its direction and reveal cause-and-effect relationships between the factors. Correlation computation can detect the presence of a stochastic relationship between two or more quantitative attributes, as well as its strength and direction. Regression serves to describe the nature of a relationship between attributes in mathematical terms.

The purpose of this unit is to familiarize students with methods of demonstrating and describing relationships between quantitative attributes. Students will learn to be able to identify the areas where correlation and regression computation can be applied and understand the internal structure of the analyses. Students will acquire the knowledge of correlation ratios, and will know what kind of information particular indices offer about the relationship. They will know how to interpret the parameters of a regression function that describes a linear relationship between two attributes, as well as how to select the appropriate form of a regression function that best fits the empirical values.

10.2 TOPICS

Correlation and regression computation is a statistical method for the description of a stochastic relationship between two or more quantitative attributes. The direction of the relationship between attributes may be either positive or negative, which may be described either with a linear or a non-linear regression function. Observations that are involved in such examinations are typically non-exhaustive: all we have is a sample of data composed of a certain number of elements, which we can use to draw general inferences.

What this unit describes in detail is the method of bivariate correlation and regression computation, i.e., the situation in which one variable accounts for changes in another. We use correlation ratios to determine the presence, strength, and direction of the stochastic relationship between the

two variables. It is practical to test the goodness of fit of a regression function, if there is a demonstrated significant relationship between the attributes. In addition to the determination and interpretation of the parameters of a function, the unit also discusses the elasticity coefficient, which quantifies relative changes. We also say a few words about non-linear functions and multivariate linear regression.

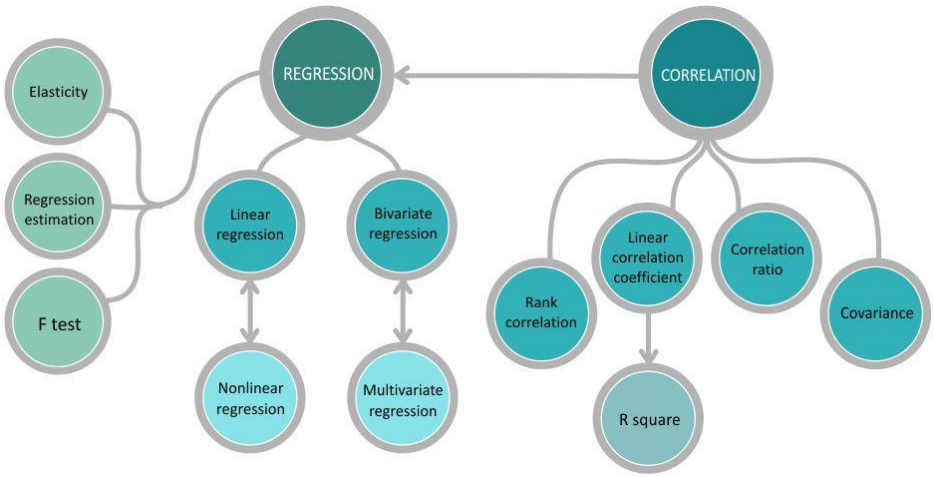


Figure 33 Structure of correlation and regression calculation

10.2.1 Correlation ratios

Correlation is a method for the demonstration of the presence, strength, and direction of a stochastic relationship between two or more quantitative attributes. Of the correlation ratios, covariance is chiefly used to determine the direction of the relationship. Depending on the level of measurement of quantitative attributes, we use different correlation ratios for the description of the strength of the relationship. The Pearson linear correlation coefficient, or its square, the linear coefficient of determination, is used for attributes measurable on a ratio scale. These coefficients, however, can be used confidently only for linear relationships. The strength of the relationship between ordinal scale variables is best demonstrated by ordinal rank correlation ratios.

Covariance (C)

Covariance is primarily used to determine the direction of the relationship between quantitative attributes. The index does not have a lower or upper bound, therefore what really matters is its sign, which, if posi-

tive, shows a relationship of the same direction and if negative, it shows a relationship in opposite directions. A relationship of the same direction means that if the values of one of the attributes rise, so do the values of the other (e.g., if the temperature rises, water consumption will also rise). In a relationship of opposite directions, a rise in the values of one attribute is followed by falls in the values of the other (e.g., the higher the mileage of a second-hand car, the lower its price). If there is no relationship, the value of the index is 0. If the value is different from 0, that is a sign of a relationship between the attributes.

$$C = \frac{\sum xy - n \cdot \bar{x} \cdot \bar{y}}{n-1} = \frac{\sum d_x d_y}{n-1} \text{ or } C = \frac{\sum xy - n \cdot \bar{x} \cdot \bar{y}}{N} = \frac{\sum d_x d_y}{N}$$

The denominator of the index formula will differ depending on whether we have all the data concerning the phenomenon of interest or only a part of it, i.e., a sample. If we are concerned with a complete population, then the denominator will contain the total number of elements (N). If, in contrast, we are examining only a part of the entire population, i.e. a sample, then the denominator contains n-1. In practice, we rarely have data about the entire population. Therefore, in what follows, we will discuss examinations of phenomena conducted on the basis of samples.

Pearson liner correlation coefficient (r)

For attributes measured on a ratio scale, we use the Pearson liner correlation coefficient to determine the strength of the relationship. The index can assume values between -1 and 1, its sign denoting the direction of the relationship, while its absolute value represents the strength of the relationship. If the value of the index is 0, the two variables are mutually independent; if it is -1, it signals a negative deterministic relationship, and if it is +1, it signals a positive deterministic relationship.

If the value of the index approximates 0, it is a sign of weak relationship, and the closer it is to either of the extreme values, the stronger is the relationship.

$$r = \frac{C}{s_x \cdot s_y} = \frac{C}{\sqrt{\frac{\sum d_x^2}{n-1}} \cdot \sqrt{\frac{\sum d_y^2}{n-1}}} = \frac{C}{\sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} \cdot \sqrt{\frac{\sum (y-\bar{y})^2}{n-1}}}$$

We derive the index from covariance, therefore the signs of the two indicators are identical. The denominator contains the product of the corrected empirical (cf. sample) standard deviation of the variables x and y.

Linear coefficient of determination (r^2)

The linear coefficient of determination is the square of the linear correlation coefficient. Its value falls between 0 and 1. It is multiplied by 100 and is interpreted in percentage form. For the interpretation of the index, we need to distinguish between the variables, which will lay the foundation for the determination of a cause-and-effect relationship, the job of regression. One of the variables is x, called the independent, or explanatory, variable, or simply the cause, while y is the dependent variable, or result variable, or simply the effect. The linear coefficient of determination is the number that shows the extent (in percentage) to which the explanatory variable (x) affects the standard deviation of the result variable (y).

Correlation quotient (η)

The correlation quotient represents the changes in the standard deviation of y grouped according to x. The square of the index can also be interpreted (exactly like H and H^2, familiar from the discussion of the analysis of variance). Both indicators can assume a value between 0 and 1: $0 \leq |\eta;\ \eta^2| \leq 1$. For their calculation, we first group the values of the variable X (class interval frequency array), and then we assign the average values of Y to the groups.

$$\eta = \sqrt{\frac{S_K(y)}{S\ (y)}}$$

Rank correlation

Rank correlation can be used to diagnose a relationship between quantitative attributes measured on an ordinal scale. We can use the rank correlation coefficients to determine the degree to which changes in the position on one rank are affected by the position on another rank, that is, the degree to which one array accords with the other.

In market-research questionnaires respondents are often asked to rank different criteria. A typical question is "Why do you choose that particular product?" The answer to this question is given by ranking alternative answers, such as price, brand, advertisement, friends' advice, etc.

In rank correlation measures, we need ranks of the same number of elements, i.e., the sum of ranking numbers must be identical in the two sets of data to compare. In the simplest case, we have clear rankings, where each rank occurs only once, i.e. there is no identity. Calculation becomes more difficult if there are identical rank numbers, or if we want to examine relationships between more than two rankings. Of the rank corre-

lation coefficients, *Spearman's ρ (rho)* coefficient is suitable for the examination of relationships between two rankings, while *Kendall's τ (tau)* coefficient can be used for the examination of more than two rankings.

Spearman's rank correlation coefficient

Spearman's ρ (rho) coefficient is the most commonly used ratio in measuring rank correlation, the value of which can fall between -1 and 1, where -1 means that the ranking is completely reversed in the two cases, while +1 means that they are completely identical. If the value of the coefficient is 0, there is no relationship.

$$\rho = 1 - \frac{6 \cdot \sum d^2}{n \cdot (n^2 - 1)} = 1 - \frac{6 \cdot \sum d^2}{n^3 - n}$$

In this formula, n is the number of observed elements, and d is the difference between the positions occupied on the two rankings.

❁ Ten applications have been submitted to a thesis tender, which were evaluated by two referees, each. The two referees arranged the 10 theses in the following rankings.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Referee A	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
Referee B	2.	3.	1.	5.	6.	9.	4.	8.	10.	7.

Determine the degree to which the rankings of the two referees accord with each other.

For the computation of rank correlation, we need to know the differences[10] between the positions occupied on the two rankings and the sum of their squares.

24. *Work table for calculation of sum of squares required for rank correlation*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	
Referee A	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	
Referee B	2.	3.	1.	5.	6.	9.	4.	8.	10.	7.	
difference (d)	-1	-1	2	-1	-1	-3	3	0	-1	3	$\sum$
square of difference (d^2)	1	1	4	1	1	9	9	0	1	9	**36**

[10] The direction of the subtraction is arbitrary, because we need the sum of squares for th e computation of rank correlation.

The relationship between the evaluations of the two referees:

$$\rho = 1 - \frac{6 \cdot \sum d^2}{n^3 - n} = 1 - \frac{6 \cdot 36}{10^3 - 10} = 1 - \frac{216}{990} = \mathbf{0.7818}$$

The evaluations of the two referees appear to be very similar.

In some cases, ranks "tie", because the ranks for two values are identical. In such cases we assign to both values the simple arithmetic mean of the two consecutive ranking values, which we would get, if the two values were not identical. Such rank numbers are called tied ranks. For example, if the 4th and 5th elements on a ranking coincide, then the 4th would occur twice, so, instead, both are assigned the rank 4.5. Thus, a tied rank is the simple arithmetic mean of two adjacent ranks. In such a case, the formula is modified as follows:

$$\rho = \frac{\frac{1}{6}(n^3 - n) - \left(T_x + T_y\right) - \sum d^2}{\sqrt{\left[\frac{1}{6}(n^3 - n) - 2T_x\right] \cdot \left[\frac{1}{6}(n^3 - n) - 2T_y\right]}}$$

T $= \sum \frac{1}{12}(t_j^3 - t_j)$, t is the number of tied ranks, j is the group with the same rank number

We determine the number of tied ranks by considering the number of cases with identical rank numbers. So, e.g., if we have two items ranked 4th, then both are assigned the rank number 4.5, but we count only the element which otherwise ranked 5th as tied. If there are three elements with the same rank number in the ranking, then two may be considered tied. The number of groups with the same rank number is the number of groups where correction is necessary.

- Ten applications have been submitted to a thesis tender, which were evaluated by two referees, each. The two referees arranged the 10 theses in the following rankings.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Referee A	1.	1.	3.	4.	5.	5.	7.	8.	9.	10.
Referee B	3.	1.	1.	5.	6.	7.	4.	7.	10.	7.

Determine the degree to which the rankings of the two referees accord with each other.

It is obvious that rankings coincide in several cases. Therefore, values with identical ranks are assigned tied ranks, and we will determine the sum of squares of the differences on such a basis.

25. *Work table for computing sums of squares required for rank correlation*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	
Referee A	1.5.	1.5.	3.	4.	5.5.	5.5.	7.	8.	9.	10.	
Referee B	3.	1.5.	1.5.	5.	6.	8.	4.	8.	10.	8.	
difference (d)	-1.5	0	1.5	-1	-0.5	-2.5	3	0	-1	2	Σ
square of difference (d^2)	2.25	0	2.25	1	0.25	6.25	9	0	1	4	**26**

$$T_A = \frac{1}{12}[(2^3 - 2) + (2^3 - 2)] = 1 \qquad T_B = \frac{1}{12}[(3^3 - 3) + (2^3 - 2)] = 2.5$$

Spearman's rank correlation coefficient:

$$\rho = \frac{\frac{1}{6}(10^3 - 10) - (1 + 2.5) - 26}{\sqrt{\left[\frac{1}{6}(10^3 - 10) - 2 \cdot 1\right] \cdot \left[\frac{1}{6}(10^3 - 10) - 2 \cdot 2.5\right]}} = \frac{165 - 3.5 - 26}{\sqrt{163 \cdot 160}} = \frac{135.5}{161.493}$$

$$\rho = \mathbf{0,8391}$$

The evaluations of the two referees appear to be very similar in this case as well.

Kendall's rank correlation method

Kendall's rank correlation method can be used even in cases that involve several variables. In terms of the example just discussed, with this method we can measure relationships between the evaluations of referees and compare how they accord with each other if we have *m* number of referees reviewing *n* number of theses. The point of the method is that it places emphasis not on the differences but on sums of rank numbers, i.e., it sums the ranks of particular elements. The value of the index varies between 0 and 1, where 0 corresponds to complete disagreement and 1 expresses complete agreement.

❁ Ten applications have been submitted to a thesis tender, which were evaluated by four referees, each. The four referees arranged the 10 theses in the following rankings.

Theses	Referee A	Referee B	Referee C	Referee D
	Referees' rankings			
(1)	1.	2.	3.	4.
(2)	2.	3.	1.	1.
(3)	3.	1.	2.	3.
(4)	4.	5.	4.	5.

(5)	5.	6.	6.	7.
(6)	6.	4.	7.	2.
(7)	7.	9.	5.	6.
(8)	8.	8.	10.	9.
(9)	9.	7.	8.	10.
(10)	10.	10.	9.	8.

Determine the degree to which the rankings of the two referees accord with one another.

26. *Work table for using Kendall's rank method*

Theses	**Ref A**	**Ref B**	**Ref C**	**Ref D**	***Sum of ran numbers (R)***	***Deviation from mean***	***Square of deviation from mean***
	Referees' rankings						
(1)	1.	2.	3.	4.	10	-12	144
(2)	2.	3.	1.	1.	7	-15	225
(3)	3.	1.	2.	3.	9	-13	169
(4)	4.	5.	4.	5.	18	-4	16
(5)	5.	6.	6.	7.	24	2	4
(6)	6.	4.	7.	2.	19	-3	9
(7)	7.	9.	5.	6.	27	5	25
(8)	8.	8.	10.	9.	35	13	169
(9)	9.	7.	8.	10.	34	12	144
(10)	10.	10.	9.	8.	37	15	225
					220	*Square of squares of deviation*	**1130**
				MEAN	**22**		

In carrying out the calculation, we first need to sum the rank numbers of each of the theses, i.e., those ranking values which were assigned to the theses by the referees. This is going to be the sum of rank numbers (R). If we sum the sums of rank numbers and divide it by the number of theses, we get the average sum of rank numbers for each thesis ($\sum R/n$ = 220/10). Then we compute the deviation of each thesis's sum of rank numbers from the average, and then, in order to eliminate the difference between positive and negative numbers, we square these deviations too. The sum of squares of deviations at the end is going to be the *d* value in Kendall's index, where d_{max} represents complete agreement among ref-

erees, which we can determine by considering the number of referees and the number of theses.

$$\boldsymbol{W} = \frac{\boldsymbol{d}}{\boldsymbol{d_{max}}} \leftarrow d_{max} = \frac{m^2(n^3 - n)}{12} = \frac{4^2 \cdot (10^3 - 10)}{12} = 1320$$

$$W = \frac{1130}{1320} = \mathbf{0.8561}$$

As the result shows, referees' opinions are very similar, i.e., they seem to agree about the rank of particular theses.

10.2.2 Bivariate linear regression

The regression function is used for the mathematical characterization of the relationship revealed by correlation. Since we are talking about quantitative attributes, we can statistically determine not only the presence of a relationship between attributes, and, if there *is* a relationship, its strength and direction, but we can also measure the order of magnitude of a change of the result variable caused by a change in the explanatory variable. Thus, in our study of the phenomenon of interest, we assume a cause-and-effect relationship between the selected variables, and use the regression model to verify this assumption. The linear regression function which applies to the entire population can be formulated thus:

$$\mathbf{Y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} \cdot \mathbf{X} + \boldsymbol{\varepsilon}$$

In the equation of the function, Y is the value of the result variable, X is the explanatory variable, and, ε represents the random effect. We need parameters β_0 and β_1 for the characterization of the relationship between the variables. Typically, however, we do not have information about the entire population, but instead, we can only use a sample and draw inferences from that. The empirical regression function, which applies to a sample of the entire population, can be formulated thus:

$$\hat{\mathbf{y}} = \mathbf{b_0} + \mathbf{b_1} \cdot \mathbf{x}$$

The sample regression offers an opportunity for us to apply the relationship between the variables revealed in the sample not only to the sample, but to the entire population. For this, the model needs to be tested (F test). With the help of the relationship revealed in the sample, we can take parameters b_0 and b_1 and estimate from them the interval into which the parameters β_0 and β_1 of the population regression function will belong with a particular degree of probability.

Determination and interpretation of the parameters of the empirical (sample) regression function

The parameters of the bivariate linear regression function can be determined in two different ways. By using what are called normal equations, we need to solve a two-variable system of equations which yields the values for b_0 and b_1.

$$\sum y = b_0 \cdot n + b_1 \cdot \sum x$$
$$\sum xy = b_0 \cdot \sum x + b_1 \cdot \sum x^2$$

The parameters may be determined directly by formulas:

$$b_1 = \frac{\sum d_x d_y}{\sum d_x^2} \qquad b_0 = \bar{y} - b_1 \bar{x}$$

The $\mathbf{b_0}$ parameter is the value which the function assumes at place x=0, the number which shows what the value of the result variable (y) is if the value of the explanatory variable (x) is 0. The interpretability of the $\mathbf{b_0}$ parameter depends in part on whether the value of the explanatory variable can be 0, i.e., whether the place x=0 makes practical sense. Furthermore, the interpretability of the parameter also depends on whether the computed value is an element of the domain of the result variable, i.e., whether the dependent variable can assume the computed value in practice. The criteria of interpretability need to be examined in every case, because the mathematically computable parameter does not always have a practical interpretation. The $\mathbf{b_1}$ parameter is always interpretable; it shows the direction and the average magnitude of the change in the result variable as a consequence of increasing the value of the explanatory variable by a unit.

Elasticity coefficient

The elasticity coefficient shows how many times the relative change of Y is of the relative change of X, i.e., what percentage of change is caused in the result variable Y by a 1% change in the explanatory variable X at an arbitrarily chosen x_0 point.

$$\boldsymbol{E\,(y, x = x_0) = b_1 \cdot \frac{x_0}{b_0 + b_1 \cdot x_0}}$$

- An instructor is interested in the possible relationship between students' achievements on a statistics test and the amount of time they spent studying for it.

  Watch the **video** about the situation and data collection.

We have the following random sample composed of 14 elements for analysis:

Preparation time (hours)	Test scores (max. 100)
26	84
12	24
42	96
10	45
6	6
31	78
19	64

Preparation time (hours)	Test scores (max. 100)
24	74
28	90
2	2
40	99
12	14
16	36
8	4

The first step in the examination of the relationship is to consider the type of attributes, which decides about the type of stochastic relationship whose ratios we can use. In this particular example, both the preparation time and the scores achieved are quantitative attributes, so we can use correlation calculation to determine whether there is a relationship, how strong the relationship is, and what the direction of the relationship between the attributes is. When we have found out about the information just indicated, we take a decision on the use of a regression function. If the relationship revealed is weak, then the use of the regression model is questionable.

For quantitative attributes, the pairs of corresponding values can be displayed on a scatter plot, which graphically represents the relationship between them.

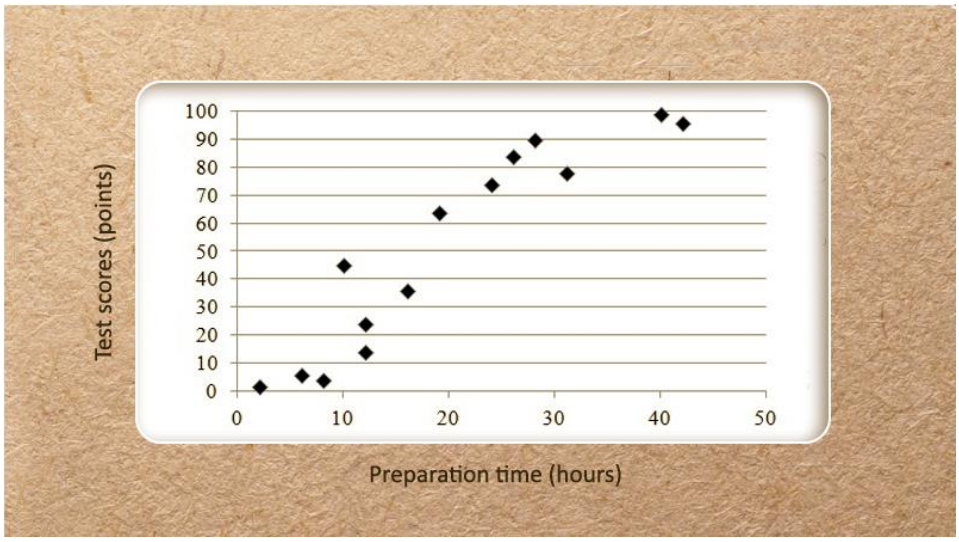


Figure 34 Relationship between preparation time for statistics exam and test scores

- In the case of linear regression, we look for the line that best fits the points which represent the observations. The most commonly used method of representing the best-fitting linear regression function is the ordinary least squared method (OLS). In this case we minimize the deviation between the real values and the values estimated from the regression function, which is to say that $\sum(y - \hat{y})^2 \rightarrow min$.

The diagram gives us a picture about the relationship between the attributes. For the interpretation of the linear coefficient of determination (r^2) and for the regression model, we need to determine which is going to be the explanatory variable and which is going to be the result variable. In this particular case, commonsense dictates that we assume that preparation time (x) explains changes in the test scores (y).

For substitution into the formulae, we need intermediate calculation results. We need to determine the mean of attribute values, as well as the deviation of particular values from that, their squares, and also the pair-by-pair product of the deviations computed for the particular variables.

27. *Work table for determining parameters of regression function*

	Preparation time x	**Test scores y**	$d_x = x - \bar{x}$	$d_y = y - \bar{y}$	$d_x d_y$	d_x^2	d_y^2
	26	84	6.286	32.857	206.531	39.510	1079.592
	12	24	-7.714	-27.143	209.388	59.510	736.735
	42	96	22.286	44.857	999.673	496.653	2012.163
	10	45	-9.714	-6.143	59.673	94.367	37.735
	6	6	-13.714	-45.143	619.102	188.082	2037.878
	31	78	11.286	26.857	303.102	127.367	721.306
	19	64	-0.714	12.857	-9.184	0.510	165.306
	24	74	4.286	22.857	97.959	18.367	522.449
	28	90	8.286	38.857	321.959	68.653	1509.878
	2	2	-17.714	-49.143	870.531	313.796	2415.020
	40	99	20.286	47.857	970.816	411.510	2290.306
	12	14	-7.714	-37.143	286.531	59.510	1379.592
	16	36	-3.714	-15.143	56.245	13.796	229.306
	8	4	-11.714	-47.143	552.245	137.224	2222.449
Σ	***276***	***716***	*0.000*	*0.000*	**5544.571**	**2028.857**	**17359.714**
	$\bar{x} = 19.714$	$\bar{y} = 51.143$					

The intermediate results which we need for further calculations:

$$\bar{x} = \frac{\sum x}{n} = \frac{276}{14} = 19.714 \qquad \bar{y} = \frac{\sum y}{n} = \frac{716}{14} = 51.143$$

$$\sum d_x^2 = 2028.857 \qquad \sum d_y^2 = 17359.714 \qquad \sum d_x d_y = 5544.571$$

Use correlation ratios to examine the relationship between students' test scores on a statistics test and their preparation time.

COVARIANCE

$$\mathbf{C} = \frac{\sum \mathrm{d_x d_y}}{\mathrm{n} - 1} = \frac{5544.571}{14 - 1} = \mathbf{426.51}$$

A *POSITIVE relationship has been found* between students' preparation time and their test scores. This means that *the longer a student spends studying for an examination, the higher will be her score* (change in the same direction).

PEARSON'S LINEAR CORRELATION COEFFICIENT

$$\boldsymbol{r} = \frac{C}{s_x \cdot s_y} = \frac{C}{\sqrt{\frac{\sum d_x^2}{n-1}} \cdot \sqrt{\frac{\sum d_y^2}{n-1}}}$$

$$\boldsymbol{r} = \frac{426.51}{\sqrt{\frac{2028.857}{14-1}} \sqrt{\frac{17359.714}{14-1}}} = \frac{426.51}{12.49 \cdot 36.54} = \mathbf{0.9345}$$

A *STRONG POSITIVE relationship has been found* between students' preparation time and their test scores. This means that *the longer a student spends studying for an examination, the higher will be her score*.

LINEAR COEFFICIENT OF DETERMINATION

$$\boldsymbol{r^2} = 0.9345^2 = \mathbf{0.8733} \overset{\cdot 100}{\longrightarrow} \mathbf{87.33}\%$$

Students' preparation time for the statistics examination (x) *has an effect of 87.33% on their test scores (y).*

Assuming a linear connection, determine the regression function that describes the relationship and interpret its parameters.

The simplest way to determine the values of the parameters is by a formula:

$$\boldsymbol{b_1} = \frac{\sum d_x d_y}{\sum d_x^2} = \frac{5544.571}{2028.857} = \mathbf{2.7329}$$

$$\boldsymbol{b_0} = \bar{y} - b_1 \cdot \bar{x} = 51.143 - 2.7329 \cdot 19.714 = \mathbf{-2.7334}$$

The equation of the bivariate linear regression function is this:

$$\hat{y} = \mathbf{-2.7334} + \mathbf{2.7329} \cdot \boldsymbol{x}$$

The b_0 parameter of the regression function is uninterpretable, because the value delivered for that parameter is not an element of the domain of *y,* which is to say that the result variable, which in this case is the test score, cannot be negative. The b_1 parameter allows us to infer that if a student spends an hour longer studying for the examination, their test score will raise by an average of 2.73 points.

The following linear regression line appears to best fit the observed values:

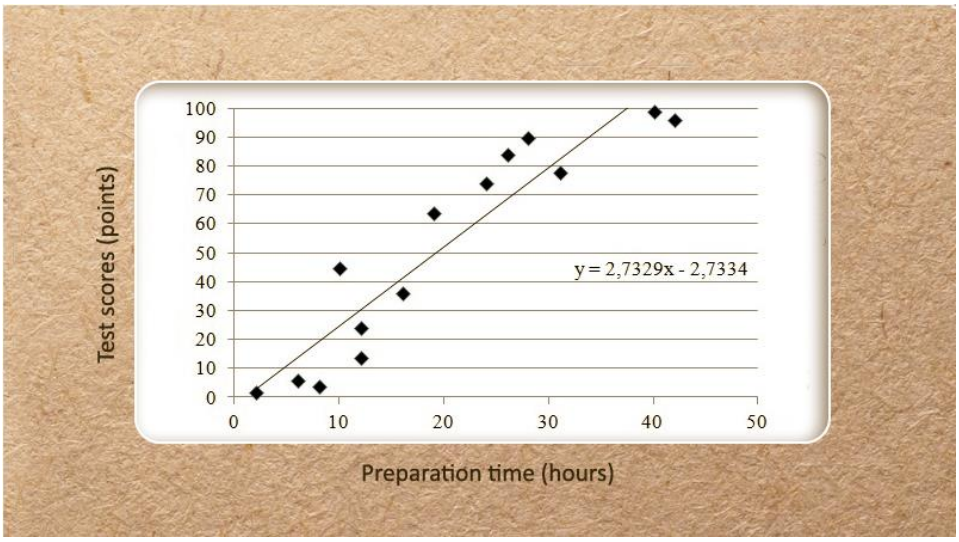


Figure 35 Linear regression function describing the relationship between preparation time and examination test scores

Determine the elasticity of test scores assuming an average amount of preparation time and assuming 10 hours preparation.

ELASTICITY COEFFICIENT

For average preparation time: $x_0 = 19.7$

$$E\,(y, x = x_0) = b_1 \cdot \frac{x_0}{b_0 + b_1 \cdot x_0}$$

$$E\,(y.x = 19.7) = 2.73 \cdot \frac{19.7}{(-2.68 + 2.73 \cdot 19.7)} = \mathbf{1.053}$$

If a student spends 19.7 hours studying for an exam, and they increase that by 1%, then their test score will raise by an average of 1.053 %.

For 10 hours preparation time: $x_0 = 10$

$$E\,(y, x = x_0) = b_1 \cdot \frac{x_0}{b_0 + b_1 \cdot x_0}$$

$$E\,(y.x = 10) = 2.73 \cdot \frac{10}{(-2.68 + 2.73 \cdot 10)} = \mathbf{1.109}$$

If a student spends 10 hours studying for an exam, and they increase that by 1%, then their test score will raise by an average of 1.109%.

The following **animation** shows how the data can be analyzed in MS Excel:

Regression analysis estimates and testing the model

The regression function in a sample allows us to make estimates about the entire population. Some of the estimates refer to unknown empirical values, while the rest approximate to the parameters of the regression function in the entire population.

- One method of drawing inferences from the sample is to make estimates. On the basis of the sample available in connection with the phenomenon of interest, we approximate to some population parameter under conditions determined by an estimation function. The concrete value of a particular parameter can be determined, too (point estimation), but we most commonly designate the interval (confidence interval) which will, with some likelihood (π), contain the value of the parameter (interval estimation). It is in the nature of this method (inference) that you must reckon with the possibility of errors.

Estimating a single value from a regression function means that for a particular value of x (x_0), you can use the regression function in a sample to estimate the interval which will, with some likelihood (π), contain the value of the result variable. In *mean value estimation*, for a particular value of x (x_0), we estimate the interval which will, with some likelihood (π), contain the mean value of the result variable. Particular estimation formulae can be used to approximate to the β_0 and β_1 parameters of the regression model in the entire population, on the basis of the b_0 and b_1 parameters of the regression in the sample. Estimates are sensitive to the number of elements in the sample; the larger the sample, the better the estimate.

In regard to the relationships between attributes, for us to be able to draw general inferences, we need to validate the regression model. The point in the validation is to verify that the relationship between the attributes holds not just for the values in the sample. Parameters b_1 and β_1 of the regression function are of particular importance in this regard, since it is those parameters that show how the explanatory variable affects changes in the result variable. In Hypothesis testing, we decide whether the β_1 parameter of the population regression function deviates from zero. Note that if the value of the β_1 parameter is zero, then the explanatory variable will be eliminated from the equation, since its coefficient is zero, the function is constant. This, in turn, means that changes

in the result variable are independent of the selected explanatory variable, in other words, there is no cause-and-effect relationship.

- Hypothesis testing is the verification of a statement concerning a parameter or some other characteristic of a population carried out by means of the data in our sample. The method essentially consists in constructing two mutually exclusive hypotheses connected to the basic claim. The null hypothesis always involves equality, while the alternative hypothesis involves an event that applies to the entire system of events from the perspective of the basic claim. We make a decision about the statement made in the null hypothesis with the help of a parameter-dependent statistical test, whether to accept or reject it, and our decision about the basic claim will be based on that.

10.2.3 Nonlinear regression

A straight line is not always the best fit for observed values, but it is common practice to use nonlinear forms that can be reduced to a line. In the case of exponential and power regression functions we can get a linear form by logarithmic transformation, whence they are sometimes also called nonlinear regression functions reducible to linear functions.

The **exponential regression** function is used when the increase in a particular phenomenon depends on the state the phenomenon has already achieved. The population regression function can be represented thus:

$\boldsymbol{Y = \alpha \cdot \beta^{x} \cdot \varepsilon}$ or, in the notation already used, $Y = \beta_0 \cdot \beta_1^{x} \cdot \varepsilon$

The empirical regression function, in analogy, is: $\hat{\mathbf{y}} = \mathbf{a} \cdot \mathbf{b}^{\mathbf{x}}$. From this, we derive the following function by logarithmic transformation: $\lg \hat{y} = \lg a + \lg b \cdot x$, where let $lg\ a = b_0$ és $lg\ b = b_1$, and so we can continue the work with a linear function.

Fit an exponential regression function to the data set above.

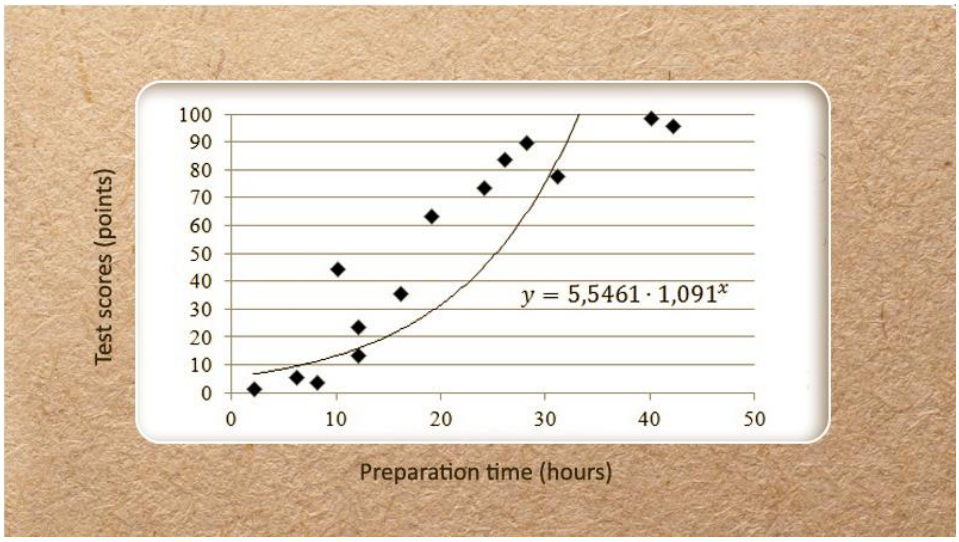


Figure 36 Exponential regression function describing relationship between preparation time and statistics test scores

Of the parameters of the exponential regression function, the value of *a* is rarely interpreted, as it is the value which the function assumes in place x=0, if interpretable. The *b* (β_1) parameter shows whether a unit of increase in the explanatory variable increases or decreases the value of the result variable.

In regard to preparation time and test scores, the value of parameter b is 1.091, which means that *students who spent 1% longer studying for the exam score higher by an average of 9.1 %.*

In the case of **power regression** we also get a linear function after logarithmic transformation. The population regression function can be represented thus: $\boldsymbol{Y = \alpha \cdot x^{\beta} \cdot \varepsilon}$
The form of the empirical regression function is: $\hat{y} = a \cdot x^b$ From this, we derive the following function by logarithmic transformation: $\lg \hat{y} = \lg a + b \cdot \lg x$, where let $lg\ a = b_0$ és $b = b_1$

Fit a power regression function to the data set above.

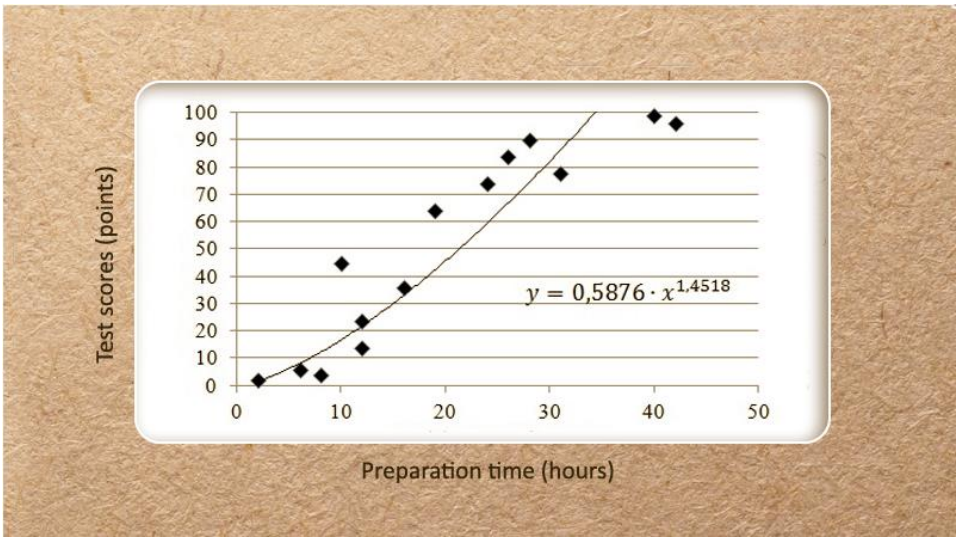


Figure 37 Power regression function describing relationship between preparation time and statistics test scores

Elasticity is identical at all points of the power regression function, same as the power. Thus, the b parameter may be interpreted as the elasticity coefficient, which shows the amount of percentage by which the value of Y is larger or smaller at a 1% increase in the value of X. The *a parameter* is rarely interpreted in practice, a value assumed in place x=1, if interpretable.

In regard to preparation time and test scores, the value of parameter b is 1.4518, which means that *students who spent 1% longer studying for the exam may expect to score by 1.4518 %.*

10.2.4 Examining goodness of fit

Although any regression function may be fitted to the empirical values, we want to choose the one that fits best. The most common choice is the linear form, as this is the easiest to work with mathematically. However, estimates from regression will be most accurate when the form that fits best is selected.

The regression function that fits best is the one in which the deviation between the empirical values and the values estimated from the regression function (residuum) is the smallest. The residual standard deviation essentially shows the average deviation of residuals (e) from the mean, which can be expressed in the unit of measurement of the Y data set. The smaller the value of the residual standard deviation, the smaller the devia-

tion between the empirical and estimated values, i.e., the better the model is.

$$s_e = \sqrt{\frac{\sum \mathbf{e^2}}{\boldsymbol{n-2}}}$$

In examining goodness of fit, we determine the value of the residual standard deviation. The regression that will fit the empirical values best is the one in which that value is the smallest. In practical work other indicators may also be considered in determining goodness of fit, such as the value of r^2, which is highest when the function best fits the empirical values.

10.2.5 Multivariate regression

As it is rare in practice that changes in a variable are due to one factor only, we mostly work with multivariate models. The multivariate linear regression function may be represented thus:

$$\boldsymbol{Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 \ldots \beta_n \cdot X_n + \varepsilon}$$

As in the case of the bivariate form, we have empirical or sample regression here too:

$$\boldsymbol{\hat{y} = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_n \cdot x_n}$$

For the determination of the parameters of the standard linear regression model, four conditions must be met. Linearity means that changes in the result variable are due to changes in the explanatory variable, i.e., the expected values of the former are the consequence of linear relationships of the latter. The criterion of independence means that the explanatory variables must be independent of one another, i.e., they account for the changes in the result variable but have no effect on one another. The homoscedasticity requirement (homogeneity of variance), which applies to the variables is also met, as well as the condition of normal distribution.

The interpretation of parameters is even more important in a multivariate model. Parameter b_0 is the value which the function assumes in place $x_1 = x_2 = \ldots = x_n = 0$, which is interpretable if the values of the explanatory variables may be 0, and if the value computed for this parameter is an element of the domain of y. Coefficients of the explanatory variables are interpreted *ceteris paribus*,[11] i.e., the values of the rest of the explanatory variables are taken as constant. Thus, for example, the b_1 variable shows by how many units on average the value of y increases or decreases, in accord with the specified direction of the relationship, as a consequence

[11] An expression often used in economics, which means “all other things being equal.”

of a one-unit increase in the value of the explanatory variable x_1, such that the values of all the other explanatory variables are taken unchanged.

- The mutual independence of the explanatory variables from one another is critical in multivariate models. If there is a high correlation among explanatory variables, the regression model is not going to be adequate.[12] This is called multicolinearity, which we must filter out from the model by recruiting variables that affect the result variable but not each other.

10.3 SUMMARY AND QUESTIONS

10.3.1 Summary

The examination of the relationship between quantitative attributes is of particular significance in practice. In correlation computation, we reveal the presence, strength and direction of a relationship between attributes. Its most important ratios include covariance, the linear correlation coefficient and the coefficient of determination for attributes measurable on a ratio scale, and Spearman's and Kendall's rank correlation coefficients for ordinal-level variables. If there is a detectable relationship among the variables, then a regression model can be used for the description of the nature of the relationship. The direction of the relationship between the attributes may be identical or reversed, to which we can fit either a linear or a nonlinear function. Regression is a mathematical device for the description of a predetermined cause-and-effect relationship among the variables. We can use a sample regression function fitted to the empirical values to determine the direction and order of magnitude of the change in the result variable caused by a unit of increase in the explanatory variable. Changes in a phenomenon are often affected by more than one factor. In such cases we use multivariate models.

10.3.2 Self-test questions

What relationship between attributes is quantified by correlation?
What does covariance show?
Rank correlation quantifies a relationship between attributes of which level of measurement, and which are its major ratios?
What is shown by the linear correlation coefficient and the coefficient of determination?

[12] An example of a method that can handle mutually dependent explanatory variables is principal component analysis.

What type of diagram can be used for the representation of two relationships between quantitative attributes?
What is the difference between population and empirical regression functions?
What types of regression functions can be fitted to the data?
How can you express a bivariate linear regression function and what do its parameters mean?
What does the elasticity coefficient show?
What is the residue and what is residual standard deviation? What are they used for?

10.3.3 Practice tests

❁ Decide whether the statements below are true (T) or false (F).

Covariance may assume a value between -1 and 1, and its sign denotes the direction of the relationship.	**F**
Spearman's coefficient determines the strength of the relationship between attributes measurable on a ratio scale.	**F**
$r^2 = 0.52$ means that the explanatory variable affects changes in the result variable to an extent of 0.52%.	**F**
A condition on multivariate regression models is that the explanatory variable and the result variable must be independent of each other.	**F**
The regression function that fits the empirical values best is the one for which the value of the residual standard deviation is highest.	**F**
For a power regression function, the elasticity coefficient is identical with the value of the b parameter of the function.	**T**
The linear coefficient of determination represents the degree in % to which the explanatory variable determines changes in the result variable.	**T**
Based on the regression function y=325-12x, we can say that if the value of the explanatory variable is increased by one unit, then the value of the result variable will increase by 325 units.	**F**
r=0.894 means that there is a strong positive relationship between quantitative attributes.	**T**
Based on the regression function y=325-12x, we can say that if the value of the explanatory variable is increased by one unit, then the value of the result variable will decrease by 12 units.	**T**

11. TIME SERIES ANALYSIS TECHNIQUES

11.1 GOALS AND COMPETENCIES

There is a variety of different statistical tools to use in the examination of time series. So far, we have discussed methods for the analysis of data relating to a particular period of time. The purpose of temporal examinations, however, is to be able to say something about what to expect in the future on the basis of the data we have about the present or past. This unit describes techniques of time series analysis which allow us to make predictions on the basis of tendencies discovered in the empirical data. The point in the decompositional analysis of time series is to decompose time series into factors and quantify the effect of particular factors given particular empirical data. In deterministic time series analysis, the components of a time series may be clearly distinguished on the basis of the duration of the period of time in which they affect changes in values. Stochastic time series analysis, by contrast, is based on the assumption that a particular time-related value has an effect on how values change in future. So, elements in a time series can be described as a function of previous values and chance.

The purpose of this unit is to familiarize students with the basics of deterministic time series analysis, and to clarify the essence of the method through a number of examples. Students will learn about the components of time series and their properties. They will be able to interpret tendencies and make forecasts on the basis of their knowledge of the relationships between factors and the methods of their calculation. Students will understand the nature of the parameter that helps them decide about the trend that describes the long-term tendency that fits the empirical values best.

11.2 TOPICS

The simplest tool in the analysis of time series data is the dynamic ratio, which allows us to compare data connected to different points in time and identify the degree and direction of changes across two time-related pieces of data. We can compute indices for heterogeneous populations that make the comparison of data from two different periods of time possible. These methods are, however, unsuitable for the detection of tendencies in a time series. One method for the identification of regularities whose knowledge allows us to make predictions is deterministic time series analysis.

In this analysis, we distinguish between two different types of time series, which will serve as the basis for later examinations. The *empirical time series* is composed of empirical values, for which a temporal tendency may be de-

scribed. The *theoretical time series* is composed of elements of a particular phenomenon over a finite period of time, for the unknown values of which a prediction can be made on the basis of a tendency revealed in the empirical time series. Thus, an empirical time series can be regarded as a sample, and the theoretical time series can be regarded as the population.

First we will introduce the ratios of average absolute change and average relative change, which allow us to quantify the temporal direction of changes in data both in absolute terms and in percentages. However, the values of these indicators can be used to make predictions confidently only if elements of the time series change in the same direction and to approximately identical degrees.

When we intend to reveal a tendency in the data in deterministic time series analysis, we decompose the time series into components and quantify the effects of these factors separately. This method allows us to make a forecast for the future, as well as for a period within the observed period of time.

On the basis of known tendencies, a forecast for the future is called extrapolation and an estimation for an unknown point in time within the period examined is called interpolation.

It is worth noting in connection with extrapolation that a confident forecast can be made for a period of time which is not longer than one and a half times the length of the period examined. To take a simple example, on the basis of data about a 10-year long time series you can estimate data about the next 5 years at most. The further away into the future you make an estimate the less reliable it is going to be.

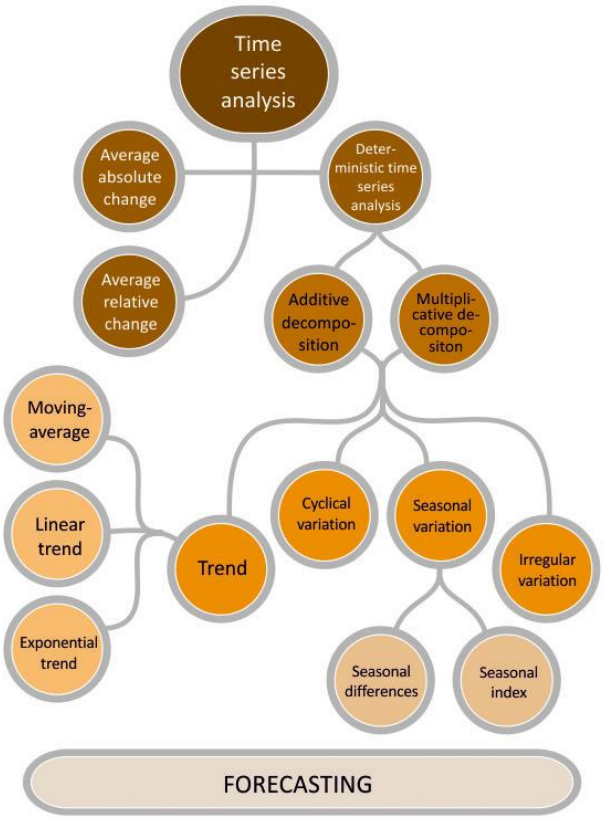


Figure 38 Time series analysis techniques

11.2.1 Average absolute and relative change

Regularities in time series may be easily described by the ratios of average absolute and relative change.[13] The ratios represent the average change in empirical values from one period of time to the next. They can be used to determine the direction and average degree of change in the observed data. We use an empirical time series (a sample) in the examination of a phenomenon, which must be borne in mind in the calculations.

- **Average absolute change expresses the average amount of change in units of measurement of the base data.**

[13] Average absolute change is sometimes called the average amount of development, and average relative change is sometimes termed the average rate of change.

For the calculation of average absolute change we use the differences between the values of consecutive periods of time. We compute their arithmetic mean, bearing in mind that the observations are not exhaustive (n-1).

$$\bar{d} = \frac{(y_2 - y_1) + (y_3 - y_2) + \cdots + (y_n - y_{n-1})}{n-1} = \frac{y_n - y_1}{n-1}$$

Average relative change expresses the average amount of change in percentage form.

For the calculation of average relative change we use the quotients of the values of consecutive time periods, which we average. Given that we are working with dynamic (chain) ratios, we compute a geometric mean.

$$\bar{l} = \sqrt[n-1]{\frac{y_2}{y_1} \cdot \frac{y_3}{y_2} \cdot \ldots \cdot \frac{y_n}{y_{n-1}}} = \sqrt[n-1]{\frac{y_n}{y_1}}$$

The direction of the comparison is fixed for both ratios—we compare a later period to an earlier period. The formulas can be simplified by removing the brackets and by what we know about the multiplication of fractions, which will show that it is sufficient for us to know the data of the last and the first periods for the calculation. This implies that the values of the ratios of average absolute and average relative change are sensitive to the two extreme values. If the data of the first or last period is outstanding, then the tendency expressed by the ratios will not describe the changes in the data adequately. The tendency of the time series will also be distorted by the values of the indicators if there is a considerable amount of fluctuation in the data within the period examined.

We have some data on the number of products sold by a company across 2009 and 2013. *Determine the average change in sales in period examined.*

28. *Number of products sold by a company across 2009 and 2013*

Year	Number of products sold (thousand pcs)
2009	240
2010	246
2011	257
2012	289
2013	320

We can determine the average amount and rate of change in sales in the period examined, i.e. over the past 5 years (n=5).

$$\bar{d} = \frac{y_n - y_1}{n-1} = \frac{320 - 240}{5-1} = \mathbf{20\ thousand\ pcs}$$

$$\bar{l} = \sqrt[n-1]{\frac{y_n}{y_1}} = \sqrt[5-1]{\frac{320}{240}} = \mathbf{1.0745} \rightarrow +7.45\%$$

The company's sales increased across 2009 and 2013 on average by 20 000 products, i.e. by 7.45%.

11.2.2 Deterministic time series analysis

In deterministic time series analysis we decompose the time series into its components. The components identified can be distinguished in terms of the length of time over which they exert their influence:

- trend ($\hat{Y}$): long-term tendency
- cyclic fluctuation or cycle (K): mid-term effect
- seasonal fluctuation or seasonality (S): short-term effect
- effect of chance (V)

These factors may interrelate in an additive and a multiplicative manner. In an additive relationship, the actual value (Y_t) is the sum of factors, where components are expressed in units of measurement of the base data. In a multiplicative relationship, the trend is specified in units of measurement of the base data, to which the other components bear a relative relationship, which can be expressed mathematically as the coefficient product of ratios compared to the trend.

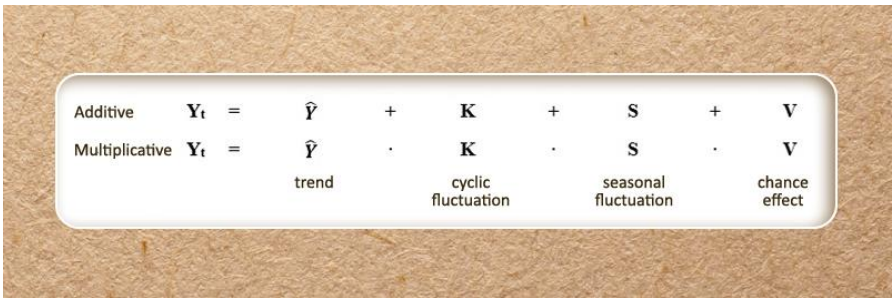


Figure 39 Relationships among elements of a time series

A **trend** is a long-term tendency in a time series, which determines the main direction of changes. The interpretation of what a long term is depends on the nature of the phenomenon examined, so its length cannot be defined universally.

Cyclic fluctuation (cycle) is prolonged undulation around the trend, movement above or below the trend over a relatively long (rather than a transitory) period of time. Fluctuation is periodical, not quite regular, and difficult to predict. Mid-term tendencies play a particularly important role in economy. You can read more about the major types of trends, named after renowned economists, on the following sites:

19. Business cycles: http://www.britannica.com/EBchecked/topic/86233/business-cycle
20. Kitchin cycle (3-5 years): http://www.policonomics.com/joseph-kitchin/
21. Juglar cycle (7-11years): http://www.policonomics.com/clement-juglar/
22. Kuznets cycle (15-25 years): http://www.dictionaryofeconomics.com/article?id=pde2008_K000045
23. Kondratyev cycle (45-60 years): http://www.newworldencyclopedia.org/entry/Nikolai_Kondratiev

Seasonal fluctuation is short-term, regular undulation around the trend. Fluctuation occurs at regular intervals of time, in periods of equal length, and it diverts values predictably in the same direction. A typical example of seasonality is tourist trade, where even its name is suggestive of the phenomenon. It is predictable that fewer visitors arrive in the preseason and in the postseason, and more in the high season, than what the trend would predict. Seasonality manifests itself in various fields of economy, such as agriculture or the area of seasonal products.

Chance fluctuation is short-tem, irregular movement. It is particularly important in the analysis of temporal data that we reckon with the possibility that an unexpected event may cause the values in the data set to differ from what is otherwise expected on the basis of the trend. Such an event may be, e.g., the devaluation of the national currency, an unexpected occurrence in the economy, or a crisis.

We will discuss trends and seasonality in some detail. Business cycle research is a special branch of economic statistics, and the effect of chance may be quantified residue-theoretically on the basis of the remaining three components.

Definition of trend

The purpose of defining the trend is to quantify the long-term tendency. The principal direction of changes is defined by the method of moving average calculation, which averages the values of a time series. In analytic trend calculation we specify the equation of the function which best approximates the values. We discuss the most frequently used trend functions, the linear and exponential forms.

Moving average calculation

In calculating the moving average, the long-term tendency will be the dynamic mean of the data we have. The first step in the calculation is to determine the number of terms in the moving average (k), i.e., the number of values to be averaged. Undulation in and the resolution of the time series are critical for the choice of the number of terms. Thus, for example, if we have data broken down to quarters, then it is a good idea to compute a four-term moving average. In regard to tourist trade, we have seasonal data, so it makes sense to compute a three-term moving average for the pre-, high, and postseason. When determining the number of terms, you may want to consider the measurability of the seasonal effect later. The method of calculating a moving average is less practical in cases where you only have yearly data or an undivided time series, because the determination of the number of terms will be arbitrary, and it may not be able to smooth out the data very well.

Once we have determined the number of terms, data may be averaged by first determining the arithmetic mean of the first *k* number of data, which we enter in the middle of the averaged period of time, i.e., in position (k+1)/2. For the determination of the other values, we exclude the first of the values already averaged and we instead enter the (k+1) period in the average, and the place of this value gets moved. We move the values entered in the average in a similar fashion along the data set and enter them in the appropriate places. In averaging, there will be no values calculated at the beginning and end of the time series, so the data series is shortened. If we have an odd number of terms in the data series, then it will be shorter by k-1 terms, and if the number of terms is even, then it will be k terms shorter.

In the event of an even number, the values calculated in moving averaging will fall in between pairs of values (in a four-term moving average, for example, the first average will fall between the 2^{nd} and 3^{rd} elements – (4+1)/2=2.5), which will need to be readjusted to the data series. This is done by truing. The point of the method is that data that fall between two values are averaged pair by pair, also by the method of moving. We will explain the method of the calculation through examples.

- We have the following data on guests' turnover for a chain of hotels. *Let us represent the time series and determine the values of the moving averaged trend. Let us calculate the values of the average absolute and relative change.*

29. *Changes in guests' seasonal turnover in a chain of hotels across 2010 and 2013*

Year	Season	Number of guests (thousand prsns)
2010	*Preseason*	25
	High season	64
	Postseason	34
2011	*Preseason*	26
	High season	66
	Postseason	32

Year	Season	Number of guests (thousand prsns)
2012	*Preseason*	24
	High season	70
	Postseason	40
2013	*Preseason*	27
	High season	72
	Postseason	38

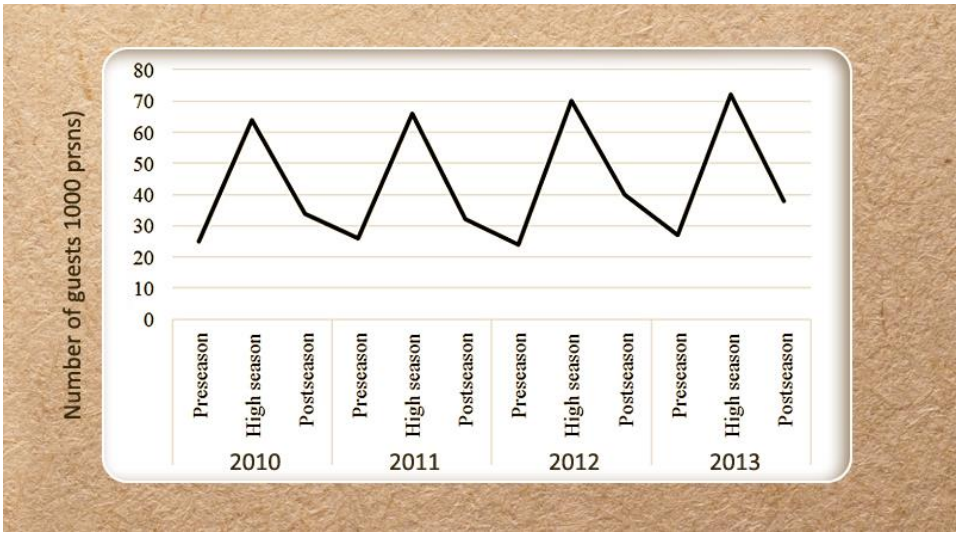


Figure 40 Changes in guests' seasonal turnover in a chain of hotels across 2010 and 2013

It is clear from the diagram that there is a marked seasonality effect in the time series. The number of guests was highest in the high season, it was considerably lower in the postseason, and even lower in the preseason. In order to reveal the long-term tendency, we need to smooth out the time series. This is done by moving averaging. First we need to determine the number of terms in the moving average, which is going to be three in this case, due to the seasonal segmentation, so we are going to

work with a data series composed of an odd number of data. We average 3 values in each case, and the averages will be placed in the same series as the second value (k+1)/2 = (3+1)/2= 2.

30. *Work table for calculating moving averages*

Year	*Season*	*Number of guests (thousand prsns)* y	*MOVING AVERAGE (MA)* $\hat{y}$
2010	Preseason	25	-
	High season.	64	$\frac{25+64+34}{3}=41$
	Postseason	34	$\frac{64+34+26}{3}=41.33$
2011	Preseason	26	$\frac{34+26+66}{3}=42$
	High season.	66	$\frac{26+66+32}{3}=41.33$
	Postseason	32	$\frac{66+32+24}{3}=40.67$
2012	Preseason	24	$\frac{32+24+70}{3}=42$
	High season.	70	$\frac{24+70+40}{3}=44.67$
	Postseason	40	$\frac{70+40+27}{3}=45.67$
2013	Preseason	27	$\frac{40+27+72}{3}=46.33$
	High season.	72	$\frac{27+72+38}{3}=45.67$
	Postseason	38	-
	Total	*518*	

In calculating the moving average, for determining the values of average absolute and average relative change, we take the values of the moving average and the number of terms in the shortened time series (shortened by k-1, which in this case is 3-1, i.e. 2) as the basis.

$$\bar{d} = \frac{\hat{y}_n - \hat{y}_1}{n-1} = \frac{45.67 - 41}{10-1} = \mathbf{0.519}\ thousand\ prsns$$

On the basis of moving average calculation, the number of guests increased on average by 519 people for each season in the period examined.

$$\bar{l} = \sqrt[n-1]{\frac{\hat{y}_n}{\hat{y}_1}} = \sqrt[10-1]{\frac{45.67}{41}} = \mathbf{1.0121} \rightarrow +1.21\%$$

On the basis of moving average calculation, the number of guests increased on average by 1.21 % for each season in the period examined.

- We have the following sales data of a company for the past 3 years, broken down to quarters. *Let us represent the time series and determine the values of the moving averaged trend. Let us calculate the values of the average absolute and relative change.*

31. *Changes in a company's sales across 2011 and 2013 broken down to quarters*

Period		Sales (1000 pcs)	Period		Sales (1000 pcs)	Period		Sales (1000 pcs)
2011.	I.	25	2012.	I.	34	2013.	I.	42
	II.	52		II.	55		II.	60
	III.	68		III.	70		III.	70
	IV.	54		IV.	54		IV.	58

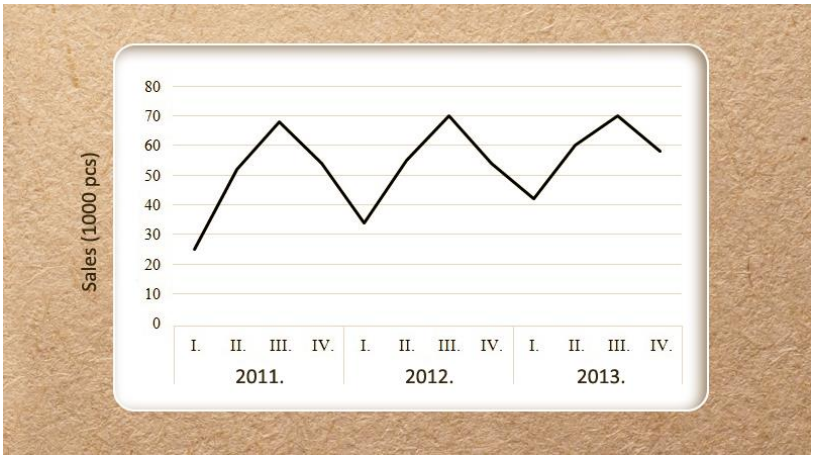


Figure 41 Changes in company's sales across 2011 and 2013 broken down to quarters

The diagram clearly shows the quarterly tendencies. The first quarters are the weakest, sales rise toward the middle of the year and culminate in the third quarter, to be followed by a decline toward the end of the year. Because of the obvious undulation in the quarterly data, the ideal choice is a four-term moving average. Thus, in this case, we are going to get an even number of terms in the moving average, which will require truing. First, we calculate the moving averages by averaging sets of 4 values, and then we will true the data series by averaging the calculated values in pairs.

32. *Work table for calculating trued moving averages*

		y_i (1000 pcs)	Moving average	$\hat{y}$ c.m.á trued moving average
2011.	I.	25		—
	II.	52	—	—
	III.	68	$\frac{25+52+68+54}{4} = 49.75$	$\frac{49.75+52}{2} = 50.875$
	IV.	54	$\frac{52+68+54+34}{4} = 52$	$\frac{52+52.75}{2} = 52.375$
2012.	I.	34	$\frac{68+54+34+55}{4} = 52.75$	$\frac{52.75+53.25}{2} = 53$
	II.	55	$\frac{54+34+55+70}{4} = 53.25$	$\frac{53.25+53.25}{2} = 53.25$
	III.	70	$\frac{34+55+70+54}{4} = 53.25$	$\frac{53.25+55.25}{2} = 54.25$
	IV.	54	$\frac{55+70+54+42}{4} = 55.25$	$\frac{55.25+56.5}{2} = 55.875$
2013.	I.	42	$\frac{70+54+42+60}{4} = 56.5$	$\frac{56.5+56.5}{2} = 56.5$
	II.	60	$\frac{54+42+60+70}{4} = 56.5$	$\frac{56.5+57.5}{2} = 57$
	III.	70	$\frac{42+60+70+58}{4} = 57.5$	—
	IV.	58	—	—

For the determination of the average absolute and average relative change, the values of the trued moving average and the number of terms in the shortened time series (which in this case is shortened by the number of terms, i.e. four) will be taken as the basis.

$$\bar{d} = \frac{\hat{y}_n - \hat{y}_1}{n-1} = \frac{57 - 50.875}{8-1} = \mathbf{0.875}\ thousand\ pcs$$

On the basis of calculating the trued moving average, sales increased quarterly by an average of 875 products in the period examined.

$$\bar{l} = \sqrt[n-1]{\frac{\hat{y}_n}{\hat{y}_1}} = \sqrt[8-1]{\frac{57}{50.875}} = \mathbf{1.0164} \rightarrow +1.64\%$$

On the basis of calculating the trued moving average, sales increased quarterly by an average of 1.64 % in the period examined.

Analytic trend calculation: linear and exponential trends

In analytic trend calculation we use a function to describe the long-term tendency observed in the empirical data. Most commonly, we fit a linear and an exponential function to the data. A trend function is only formally similar to a regression function: we do not necessarily assume a cause-and-effect relationship between time and the observed phenomenon.[14] A trend, then, reveals the temporal regularity in the data by a function, offering information on the direction and degree of changes.

In the determination of trend functions, the first thing we do is assign numerical values to the periods, which will allow the use of methods developed for the analysis of quantitative attributes in the case of time series. These values are called t values,[15] which may be assigned to periods in two different ways.

The **t=1,2..n** method, or monotonous t method, assigns values to periods by assigning the value 1 first and then values increasing in equal increments are assigned to the periods. So, the value assigned to the first period is 1, the value assigned to the second period is 2, and so on.

The **∑t=0** method assigns values to periods in such a way that the sum of those values is 0. In practice, this is carried out by assigning the value 0 to the (n+1)/2 element of the time series, and assigning negative values moving backwards and positive values moving forwards, in equal increments. For example, in the case of a 9-year time series, 0 will be assigned to the datum of the 5th year, -1 to the datum of the 4th year, and 1 to the datum of the 6th year. In the case of an even-numbered time series, the (n+1)/2 ele-

[14] Due to the formal similarity, the trend function is often called special regression, whose explanatory variable is time. However, this would mean, in analogy with a regression rel ationship, that changes in the data are always determined by time, i.e., data change du e to the passage of time. But in trend calculation, what we take as a basis is not time bu t the numerical values assigned to periods of time. So, time is merely an indirect factor.

[15] where “t” abbreviates “time.”

ment falls between two values. In this case, the value 0 will not be assumed by any period. The immediately preceding period is assigned the value -1 and the immediately following period is assigned the value 1. Since the distance between these two values is two steps, values will change from period to period in increments of two. For example, in the case of a 10-year time series, the 5.5th element would be 0. To avoid this, the 5th element is assigned the value -1 and the 6th element is assigned the value 1, and the values assigned to the 4th and 7th elements are -3 and +3, respectively.

33. *Example of assigning t values with $\sum t = 0$ method*

Period	t value assigned
2010	-1
2011	0
2012	1

Period	t value assigned
2009	-3
2010	-1
2011	1
2012	3

A **linear trend** is a straight line approximating the associated pairs of values.

The population trend function for the theoretical time series is this:

$$\boldsymbol{Y = \beta_0 + \beta_1 t + \varepsilon}$$

The interpretation of the parameters β_0 and β_1 will differ depending on whether we used the t=1,2…n method or the $\sum t=0$ method to fit the trend to the data. ε is the effect of chance, which, given that it is an estimated value, needs to be considered.

The empirical trend function that can be fitted to the observed values is this:

$$\boldsymbol{\hat{y} = b_0 + b_1 \cdot t}$$

Of the parameters of the trend function, the sign of β_1 and b_1 is of particular importance, as it shows the direction of the change in observed values.

Determination of the parameters of the linear trend function

In the **t = 1,2 … n** (monotonous t) method, the equation of the trend function may be expressed by normal equations. The number of observed periods is also important in the calculation. The solution to the two-variable system of equations yields the values of the b_0 and b_1 parameters of the function.

$$\boldsymbol{\sum y = b_0 \cdot n + b_1 \cdot \sum t}$$

$$\sum ty = b_0 \cdot \sum t + b_1 \cdot \sum t^2$$

Parameter b_0 of the linear trend function expressed by the **t = 1,2 … n** (monotonous t) method is the value of the period immediately preceding the observed period, i.e., the value that the function assumes in place t=0, and parameter b_1 represents the average absolute change from period to period.

In the **∑t = 0** method, the equation of the trend function may be expressed by formulae.

$$b_0 = \frac{\sum y}{n} \qquad\qquad b_1 = \frac{\sum ty}{\sum t^2}$$

Parameter b_0 of the linear trend function expressed by the ∑t = 0 method is the value of the (n+1)/2 element of the observed period, the arithmetic mean of the empirical values, and parameter b_1 represents the average absolute change from period to period. For an even-numbered data series, where the values of t increase in increments of two, the change from one period to the next may be interpreted as a degree of $2 \cdot b_1$.

❁ We have the following sales data of a company for the past 3 years, broken down to quarters (thousand pcs). *Determine the equation of the linear trend with the t=1,2… n method and with the ∑t=0 method.*

34. *Work table for determining the linear trend*

		y_i (1000 pcs)	t = 1,2 …n			∑t = 0		
			t	t · y	t^2	t	t · y	t^2
2011.	I.	25	1	25	1	-11	-275	121
	II.	52	2	104	4	-9	-468	81
	III.	68	3	204	9	-7	-476	49
	IV.	54	4	216	16	-5	-270	25
2012.	I.	34	5	170	25	-3	-102	9
	II.	55	6	330	36	-1	-55	1
	III.	70	7	490	49	1	70	1
	IV.	54	8	432	64	3	162	9
2013.	I.	42	9	378	81	5	210	25
	II.	60	10	600	100	7	420	49
	III.	70	11	770	121	9	630	81
	IV.	58	12	696	144	11	638	121
	Σ	642	78	4415	650	0	484	572

*Determining the equation of the linear trend function using the **t=1,2…n** method - normal equations*

$$\sum y = b_0 \cdot n + b_1 \cdot \sum t$$
$$\sum ty = b_0 \cdot \sum t + b_1 \cdot \sum t^2$$

$$642 = b_0 \cdot 12 + b_1 \cdot 78 \quad \rightarrow \quad b_0 = \frac{642 - b_1 \cdot 78}{12}$$
$$4415 = b_0 \cdot 78 + b_1 \cdot 650$$

SOLUTION: $4415 = \frac{642 - b_1 \cdot 78}{12} \cdot 78 + b_1 \cdot 650$ one-variable equation

$$4415 = (642 - b_1 \cdot 78) \cdot 6.5 + b_1 \cdot 650 \rightarrow 4415 = 4173 - b_1 \cdot 507 + b_1 \cdot 650$$

$$242 = b_1 \cdot 143 \rightarrow \boldsymbol{b_1 = 1.6923} \rightarrow \boldsymbol{b_0} = \frac{642 - 1.6923 \cdot 78}{12} = \boldsymbol{42.5}$$

Equation of linear trend function (t=1,2…n): $\hat{y} = 42.5 + 1.6923 \cdot t$

According to the linear trend function, the company sold 42.5 thousand products in the 4th quarter of 2010, and sales increased quarterly by an average of 1.6923 thousand products in the period examined.

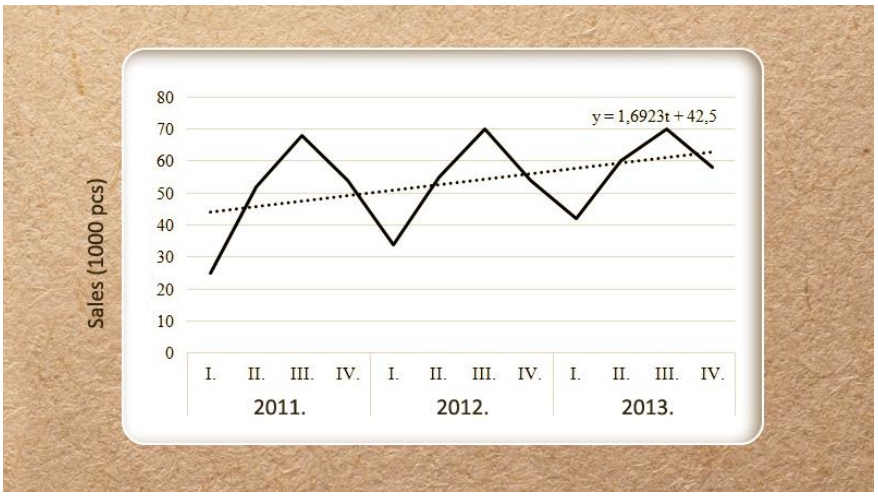


Figure 42 Changes in sales of the company across 2011 and 2013 broken down to quarters and the fitted linear trend function (t=1,2 …n)

Determining the equation of the linear trend function using the $\boldsymbol{\sum t = 0}$ method – formulae

$$b_0 = \frac{\sum y}{n} = \frac{642}{12} = 53.5 \qquad b_1 = \frac{\sum ty}{\sum t^2} = \frac{484}{572} = 0.84615$$

Equation of linear trend function ($\sum t = 0$): $\hat{y} = 53.5 + 0.84615 \cdot t$

According to the linear trend function, 53.5 thousand products were sold across the 2nd and 3rd quarters of 2012, and sales increased quarterly by an average of 1.6923 thousand products in the period examined (as values of t increase in increments of 2, therefore 2· 0.84615 = 1.6923).

The **exponential trend** is an accelerating function that approximates to the associated pairs of values. The population trend function for the theoretical time series is

$$Y = \alpha \cdot \beta^t \cdot \varepsilon$$

The interpretation of the α and β parameters will differ in this case too for the two different methods, t=1,2…n and $\sum t=0$. The effect of chance (ε) must be considered here too for a population function.

The empirical exponential trend function fitted to the observed values is this:

$$\hat{y} = a \cdot b^t$$

Determining the parameters of the exponential trend function

In order to determine the parameters, the function needs to be logarithmically transformed:

$$\lg \hat{y} = \lg a + \lg b \cdot t$$

We can express a linear relationship on the basis of the logarithmic function, if we substitute new variables for the values *lga* and *lgb*.

$$\text{Let } lg\ a = b_0 \text{ and } lg\ b = b_1$$

This yields the following relationship:

$$\lg \hat{y} = b_0 + b_1 \cdot t$$

We can determine the normal equations and formulae for this equation in a manner familiar from the discussion of the linear trend function.

In the **t = 1,2 … n** (monotonous t) method, the exponential equation of the trend function may also be expressed by normal equations. The solution to the two-variable system of equations yields the values of the b_0 and b_1 parameters of the function, which must be reconverted into parameters a and b.

$$\sum \lg y = b_0 \cdot n + b_1 \cdot \sum t$$
$$\sum t \cdot \lg y = b_0 \cdot \sum t + b_1 \cdot \sum t^2$$

Reconversion of the parameters:

$$b_1 = lg\ b \rightarrow \mathbf{b} = 10^{b_1}$$
$$b_0 = lg\ a \rightarrow \mathbf{a} = 10^{b_0}$$

Parameter a of the linear trend function expressed by the **t = 1,2 … n** (monotonous t) method is the value of the period immediately preceding the observed period, i.e., the value that the function assumes in place t=0, and parameter b represents the average absolute change from period to period.

In the **∑t = 0** method, the equation of the trend function may be expressed by formulae.

$$b_0 = \frac{\sum lg\ y}{n} \qquad b_1 = \frac{\sum t \cdot lgy}{\sum t^2}$$

Parameter a of the exponential trend function expressed by the ∑t = 0 method is the value of the (n+1)/2 element of the observed period, the geometric mean of the empirical values, and parameter b represents the average relative change from period to period. For an even-numbered data series, where the values of t increase in increments of two, the change from one period to the next may be interpreted as a degree of b^2.

❁ We have the following sales data of a company for the past 3 years, broken down to quarters. *Determine the equation of the exponential trend with the t=1,2… n method and with the ∑t=0 method.*

35. *Work table for determining the exponential trend*

		y_i (1000 pcs)	lg y	t = 1.2 …n			∑t = 0		
				t	t · lgy	t^2	t	t · lgy	t^2
2011.	I.	25	1.3979	1	1.3979	1	-11	-15.3773	121
	II.	52	1.7160	2	3.4320	4	-9	-15.4440	81
	III.	68	1.8325	3	5.4975	9	-7	-12.8276	49
	IV.	54	1.7324	4	6.9296	16	-5	-8.6620	25
2012.	I.	34	1.5315	5	7.6574	25	-3	-4.5944	9
	II.	55	1.7404	6	10.4422	36	-1	-1.7404	1
	III.	70	1.8451	7	12.9157	49	1	1.8451	1
	IV.	54	1.7324	8	13.8592	64	3	5.1972	9

2013.	I.	42	1.6232	***9***	14.6092	*81*	***5***	8.1162	*25*
	II.	60	1.7782	***10***	17.7815	*100*	***7***	12.4471	*49*
	III.	70	1.8451	***11***	20.2961	*121*	***9***	16.6059	*81*
	IV.	58	1.7634	***12***	21.1611	*144*	***11***	19.3977	*121*
	∑	*642*	20.5381	***78***	135.9794	*650*	***0***	4.9635	*572*

Determining the equation of the exponential trend function using the **t=1,2…n** method - normal equations

$$\sum \mathbf{lg}\, \boldsymbol{y} = \boldsymbol{b_0} \cdot \boldsymbol{n} + \boldsymbol{b_1} \cdot \sum \boldsymbol{t}$$
$$\sum \boldsymbol{t} \cdot \mathbf{lg}\, \boldsymbol{y} = \boldsymbol{b_0} \cdot \sum \boldsymbol{t} + \boldsymbol{b_1} \cdot \sum \boldsymbol{t^2}$$

$$20.5381 = b_0 \cdot 12 + b_1 \cdot 78 \rightarrow b_0 = \frac{20.5381 - b_1 \cdot 78}{12}$$
$$135.9794 = b_0 \cdot 78 + b_1 \cdot 650$$

$$135.9794 = \frac{20.5381 - b_1 \cdot 78}{12} \cdot 78 + b_1 \cdot 650$$

$$b_1 = 0.01735 = \lg b \rightarrow \boldsymbol{b} = 10^{0.01735} = \mathbf{1.0408}$$

$$\downarrow$$

$$b_0 = \frac{20.5381 - 0.01735 \cdot 78}{12} = 1.5987 = lg\ a \rightarrow \mathbf{a} = 10^{1.5987} = \mathbf{39.692}$$

Equation of exponential trend function (t=1,2…n): $\hat{y} = \mathbf{39.692} \cdot \mathbf{1.0408}^{t}$

According to the exponential trend function, the company sold 39.962 thousand products in the 4th quarter of 2010, and sales increased quarterly by an average of 4.08 % in the period examined.

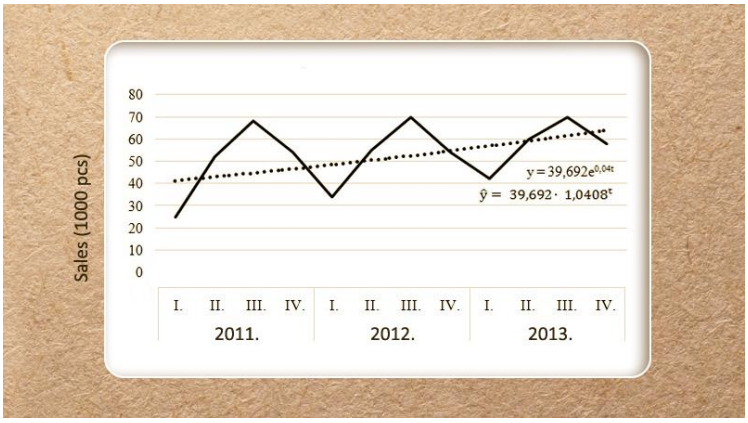


Figure 43 Changes in sales of the company across 2011 and 2013 broken down to quarters and the fitted exponential trend function (t=1,2 …n)

Determining the equation of the exponential trend function using the $\sum \mathbf{t} = \mathbf{0}$ method

$$\boldsymbol{b_0} = \frac{\sum \boldsymbol{lg\ y}}{\boldsymbol{n}} = \frac{20.5381}{12} = 1.7115 \rightarrow \mathbf{a} = 10^{1.7115} = \mathbf{51.46}$$

$$\boldsymbol{b_1} = \frac{\sum \boldsymbol{t} \cdot \boldsymbol{lgy}}{\sum \boldsymbol{t^2}} = \frac{4.9635}{572} = 0.00868 \rightarrow \boldsymbol{b} = 10^{0.00868} = \mathbf{1.0202}$$

According to the exponential trend function, 51.46 *thousand products were sold across the 2nd and 3rd quarters of 2012, and sales increased quarterly by an average of* 4.08% (1.0202^2) *in the period examined.*

Seasonality

Seasonality is short-term periodic undulation. Depending on the relationship between the elements of a time series, we calculate seasonal deviations (additive) and season indices (multiplicative). Seasonal deviations are determined from the pair by pair differences of the actual values and the values estimated by the trend (residue), and season indices are determined from the pair by pair quotients of the values. When seasonal undulation is filtered out from the elements of the time series, seasonal effects must balance each other out.

In determining the **seasonal deviation**, we determine the arithmetic mean of the differences in recurring periods, which is called raw seasonal deviation. A sum of means must come out to 0, therefore, if the sum of raw seasonal deviations is not 0, values must be corrected. The correction factor (k) is the simple arithmetic mean of the raw seasonal deviations. Corrected seasonal deviations are derived by reducing the values of the raw seasonal deviations by the amount of the correction factor. In the case of a linear trend and additive seasonality, raw and corrected seasonal deviations are equal.

In determining the **seasonal indices,** we determine the geometric mean of the quotients of recurring periods, which is called raw season index. The product of the means must be 1, therefore, if the product of the raw season indices is not 1, the values must be corrected. The correction factor is the geometric mean of the raw season indices. The values of the corrected season indices are derived by dividing the values of raw season indices by the amount of the correction factor. In the case of an exponential trend and multiplicative seasonality, raw and corrected seasonal indices are equal.

- We have the following quarterly sales data of a company across 2011 and 2013, and the values of the trued moving average. *Determine the degree of seasonality.*

36. *Pair-by-pair quotients and differences of the values of actual and the moving averaged trends*

		y_i (1000 pcs)	Moving average	$\hat{y}$ c.m.á trued moving average	$\frac{y_i}{\hat{y}\,c.m.á}$	$e = y_i - \hat{y}\,c.m.á$ (residue)	$(y_i - \hat{y}\,c.m.á)^2$ e^2
2011.	I.	25	—	–	—	—	—
	II.	52	49.75	–	—	—	—
	III.	68	52	50.875	1.3366	17.125	293.2656
	IV.	54	52.75	52.375	1.0310	1.625	2.6406
2012.	I.	34	53.25	53	0.6415	-19	361
	II.	55	53.25	53.25	1.0329	1.75	3.0625
	III.	70	55.25	54.25	1.2903	15.75	248.0625
	IV.	54	56.5	55.875	0.9664	-1.875	3.5156
2013	I.	42	56.5	56.5	0.7434	-14.5	210.25
	II.	60	57.5	57	1.0526	3	9
	III.	70	—	—	—	—	—
	IV.	58	—	–	—	—	—
						$\sum e^2$	1130.7968

For additive seasonality, we use the pair by pair differences of the values ($y_i - \hat{y}\,c.m.á$).

37. *Work table for determining seasonal deviations in the case of moving average calculation*

	I.	II.	III.	IV.	
2011.	—	—	17.125	1.625	
2012.	-19	1.75	15.75	-1.875	
2013.	-14.5	3	—	—	*Totals*
Raw seasonal deviation (arithmetic mean of deviations)	-16.75	2.375	16.4375	-0.125	1.9375
Corrected seasonal deviation: $\boldsymbol{S_i}$	*-17.234*	*1.891*	*15.9535*	*-0.609*	***0***

Correction factor: $k = \frac{1.9375}{4} = 0.484 \rightarrow Corrected = \text{Raw} - k$

Actual sales are lower by an average of 17.234 thousand products in the 1st quarters, by an average of 609 products in the 4th quarters, higher by an average of 1.891 thousand products in the 2nd quarters, and by an average of 15.9535 thousand products in the 3rd quarters than the value estimated from the trued moving average.

For *multiplicative seasonality,* we use the pair by pair quotients of the values.[16]

38. *Work table for determining season indices in the case of moving average calculation*

$\frac{y_i}{\hat{y}\ c.m.á}$		I.	II.	III.	IV.	
	2010.	—	—	1.3366	1.0310	
	2011.	0.6415	1.0329	1.2903	0.9664	
	2012.	0.7434	1.0526	—	—	*Product*
Raw season index (geometric mean of changes)		0.6906	1.0427	1.3132	0.9982	0.9439
Corrected season index: $\boldsymbol{S_i}$		*0.7006*	*1.0578*	*1.3323*	*1.0127*	***1***

Correction factor: $k = \sqrt[4]{0.9439} = 0.9857 \rightarrow \text{Korrigált} = \frac{\text{Nyers}}{\text{k}}$

Actual sales are lower by an average of 29.94% in the 1st quarters and higher by an average of 5.78% in the 2nd quarters, by an average of 33.23% in the 3rd quarters, and by an average of 1.27% in the 4th quarters than the value estimated from the trued moving average.

[16] For the sake of accuracy of calculations, it is a good idea to specify values to the 4th decimal.

❁ We have the following quarterly sales data of a company across 2011 and 2013, and the equation of the linear trend $\hat{y} = 42.5 + 1.6923 \cdot t$ (t=1,2,…n). *Determine the quarterly values estimated from the trend and the degree of seasonality.*

39. *Pair by pair differences and quotients of actual values and values estimated from the linear trend*

		y_i (1000 pcs)	t	$\hat{y} = 42.5 + 1.6923 \cdot t$	$e = y_i - \hat{y}$	$(y_i - \hat{y})^2 = e^2$	$\frac{y_i}{\hat{y}}$
2011.	I.	25	1	*44.19*	*-19.19*	*368.344*	*0.5657*
	II.	52	2	*45.88*	*6.12*	*37.398*	*1.1333*
	III.	68	3	*47.58*	*20.42*	*417.103*	*1.4293*
	IV.	54	4	*49.27*	*4.73*	*22.380*	*1.0960*
2012.	I.	34	5	*50.96*	*-16.96*	*287.692*	*0.6672*
	II.	55	6	*52.65*	*2.35*	*5.505*	*1.0446*
	III.	70	7	*54.35*	*15.65*	*245.045*	*1.2880*
	IV.	54	8	*56.04*	*-2.04*	*4.155*	*0.9636*
2013.	I.	42	9	*57.73*	*-15.73*	*247.455*	*0.7275*
	II.	60	10	*59.42*	*0.58*	*0.333*	*1.0097*
	III.	70	11	*61.12*	*8.88*	*78.938*	*1.1454*
	IV.	58	12	*62.81*	*-4.81*	*23.113*	*0.9235*
	Σ	*642*		*642*		*1737.462*	

Additive seasonality

40. *Work table for determining seasonal deviations in the case of a linear trend*

	I.	II.	III.	IV.	
2011.	-19.19	6.12	20.42	4.73	
2012.	-16.96	2.35	15.65	-2.04	
2013.	-15.73	0.58	8.88	-4.81	*Total*
Raw =corrected seasonal deviation: S_i	*-17.29*	*3.02*	*14.98*	*-0.71*	***0***

In the case of a linear trend and additive seasonality, the sum of raw seasonal deviations is 0, i.e. raw and corrected seasonal deviations are equal.

Actual sales are lower by an average of 17.29 thousand products in the 1st quarters, by an average of 0.71 thousand products in the 4th quarters, higher by an average of 3.02 thousand products in the 2nd quarters, and by an average of 14.98 thousand products in the 3rd quarters than the value estimated from the linear trend.

Multiplicative seasonality

41. *Work table for determining season indices in the case of a linear trend*

	I.	II.	III.	IV.	
2010.	0.5657	1.333	1.4293	1.096	
2011.	0.6672	1.0446	1.288	0.9636	
2012.	0.7275	1.0097	1.1454	0.9235	*Products*
Raw season index	0.6500	1.1203	1.2823	0.9917	0.9260
Corrected season index: S_i	*0.6626*	*1.1420*	*1.3072*	*1.0109*	1

Correction factor: $k = \sqrt[4]{0.926} = 0.981 \rightarrow Corrected = \frac{\text{Raw}}{\text{k}}$

Actual sales are lower by an average of 33.74% in the 1st quarters and higher by an average of 14.2% in the 2nd quarters, by an average of 30.72% in the 3rd quarters, and by an average of 1.09% in the 4th quarters than the value estimated from the linear trend.

❁ We have the following quarterly sales data of a company across 2011 and 2013, and the equation of the exponential trend $\hat{y} = 39.692 \cdot 1.0408^t$ (t=1,2,…n). *Determine the quarterly values estimated from the trend and the degree of seasonality.*

42. *Pair by pair differences and quotients of actual values and values estimated from the exponential trend*

		y_i (1000 pcs)	t	$\hat{y} = 39.692 \cdot 1.0408^t$	$e = y_i - \hat{y}$	$(y_i - \hat{y})^2 = e^2$	$\frac{y_i}{\hat{y}}$
2011.	I.	25	1	41.311	-16.311	266.063	0.6052
	II.	52	2	42.997	9.003	81.055	1.2094
	III.	68	3	44.751	23.249	540.506	1.5195
	IV.	54	4	46.577	7.423	55.100	1.1594
2012.	I.	34	5	48.477	-14.477	209.595	0.7014
	II.	55	6	50.455	4.545	20.654	1.0901
	III.	70	7	52.514	17.486	305.765	1.3330
	IV.	54	8	54.656	-0.656	0.431	0.9880
2013.	I.	42	9	56.886	-14.886	221.605	0.7383
	II.	60	10	59.207	0.793	0.628	1.0134
	III.	70	11	61.623	8.377	70.174	1.1359
	IV.	58	12	64.137	-6.137	37.666	0.9043
	Σ	*642*				1809.242	

Additive seasonality

43. *Work table for determining seasonal deviations in the case of an exponential trend*

	I.	II.	III.	IV.	
2011.	-16.311	9.003	23.249	7.423	
2012.	-14.477	4.545	17.486	-0.656	
2013.	-14.886	0.793	8.377	-6.137	*Totals*
Raw seasonal deviation	*-15.225*	*4.780*	*16.371*	*0.210*	*6.136*
Corrected seasonal deviation: S_i	*-16.759*	*3.246*	*14.837*	*-1.324*	***0***

Correction factor: $k = \frac{6.136}{4} = 1.534 \rightarrow Corrected = \text{Raw} - k$

Actual sales are lower by an average of 16.759 thousand products in the 1st quarters, by an average of 1.324 thousand products in the 4th quarters, higher by an average of 3.246 thousand products in the 2nd quarters, and by an average of 14.837 thousand products in the 3rd quarters than the value estimated from the exponential trend.

Multiplicative seasonality

44. *Work table for determining season indices in the case of an exponential trend*

	I.	II.	III.	IV.	
2011.	0.6052	1.2094	1.5195	1.1594	
2012.	0.7014	1.0901	1.333	0.988	
2013.	0.7383	1.0134	1.1359	0.9043	*Product*
Raw=Corrected season index: S_i	*0.6793*	*1.1014*	*1.3202*	*1.0118*	***1***

In the case of an exponential trend and multiplicative seasonality, the product of the raw season indices is 1, i.e. raw and corrected season indices are equal.

Actual sales are lower by an average of 32.07% in the 1st quarters and higher by an average of 10.14% in the 2nd quarters, by an average of 32.02% in the 3rd quarters, and by an average of 1.18% in

the 4th quarters than the value estimated from the exponential trend.

Forecasts

The purpose of time series analysis is to gain information for an unknown period of time in regard to a phenomenon of interest. We most commonly make estimates about the future, i.e. extrapolate. In this process, we determine the values for the period that follows the period examined on the basis of the long-term tendency, or trend, and seasonality.[17] In making a forecast we assume that the tendency and periodic undulation revealed within the period examined will continue. The trend and seasonality may interrelate either additively or in a multiplicative manner, which allows us to make two different types of forecasts in regard to the data examined.

- Consider the examples used above and make forecasts for 2014 by combining all the trends and seasonality data.

Forecast for 2014 on the basis of calculating the moving average and additive seasonality

$$\bar{\mathbf{d}} = \frac{\hat{y}_n - \hat{y}_1}{n-1} = \frac{57 - 50.875}{8-1} = \mathbf{0.875}\ thousand\ pcs$$

On the basis of trued moving average calculation, the number of products sold in the period examined increased by 875 pcs. The last period with a trued moving average was the 2nd quarter of 2013, therefore the number of products sold in the 3rd quarter of 2013 was 57+0.875= 57.875 pcs, and in the 4th quarter of 2013 it was 57.875+0.875=58.75 or $57 + 0.875 \cdot 2 = 58.75$ pcs. Hence, the number of products sold is going to be $57 + 0.875 \cdot 3 = 59.625$ pcs for the 1st quarter of 2014.

45. *Work table for forecast – moving average and additive seasonality*

		$\hat{y}\ c.m.$ á	S_i	$Forecast: \hat{y}\ c.m.$ á $+ S_i$
2014.	I.	$57 + 0.875 \cdot 3 = 59.625$	-17.234	$59.625 - 17.234 = \mathbf{42.391}$
	II.	$57 + 0.875 \cdot 4 = 60.5$	1.891	$60.5 + 1.891 = \mathbf{62.391}$
	III.	$57 + 0.875 \cdot 5 = 61.375$	15.9535	$61.375 + 15.9535 = \mathbf{77.3285}$
	IV.	$57 + 0.875 \cdot 6 = 62.25$	-0.609	$62.25 - 0.609 = \mathbf{61.641}$

[17] Effects of cyclic fluctuation and chance are disregarded now.

On the basis of calculating the moving average and additive seasonality, the company is expected to sell 42.391 thousand products in the 1st quarter of 2014, 62.391 thousand products in the 2nd quarter, 77.3285 thousand products in the 3rd quarter, and 61.641 thousand products in the 4th quarter.

Forecast for 2014 on the basis of calculating the moving average and multiplicative seasonality

We can use the values of the trued moving average just computed above (the computation is identical) in making a forecast based on multiplicative seasonality.

46. *Work table for forecast – moving average and multiplicative seasonality*

		$\hat{y}\,c.m.$á	S_i	$Forecast: \hat{y}\,c.m.$á $\cdot S_i$
2014.	I.	59.625	0.7006	$59.625 \cdot 0.7006 = \mathbf{41.77}$
	II.	60.5	1.0578	$60.5 \cdot 1.0578 = \mathbf{64}$
	III.	61.375	1.3323	$61.375 \cdot 1.3323 = \mathbf{81.77}$
	IV.	62.25	1.0127	$62.25 \cdot 1.0127 = \mathbf{63.04}$

On the basis of calculating the moving average and multiplicative seasonality, the company is expected to sell 41.77 thousand products in the 1st quarter of 2014, 64 thousand products in the 2nd quarter, 81.77 thousand products in the 3rd quarter, and 63.04 thousand products in the 4th quarter.

Forecasts on the basis of a linear trend and additive seasonality

In the case of a linear trend, with the numbering t=1,2..n, the 4th quarter of 2013 assumed the value t=12. For 2014, the numbering continues with the 13th value, which, when substituted into the equation of the trend, will yield the values estimated from the trend.

47. *Work table for forecasts – linear trend and additive seasonality*

		t	value estimated from trend $\hat{y}$	S_i	$Forecast: \hat{y} + S_i$
2014.	I.	13	$42.5 + 1.6923 \cdot 13 = \mathbf{64.5}$	- 13.03	64.5-13.03 = **51.47**
	II.	14	$42.5 + 1.6923 \cdot 14 = \mathbf{66.2}$	2.43	66.2+2.43 = **68.63**
	III.	15	$42.5 + 1.6923 \cdot 15 = \mathbf{67.9}$	12.57	67.9+12.57 = **80.47**
	IV.	16	$42.5 + 1.6923 \cdot 16 = \mathbf{69.6}$	–1.97	69.6-1.97 = **67.63**

On the basis of a linear trend and additive seasonality, the company is expected to sell 51.47 thousand products in the 1st quarter of 2014, 68.63 thousand products in the 2nd quarter, 80.47 thousand products in the 3rd quarter, and 67.63 thousand products in the 4th quarter.

Forecasts on the basis of a linear trend and multiplicative seasonality

48. *Work table for forecasts – linear trend and multiplicative seasonality*

		t	$\hat{y}$	S_i	$Forecast: \hat{y} \cdot S_i$
2014.	I.	13	64.5	0.6626	$64.5 \cdot 0.6626 = \mathbf{42.74}$
	II.	14	66.2	1.142	$66.2 \cdot 1.142 = \mathbf{75.6}$
	III.	15	67.9	1.3072	$67.9 \cdot 1.3072 = \mathbf{88.76}$
	IV.	16	69.6	1.0109	$69.6 \cdot 1.0109 = \mathbf{70.36}$

On the basis of a linear trend and multiplicative seasonality, the company is expected to sell 42.74 thousand products in the 1st quarter of 2014, 75.6 thousand products in the 2nd quarter, 88.76 thousand products in the 3rd quarter, and 70.36 thousand products in the 4th quarter.

Forecasts on the basis of an exponential trend and additive seasonality

In the case of an exponential trend, with the numbering t=1,2..n, the 4th quarter of 2013 assumed the value t=12. For 2014, the numbering continues with the 13th value here too, which, when substituted into the equation of the trend, will yield the values estimated from the trend.

49. *Work table for forecasts – exponential trend and additive seasonality*

		t	value estimated from trend $\hat{y}$	S_i	$Forecast: \hat{y} + S_i$
2014	I.	13	$39.692 \cdot 1.0408^{13} = 66.75$	-16.759	66.75-16.759=**49.99**
	II.	14	$39.692 \cdot 1.0408^{14} = 69.48$	3.246	69.48+3.246=**72.73**
	III.	15	$39.692 \cdot 1.0408^{15} = 72.31$	14.837	72.31+14.837=**87.15**
	IV.	16	$39.692 \cdot 1.0408^{16} = 75.26$	-1.324	75.26-1.324=**73.94**

On the basis of an exponential trend and additive seasonality, the company is expected to sell 49.99 thousand products in the 1st quarter of 2014, 72.73 thousand products in the 2nd quarter, 87.15 thousand products in the 3rd quarter, and 73.94 thousand products in the 4th quarter.

Forecasts on the basis of an exponential trend and multiplicative seasonality

50. *Work table for forecasts – exponential trend and multiplicative seasonality*

		t	$\hat{y}$	S_i	*Forecast*: $\hat{y} \cdot S_i$
2014.	I.	13	66.75	0.6793	$66.75 \cdot 0.6793 = \mathbf{45.34}$
	II.	14	69.48	1.1014	$69.48 \cdot 1.104 = \mathbf{76.71}$
	III.	15	72.31	1.3202	$72.31 \cdot 1.3202 = \mathbf{95.46}$
	IV.	16	75.26	1.0118	$75.26 \cdot 1.0118 = \mathbf{76.15}$

On the basis of an exponential trend and multiplicative seasonality, the company is expected to sell 45.34 thousand products in the 1st quarter of 2014, 76.71 thousand products in the 2nd quarter, 95.46 thousand products in the 3rd quarter, and 76.15 thousand products in the 4th quarter.

11.2.3 Examining the best fit of trends

Several different trends may be fitted to the empirical values. The best forecast may be made with the trend that best approximates the empirical values. The way to determine which trend fits the empirical data best is to calculate the residual standard deviation based on the pair by pair deviation of actual and estimated values.

$$s_e = \sqrt{\frac{\sum(y_i - \hat{y})^2}{n-2}} = \sqrt{\frac{\sum e^2}{n-2}}$$

The trend that fits the empirical values best is the one for which the residual standard deviation is the smallest.

- In the examples used above we fitted different trends to the quarterly sales data of a company across 3 years. *Determine which trend fits best.*

Moving average $s_e = \sqrt{\frac{\sum e^2}{n-2}} = \sqrt{\frac{1130.7968}{8-2}} = \mathbf{13.728}\ thousand\ pcs$

The pair by pair deviation of actual data and data calculated from the trued moving average is 13 728 pcs.

Linear trend $\boldsymbol{s_e = \sqrt{\frac{\sum e^2}{n-2}} = \sqrt{\frac{1737.462}{12-2}} = 13.181\ thousand\ pcs}$

On the basis of the linear trend, the average pair by pair deviation of actual and estimated data is 13 181 pcs.

Exponential trend

$$\boldsymbol{s_e = \sqrt{\frac{\sum e^2}{n-2}} = \sqrt{\frac{1809.242}{12-2}} = 13.451\ thousand\ pcs}$$

On the basis of the exponential trend, the average pair by pair deviation of actual and estimated values is 13 451 pcs.

Comparison of trends:

Linear trend		**Exponential trend**		**Moving average**
$s_e = 13.181\ 1000$ pcs	<	$s_e = 13.451\ 1000$ pcs	<	$s_e = 13.728\ 1000$ pcs

Based on the values of the residual standard deviation, the LINEAR TREND fits the data best.

11.3 SUMMARY AND QUESTIONS

11.3.1 Summary

We have discussed a wide variety of different tools to be used in examinations based on temporal attributes. The choice among time series analysis techniques may be affected by the kind of information we expect to gain about the time series data. Some methods quantify only the temporal changes in the data, but this unit has presented methods that can be used for making forecasts too.

Of the methods described, moving average is the most widely criticized. It is easy to average the data, but outliers pose serious problems, as they may have a considerable amount of distortive effect. Shortening time series also leads to inaccuracy, which, however, may be mitigated by increasing the number of empirical values, since the more data we have, the less is the loss of information caused by the shortening of the time series. We most commonly fit linear or exponential trends to the data, which can be carried out easily with computer programs. Considering seasonality is particularly important in data series where periodic undulation is apparent from the empirical data.

11.3.2 Self-test questions

What is the difference between average absolute and average relative change?

Which are the principal elements of the time series in deterministic time series analysis?

How can you characterize the trend?

What methods can we use to determine the trend?

How can you decide which trend fits best?

How can you express the linear trend in general form, and what do the parameters mean?

How can you characterize the effect of cyclic fluctuation?

How can you characterize seasonality?

What methods can be used to determine seasonality?

Why is it important to reckon with the effect of chance in the time series?

11.3.3 Practice tests

❁ Decide whether the following statements are true or false.

The trend that fits best is the trend for which the value of residual standard deviation is the smallest.	**T**
In the case of a linear trend and additive seasonality, the raw and corrected season indices are equal.	**F**
In the case of an exponential trend and multiplicative seasonality, the product of the raw and corrected season indices is 1.	**T**
In the case of a linear trend and additive seasonality, the sum of the raw and corrected seasonal deviations is 1.	**F**
In the case of an exponential trend and multiplicative seasonality, the raw and corrected seasonal deviations are equal.	**F**

❁ Choose the correct answer.

The equation of the linear trend that describes the annual turnover of a restaurant (HUF thousand) across 2000 and 2013 is $y=620+12 \cdot t$. You can draw the following inference on the basis of the trend:

– the turnover of the restaurant decreased in the period examined

- the turnover of the restaurant in 2000 was HUF 620 thousand
- the turnover of the restaurant in the period examined increased annually by an average of HUF 12 thousand

On the basis of a linear trend, a restaurant expects HUF 1 400 thousand for its 2014 postseason. We also know that the turnover in the postseason is 10% lower on average than the estimated value. According to the forecast, the turnover for the high season in 2014, calculated by considering seasonality, is

- HUF 1540 thousand
- HUF 1260 thousand
- HUF 1410 thousand

The equation of the linear trend that describes the quarterly changes in the production of a company (thousand pcs) across 2009 and 2013 is y=1200-50t. You can draw the following conclusion on the basis of the trend:

- the production of the company decreased annually by an average of 50 thousand pcs
- the production of the company decreased quarterly by an average of 50 thousand pcs
- the production of the company in 2013 was 1200 thousand pcs

According to forecasts based on a linear trend and additive seasonality, a company is expected to sell 6500 products in the 2nd quarter of 2015. It would be 6200 products on the basis of the linear trend. How big is the seasonality of the 2nd quarters?

- sales are 300 pcs lower on average in the 2nd quarters
- sales are 300 pcs higher on average in the 2nd quarters
- sales are higher on average by 4.82% in the 2nd quarters

We have the following information about the residual standard deviation of the trends on a company's revenue: the average pair by pair deviation of the actual and estimated values in the case of a linear trend is 625 pcs, in the case of the moving average it is 612 pcs, and in the case of an exponential trend it is 627 pcs. Which trend fits best?

- linear trend
- exponential trend
- moving average

A department store sells an average of 12.5% less products in its 3rd quarter. This means that

- the value of the seasonal deviation is 12.5%
- the value of the season index is 112.5
- the value of the season index is 87.5

12. DISPLAYING THE RESULTS OF STATISTICAL ANALYSIS: GRAPHICAL REPRESENTATIONS

12.1 GOALS AND COMPETENCIES

Graphical representation is the visual display of the results of statistical analysis. Diagrams make it easy to quickly acquire a picture of tendencies and relationships, though they may distort the numerical results and lead to false inferences. A diagram that observes the criteria of simplicity and clarity and is precisely designed statistically may offer a lot of information that relates to a problem examined in a concise form and may prevent the observer from getting lost in the labyrinth of numbers. Graphical representations are most commonly used in temporal analyses, but they are also helpful in demonstrating frequencies, distributions and relationships. Therefore, they are widely used in statistical examinations.

The purpose of this unit is to familiarize students with the possibilities of the application of graphical representations in statistics. Students will learn about how to represent graphically the results of computations they have acquired in the chapters above. Students will learn about the major types of diagrams, their characteristics and areas of application. They will acquire the major principles of designing statistical diagrams, which they will be able to use for displaying the results of their own analyses in a statistically precise manner.

12.2 TOPICS

Simplicity and clarity are top priority in graphical representations, since our goal is to illustrate the results of our analyses in a concise and informative manner. There is a variety of different types of diagrams to choose from, depending on the topic of our inquiry, but the general principles of diagram design are to be observed universally. The use of titles, indicating the source, and a purpose-oriented design are to be mentioned first in the context of requirements of content and form. Diagrams may be divided into two groups. In diagrams based on a coordinate system, we examine a phenomenon of interest along two related attributes. A vertical or horizontal bar chart, a line chart, and a scatter plot are primarily used for the display of relationships or comparisons and tendencies. Diagrams not based on a coordinate system, such as pie charts or cartograms, display information structured along a particular dimension.

12.2.1 Requirements on the content and form of graphical representations

The first thing to do in designing a graphical representation is to think about its purpose, i.e., what you intend to demonstrate. A diagram can depict proportional relationships and tendencies mainly, so you need to consider your purpose and the phenomenon examined when selecting a particular type of diagram. There is a range of different types of diagrams and charts to choose from, depending on the amount and order of magnitude of the data to be displayed and the purpose of the demonstration.

Diagrams ought to be simple and clear. It is not a good idea to include a large amount of data of various sorts in a single diagram, because it will be too complex and abstruse. What a diagram depicts should be clear at a glance. A good title, which is suggestive of the population and its properties in connection with a phenomenon of interest, is conducive to clarity and understanding. In presenting temporal data, for example, the period examined must be clearly specified in the title. In presenting comparative data, the attribute that serves as the basis of comparison must be indicated. If a diagram presents homogeneous data, the unit of measurement may be included either in the title, or along one of the axes. Another obligatory element of a diagram is the indication of its source. You must indicate where your data come from. In the case of absolute data, it suffices to identify the source and the year, as in *OECD (2014)*, for example. If the data come from your own research, you may specifically indicate that too. If your data are derived from someone else's primary data by conducting a mathematical operation on them, for example, then you will indicate it in a form similar to this: *My own calculation, based on OECD (2014).*

Concerning formal properties, it is not obligatory to label the axes or to include a legend. A legend is necessary only when you represent the temporal changes in a variety of different types of data, but it is redundant in the representation of changes in a single phenomenon. Axes, too, must only be labeled when their meaning is not clear otherwise, from the title, for example. When representing temporal data, it is not necessary to add "years", as its meaning is otherwise clear from the text attached to the axes. The segmentation of the axes is determined by the nature of the data to be displayed. The "zero point", for instance, does not always need to be 0, though you must make sure that it is not confusing for the audience. The choice of colors and shapes is a matter of taste, though that, too, is best when it accords well with the statistical content. *A good diagram is one that is clear and easy to understand without reading any accompanying text attached to it.* In what follows, we will discuss the *ap-*

plication of the general principles through examples of different types of diagrams and charts.

12.2.2 Types of diagrams

Diagrams vary according as they display information in a coordinate system on data compared from a perspective or they display non-comparative data in some other specific form. We can fit a function to deterministically related pairs of data, most commonly in the case of a trend or regression, conveniently presented in the form of a scatter plot.

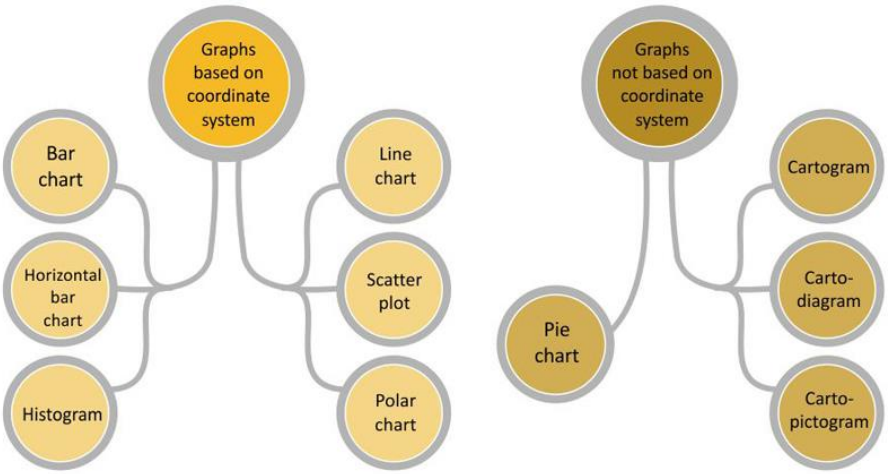


Figure 44 Types of diagrams

Vertical and horizontal bar charts

Vertical or horizontal bar charts[18] can be used to display absolute and relative frequency arrays. The height or length of bars represents frequency, so this type of diagram is suitable for any kind of comparison.

[18] A horizontal bar chart is a regular (vertical) bar chart rotated by 90 degrees.

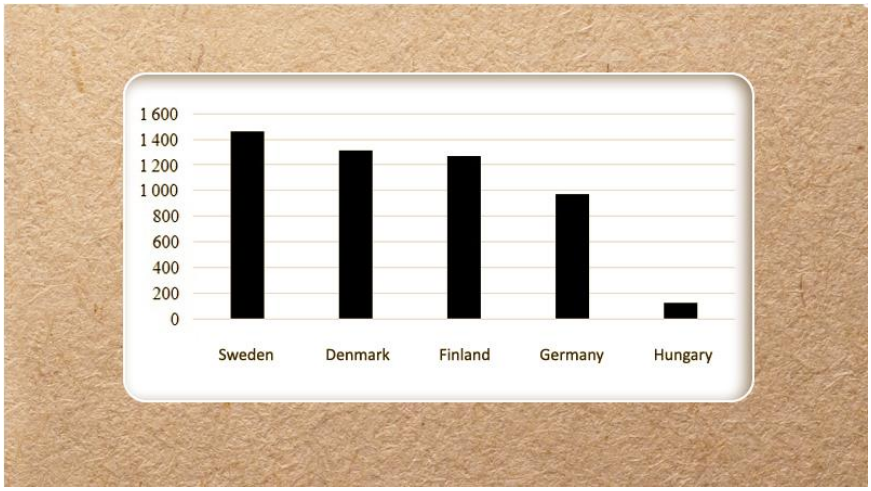


Figure 45 Per capita R&D expenditure data in Euros in some European countries in 2012

Source: *EUROSTAT* (2014)

This type of diagram can be used to display the frequency order of attribute variants. With time series data, however, the passage of time is decisive, so the direction of the changes will be clearly displayed.

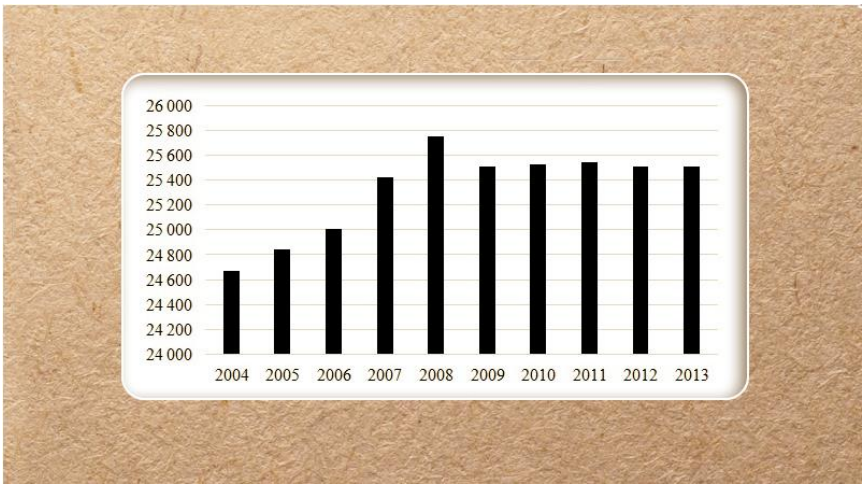


Figure 46 Changes in the number of employed of ages 15 to 64 (thousand prsns) in France across 2004 and 2013

Source: *EUROSTAT* (2014)

In figure 46, the starting point is not 0, which is why it better depicts the tendency in the temporal change of data within a narrow range. You need to be careful in your choice of the starting point, however, because the diagram may be misleading in magnifying otherwise insignificant differences.

You can display several data sets arranged according to the same attribute in one bar chart in a multi-dimensional examination, but in this case the legend is an obligatory component of the diagram, as this will clarify what sets of data are being compared. This must also be indicated in the title. Stacked bar charts are useful in displaying a comparison of distributions with or without regard to the number of elements in the populations. You can compare, e.g., changes in the qualification of employees of a company across periods of time by emphasizing only their distribution. A 100% stacked bar chart is the ideal choice for this, in which the height (or width) of a bar cannot exceed 100%. The diagram may also be designed to represent changes in the number of employees too.

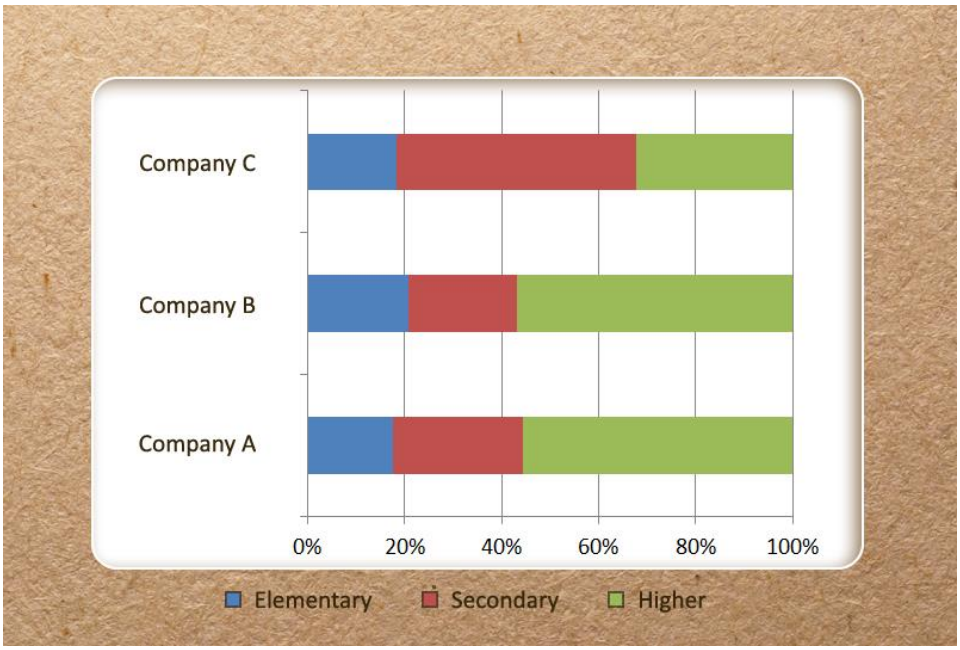


Figure 47 Distribution of employees by qualification in three companies

Source: *fictitious data*

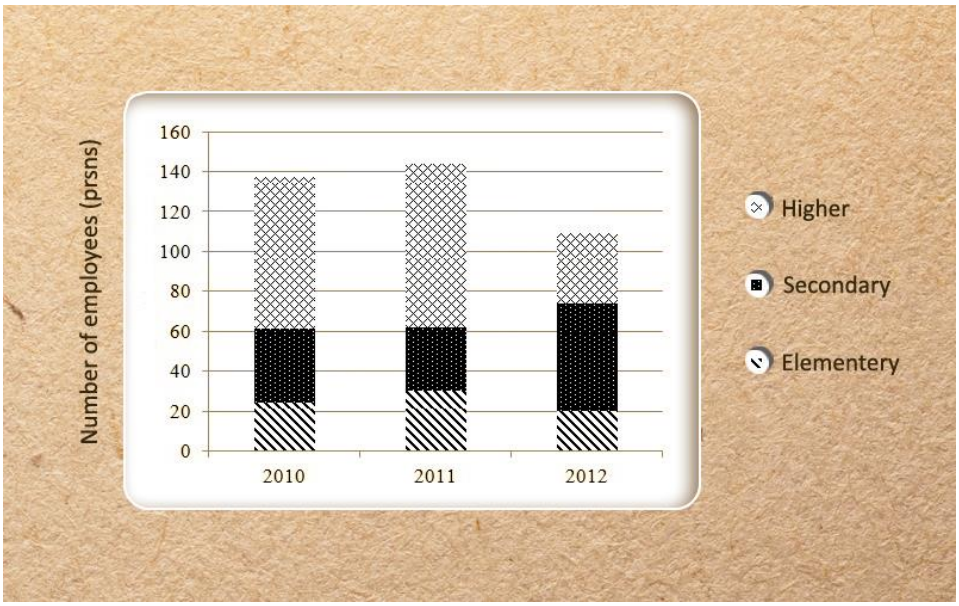


Figure 48 Changes in the composition of employees by qualification in a company across 2010 and 2012

Source: *fictitious data*

A special variant of the horizontal bar chart is the population pyramid, used in population statistics. This type of diagram is used to display the distribution of men and women by age group. It is a complex type of diagram, with the vertical axis positioned in the middle, representing age, and the number of elements in each age group, i.e. their frequency, on either side of the vertical axis. A population pyramid represents the age composition of a population by gender. You can read more about its types on the following web site:

24. Population pyramid: http://www.ehow.com/list_6872808_three-different-types-population-pyramids.html

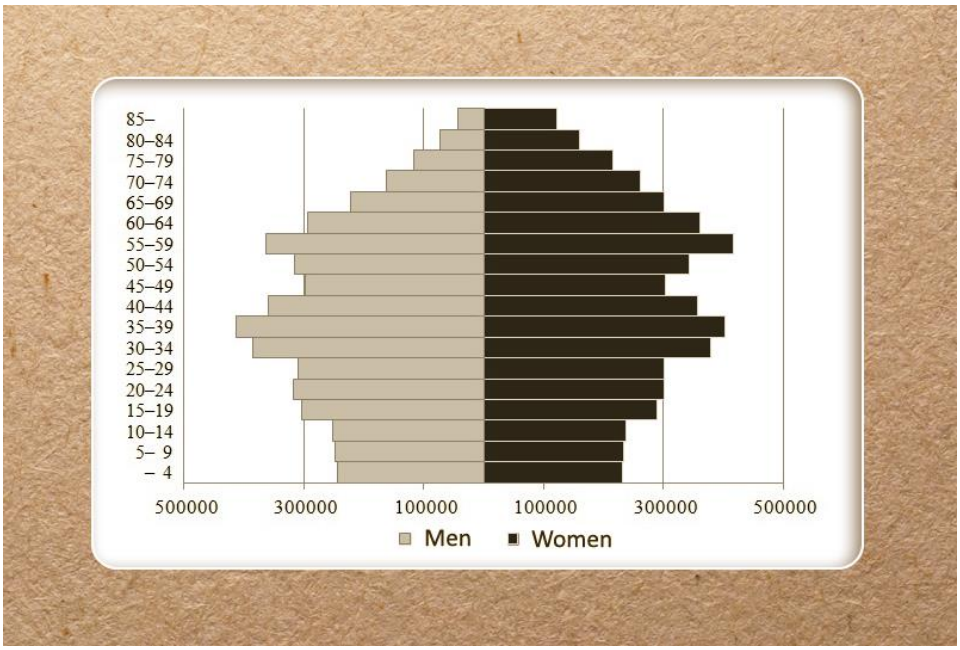


Figure 49 Population pyramid of the Hungarian population in 2011

Source: *HCSO* (2014)

Histograms

A histogram is a special bar chart, used to visually represent (primarily class interval) frequency arrays. In the case of a population with a large number of elements, non-overlapping intervals of the same or different lengths can be constructed, into which each element may be unambiguously assigned.[19] When we design a histogram, we construct a bar chart without spaces between the bars.

[19] For more about class interval frequency arrays, see section 2.2.2 Organizing statistical data.

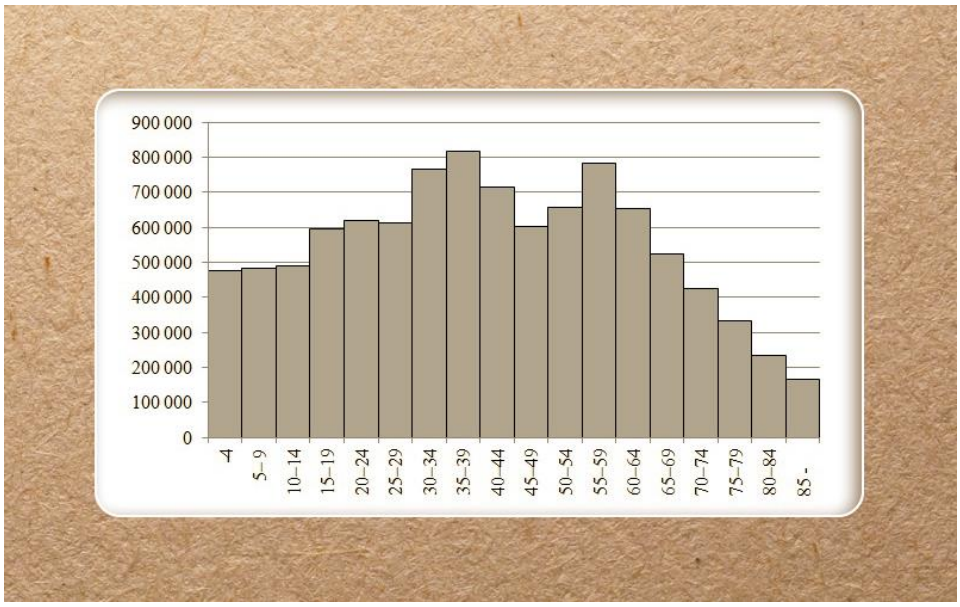


Figure 50 Hungary's population by age groups in 2011-ben

Source: *HCSO* (2014)

A histogram conveys information about the frequency of classes and the distribution of the data set too. We specify the classes on the horizontal axis, and the vertical axis contains the absolute or relative frequency values. The frequency representations of the diagram also give you a picture of the shape of the distribution function, so you can tell whether it is symmetrical or asymmetrical, and have some idea of its curtosis, which is quantified by the shape indices of descriptive statistics.

Line charts

Line charts are most commonly used in time series analysis where the time series data are non-discrete values. This type of diagram is also a good choice when we want to represent frequency arrays based on a continuous quantitative attribute. A line represents continuity.

This type of diagram is the perfect choice for the visual representation of tendencies. The direction of changes is clearly displayed, and so are the peaks and valleys in the period examined. By including primary and secondary axes additional data, perhaps of different orders of magnitude, can also be represented, although this will add to the complexity of the diagram and might make it a little more difficult to interpret. The repre-

sentation of temporal changes may involve several data sets characterized along a common attribute.

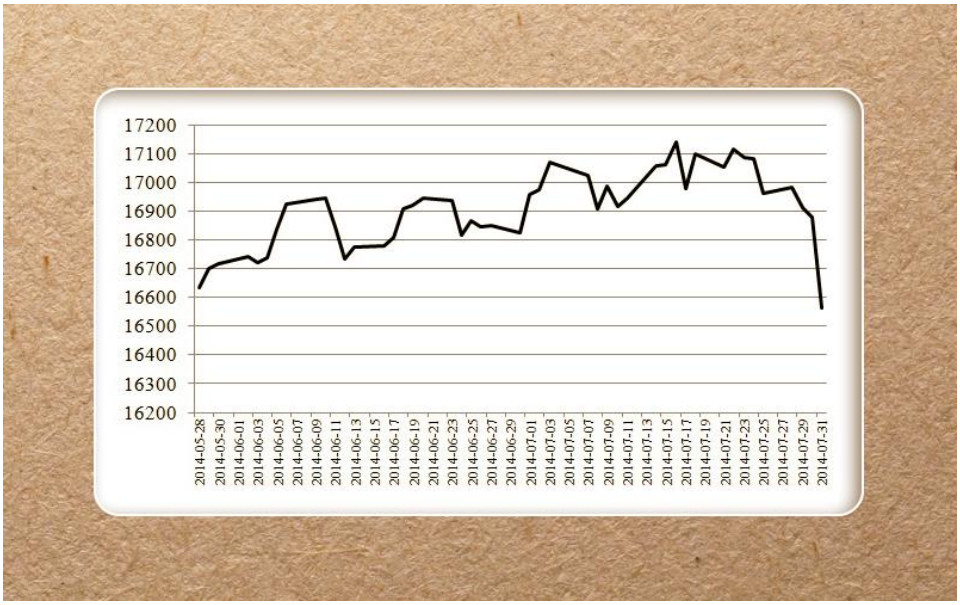


Figure 51 Changes in the DOW Jones index across May 28 and July 31, 2014

Source: *FRED* (2014)

Scatter plots

Scatter plots are most commonly used in the examination of relationships between quantitative attributes when we have discrete quantitative values. An important criterion of this type of diagram is that transitions between values are uninterpretable. That is possible only for continuous attributes, but for them, we would use a line chart instead. In the case of two attributes, the axes need to be labeled and the unit of measurement must also be indicated.

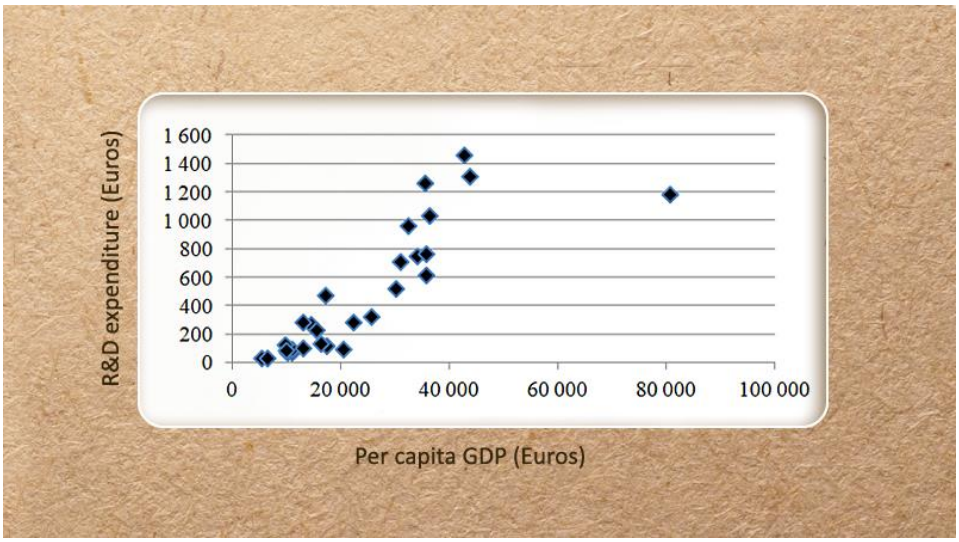


Figure 52 *Relationship between per capita GDP and R&D expenditure in the member states of the European Union*

Source: *EUROSTAT* (2014)

The scatter plot is suggestive of the shape of the function that best fits the associated pairs of values, i.e. the type of regression function that can be fitted to the data. Note that the diagram above is also indicative of the relevant outliers. This type of diagram allows the visual representation of relationships between data of different orders of magnitude.

Polar charts or polar curves

A polar chart is a type of diagram which can be constructed in a polar coordinate system, which allows us to study a particular phenomenon from various perspectives. The shape of the polygon is determined by the number of features examined. The 0 point is positioned in the center of the diagram, and the features and values examined appear at the vertices of the polygon. This multidimensional examination allows comparisons of a temporal and qualitative nature too.

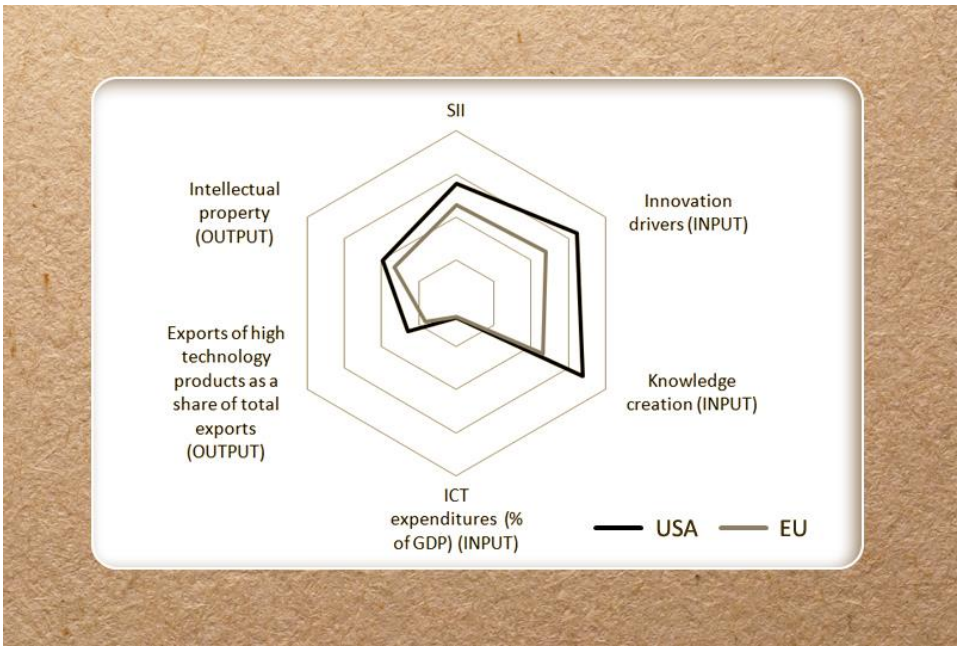


Figure 53 Performance of Europe and the USA in areas of the Summary Innovation Index

Source: *EC – Union Innovation Scoreboard* (2012)

Figure 53 displays a comparison of Europe and the USA in selected areas of the Summary Innovation Index (SII). We might specify the units of measurement of the data too, but this type of diagram is primarily used for the comparison of products in interrelated areas relevant to a phenomenon of interest, where the units of measurement of the base data are not particularly important.

Pie charts

A pie chart is a type of diagram that can be constructed on the basis of classifying rows, and is used to display distribution. Some types of computer software give you an option of the pretzel chart, which also allows you to compare the distributions of populations of identical elements too. For these types of diagrams, clarity and the identification of particular parts are especially important. It is a good idea to indicate the distribution percentages in particular parts, which makes it easier to interpret the diagram, especially if the proportions are very clearly visible in the diagram. Such diagrams are not suitable for the visual representation

of populations composed of a large number of components, as the legend may become undecipherable.

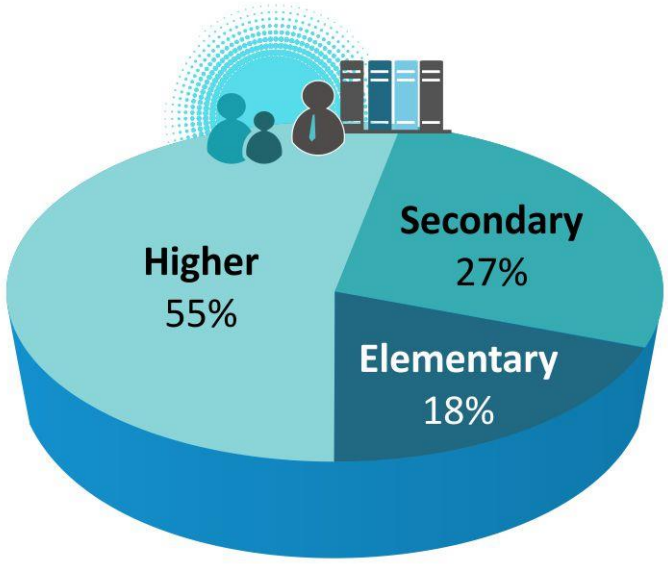


Figure 54 Distribution of employees at a company by qualification

Source: *fictitious data*

Maps as graphical tools in statistics

A map is helpful in the graphical representation of areal comparative data, as it offers a clear and easy-to-read picture with information on different geographical units. Maps are most commonly used in demographic and socio-geographic analyses, but they are also a good choice in special areal examinations.

A **cartogram** is a type of diagram in which different colors and patterns represent the intensity of a particular phenomenon of interest in different geographical areas, such as, for example, population density.

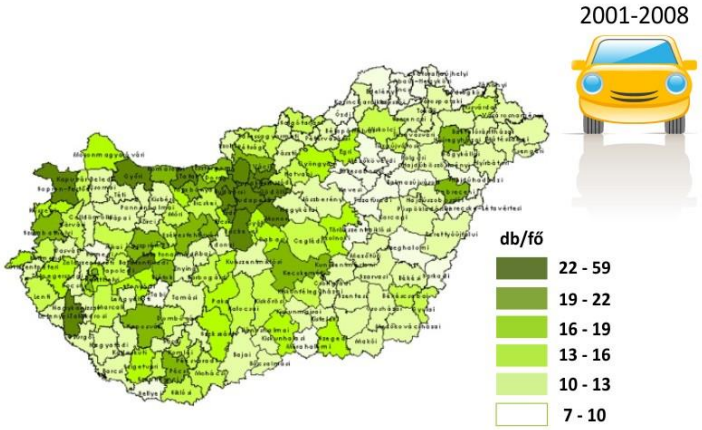


Figure 55 Number of new motorcars per 100 residents registered across 2001 and 2008 in Hungary

Source: *TeIR HCSO T-STAR and census databases*

A **carto-diagram** offers more information than a cartogram, as it represents not only the intensity of a particular phenomenon of interest but it offers additional diagrammatic information on a geographical area in terms of a different attribute.

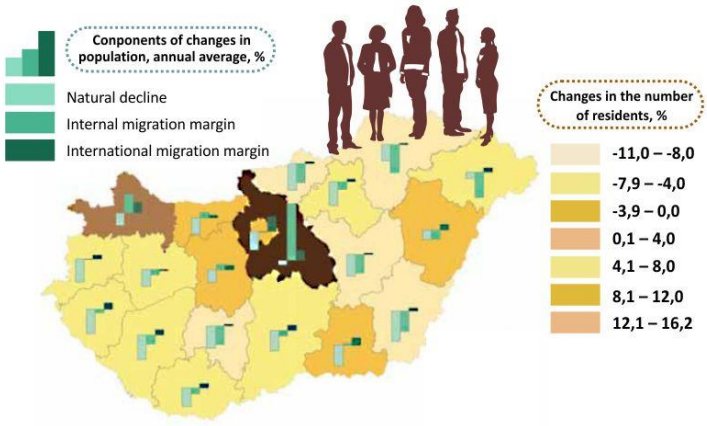


Figure 56 Changes in the number of residents and their composition in Hungary across 2001–2011

Source: *HCSO – Population atlas of Hungary (2012)*

The most commonly used types of diagrams used to supplement a cartogram are pie charts, which represent distribution, and bar charts, which are employed in the examination of temporal changes.

A **carto-pictogram** is a very special means of graphical representation. When a phenomenon of interest can be denoted by an icon, its geographical intensity may be nicely suggested by changes in its size.

Figure 57 Tick-infected areas in Hungary–photograph (June 15, 2009)
Source: *Alfahír Hírportál* (2012)

12.3 SUMMARY AND QUESTIONS

12.3.1 Summary

Results of statistical analyses can be graphically represented in a variety of different forms. Observing the general principles of simplicity, clarity and a well-defined purpose, you can design diagrams which clearly depict a tendency and relationships revealed in a data set or visually represent a comparison.

Diagrams must have a title and the source needs to be indicated too. Sometimes you need to specify the units of measurement and you may need to attach a legend. The choice of the type of diagram depends on

the purpose of presentation. Comparative statistical arrays are best depicted in vertical or horizontal bar charts, with classifying rows we most commonly use pie charts, line charts are ideal for the analysis of continuous temporal changes, and a scatter plot is the best choice for representing relationships. Maps are useful in areal comparison.

12.3.2 Self-test questions

Which are the major principles of graphical representation?

Which are the obligatory elements of all diagrams?

When do you need to add a legend?

What types of data are best displayed in a bar chart?

What is a population pyramid?

What is a histogram?

What types of data may be depicted in a line chart?

What types of data may be depicted in a scatter plot?

What types of data may be depicted in a pie chart?

What is the difference between a cartogram and a carto-diagram.

12.3.3 Practice tests

Decide whether the following statements are true (T) or false (F).

A bar chart or a line chart is best for depicting temporal comparison.	**T**
Scatter plots are most commonly used in relationship examinations, as the dots depict a curve that best fits the pairs of associated values.	**T**
A population pyramid is a type of bar chart used in population statistics.	**F**
A carto-diagram uses different colors to represent the areal intensity of the phenomenon examined.	**F**
A histogram is a special sort of bar chart with no space between the bars, which is best for the representation of class interval frequencies.	**T**
A cartogram represents changes in the phenomenon examined by diagrams in different colors.	**F**

A polar chart can represent comparison between data sets along several different dimensions.	**T**
A cartogram represents differences in the areal intensity of the phenomenon examined by different shades of colors.	**T**
The title, the source, and the legend are obligatory elements of diagrams.	**F**
In diagrams based on a coordinate system, you can define the relative position of the features examined.	**T**

13. SUMMARY

13.1 SUMMARY OF CONTENT

Statistical analyses play an important role in the examination of social and economic phenomena. Familiarity with the basic concepts, such as population, attribute, and measurement scales, is a prerequisite for the application of methods. Statistical work begins with gathering data. Data may come from primary or secondary sources. Secondary data may be collected from various databases, whose credibility is critical, therefore it makes sense to use data obtained from international organizations. In gathering primary data we obtain information about a particular phenomenon by doing our own research, by conducting a questionnaire survey, for example. Data collection may target an entire population, but most frequently we can only characterize a part of a whole population by examining a sample. Elements of a sample may be selected by various sampling techniques. In random sampling, you can specify in advance the probability with which particular elements will be selected in the sample, while in systematic sampling you specify the sampling criteria in advance, as the kind of elements that will be included in the sample will play a decisive role in later analysis. Representativeness of samples is of primary importance, as this ensures that you can draw general inferences from partial data collection. Data collected may be organized in statistical rows and the rows may be organized into statistical tables, which must be supplied with a title and the source of data.

Ratios and descriptive statistics may be regarded as simple analytic tools. Ratios, which may be derived as quotients of mutually related data, are present in all areas of life and can be easily recognized. The internal structure of a population or a sample may be characterized by distribution or coordination ratios, and comparisons may be carried out with dynamic ratios, plan accomplishment, plan task, or areal comparative ratios. Intensity ratios allow the comparison of different types of data. Ratios are used in labor-market statistics, as well as financial and population statistics. We can concisely characterize data sets with the apparatus of descriptive statistics along quantitative attributes. Central tendencies – means, mode, quantiles – characterize the typical values of a data set; indices of dispersion – range, standard deviation, relative standard deviation, mean difference and mean deviation – characterize the variation of data, and shape indices – asymmetry and curtosis indices – characterize the shape of the frequency curve.

Heterogeneous populations offer a wider range of possibilities of analysis. When applying standardization heterogeneity is handled by clustering,

which allows us to make qualitative or areal comparisons with respect to the entire population. In decomposing differences the difference between complex ratios characterizing the populations compared is reduced to two factors: differences between cluster ratios of clusters and differences in the cluster-composition of populations. The method of index calculation based on standardization can be used for temporal comparison. Changes in time are in part caused by changes in the cluster ratios and in part by changes in cluster composition. By introducing the statistical concept of value, we can analyze directly incommensurable heterogeneous data sets. With the help of value, defined as the product of quantity and price, we can determine changes in prices, quantities and value for one or more products. The method of index calculation is used in quantifying corporate and national economic performance and in external trade statistics.

An examination of relationships between attributes is especially important, because we often want to know about the particular factors that affect the changes of some phenomena as well as the roles they play. Stochastic relationships are distinguished in terms of the attributes that are interrelated. Association can be used to reveal the presence and strength of a relationship between two quantitative attributes. In studying mixed relationships (analysis of variance) we are interested to know if there is a relationship between a quantitative and a non-quantitative attribute, how strong the relationship is, and we can also quantify the degree to which the non-quantitative attribute determines changes in the quantitative attribute. In correlation calculation, we can measure the presence, strength, and direction of a relationship between two quantitative attributes. Given that they are quantitative attributes, the nature of the relationship can be described in mathematical terms. This is called regression calculation. Regression functions represent a method for the detection of an assumed cause-and-effect relationship between attributes.

We can use a variety of different devices in the examination of time series. Dynamic ratios, averages, or indices can quantify only changes in the observed data. But there are techniques in time series analysis which allow us to make forecasts too. The essence of deterministic time series analysis is to decompose time series into component factors and quantify the effects of those factors. Components are distinguished in terms of the length of the period of time they affect changes in. Elements of a time series may be related in an additive or multiplicative manner. The trend is a long-term tendency, which can be represented by moving average calculation or with the help of a linear or exponential trend. We can reveal seasonality, short-term periodic undulation, in absolute or relative form. Seasonal deviations express deviations from the trend in units of measurement of the base data,

while season indices express such deviations in percentage. Mid-term business cycle effects play an important role in economic processes. An indispensable element of time series is the effect of chance, which may be determined on the basis of the residue principle. For forecasts, we minimally need to know the trend and seasonality.

Diagrams can represent the results of statistical analyses graphically. Simplicity and clarity are top priority general principles of diagram design. Important requirements of content and form include the obligatory presence of a title, the source, and units of measurement. Different types of diagrams offer different possibilities for analysis, therefore the presentation of data must be purpose-oriented. Horizontal or vertical bar charts, line and polar charts can be used primarily for comparison, scatter plots may be used to display relationships, while pie charts are used for the visual representation of structural properties. Maps may be used to represent areal comparisons.

13.2 PRACTICAL APPLICABILITY OF STATISTICS AS A SCIENTIFIC METHOD

Statistics is both a scientific method and a practical activity. The use of methods presupposes data, so accurate data collection is just as important as the professional use of analytic tools. Sometimes you need to collect and interpret your own data, because there are a number of different phenomena for which publically available data do not exist. Statistics is not only used in scientific research, but it may be useful in the analysis of a range of different processes in everyday life. We regularly encounter problems in various fields of life which we can resolve only by quantifying data and by employing statistical tools in their analysis. Sometimes we do not realize that we are in fact using statistical tools, as in conversations about average wages or students' academic achievements, the ratio of boys and girls in schools or the cinema, or when we speculate about the amount of money we spend in shops. This course has attempted to present various statistical methods through practical examples, which may help students understand that statistics is with us in our daily lives and by the acquisition of skills in the use of analytic tools they have attained knowledge that they may find useful in practice.

14. APPENDICES

14.1 BIBLIOGRAPHY

Books

BREALEY, R. A. – MYERS, S. C.:*Principles of Corporate Finance*. Fourth Edition. McGraw Hill, 1991.

HOLLÓNÉ KACSÓ ERZSÉBET: *Vállalatértékelés*. Mutatók, modellek, látható- és láthatatlan vagyonértékelési eljárások. Főiskolai jegyzet, Eger, Eszterházy Károly Főiskola, 2011.

ILYÉSNÉ MOLNÁR EMESE – LOVASNÉ AVATÓ JUDIT: *Statisztika feladatgyűjtemény I.* Budapest, Perfekt Kiadó, 2006.

ILYÉSNÉ MOLNÁR EMESE – LOVASNÉ AVATÓ JUDIT: *Statisztika feladatgyűjtemény II.* Budapest, Perfekt Kiadó, 2006.

KERÉKGYÁRTÓ GYÖRGYNÉ – MUNDRUCZÓ GYÖRGY – SUGÁR ANDRÁS: *Statisztikai módszerek és alkalmazásuk a gazdasági, üzleti elemzésekben*. Budapest, Aula Kiadó, 2001.

KORPÁS ATTILÁNÉ DR. (szerk.): *Általános statisztika I.* Budapest, Nemzeti Tankönyvvkiadó, 1996.

KORPÁS ATTILÁNÉ DR. (szerk.): *Általános statisztika II.* Budapest, Nemzeti Tankönyvvkiadó, 1997.

SÁNDORNÉ DR. KRISZT ÉVA – DR. CSESZNÁK ANITA – ORSZÁG GÁBORNÉ: *Statisztika I.* Budapest, Nemzedékek tudása Tankönyvkiadó, 2013.

SOLT KATALIN: *Makroökonómia.* Tatabánya, TRI-Mester, 2001.

SZŰCS ISTVÁN (szerk): *Alkalmazott statisztika*. Budapest, AGROINFORM Kiadó és Nyomda Kft, 2002.

Electronic documents / sources

EURÓPAI BIZOTTSÁG: *Eurostat* [adatbázis] [2014. július] <URL: http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/>

EUROPEAN COMMISSION: *Union Innovation Scoreboard* [adatbázis] [2014. június] <URL: http://ec.europa.eu/enterprise/policies/innovation/policy/innovation-scoreboard/index_en.htm>

FEDERAL RESERVE BANK OF ST. LOUIS (FRED) Economic Research: *Dow Jones Industrial Average* (DJIA) [2014. augusztus 2.] <URL: http://research.stlouisfed.org/fred2/series/DJIA/downloaddata>

KÖZPONTI STATISZTIKAI HIVATAL (HCSO) *statisztikái* [adatbázis] [2014. július] <URL: http://www.ksh.hu/>

KÖZPONTI STATISZTIKAI HIVATAL: *Magyarország társadalmi atlasza*. Budapest, 2012. [elektronikus dokumentum] 2014. augusztus 3.] <URL: https://www.ksh.hu/docs/hun/xftp/idoszaki/pdf/tarsatlasz.pdf>
KÖZPONTI STATISZTIKAI HIVATAL: *Áralakulás*. [elektronikus dokumentum] 2014. augusztus 19.] < URL: https://www.ksh.hu/thm/1/indi1_2_3.html>
ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD) *statisztikái* [adatbázis] [2014. július] <URL: http://stats.oecd.org/>
TERÜLETI ELEKTRONIKUS INFORMÁCIÓS RENDSZER (TEIR) KSH T-STAR és Népszámlálás [adatbázis] [2014. augusztus 3.] <URL: http://www.terport.hu/kartogramok/a-szemelygepkocsi-allomany-1992-es-2008-kozotti-ill-a-magyarorszagon-elso-alkalommal-for>
VILÁGBANK *statisztikái* [adatbázis] [2014. július] <URL: http://data.worldbank.org/>
ALFAHÍR HÍRPORTÁL: Magyarország kullancsfertőzött helyei – fotó (2009. június 15) [2014. augusztus 3.] <URL: http://alfahir.hu/node/31010>

14.2 SUMMARY OF MEDIA ELEMENTS

14.2.1 List of tables

14.2.2 List of diagrams

14.2.3 External URL references

8. United Nations Conference on Trade and Development: http://unctad.org/en/Pages/Publications/Handbook-of-Statistics.aspx.....14
9. Web site of the European Central Bank: http://www.ecb.europa.eu/stats/exchange/eurofxref/html/index.en.html 46
10. P/E: http://www.investopedia.com/terms/p/price-earningsratio.asp 46
11. ROE: http://www.investopedia.com/terms/r/returnonequity.asp 47
12. ROA: http://www.investopedia.com/terms/r/returnonassets.asp ...47
13. ROS: http://www.investopedia.com/terms/r/ros.asp.....47
14. Financial indicators: http://kfknowledgebank.kaplan.co.uk/KFKB/Wiki%20Pages/Financial%20Performance%20Indicators%20%28FPIs%29.aspx ..48
15. World Bank demographic indices: http://econ.worldbank.org/WBSITE/EXTERNAL/DATASTATISTICS/0,,contentMDK:20451597~pagePK:64133150~piPK:64133175~theSitePK:239419,00.html 48
16. The relationship between GDP and inflation: http://www.investopedia.com/articles/06/gdpinflation.asp 101
17. Inflation, deflation, and stagflation: http://www.investopedia.com/exam-guide/cfp/economics-time-value/cfp6.asp 103
18. World Trade Organization: http://www.wto.org/english/res_e/statis_e/statis_e.htm 103
19. Business cycles: http://www.britannica.com/EBchecked/topic/86233/business-cycle 147
20. Kitchin cycle (3-5 years): http://www.policonomics.com/joseph-kitchin/ 147
21. Juglar cycle (7-11years): http://www.policonomics.com/clement-juglar/ 147
22. Kuznets cycle (15-25 years): http://www.dictionaryofeconomics.com/article?id=pde2008_K000045 147
23. Kondratyev cycle (45-60 years): http://www.newworldencyclopedia.org/entry/Nikolai_Kondratiev 147
24. Population pyramid: http://www.ehow.com/list_6872808_three-different-types-population-pyramids.html 179

15. TESTS

PRACTICE TESTS

1. You are conducting a survey on coffee consumption habits. Which sampling technique would you use for gathering data? *More than one answer may be correct.*

Random		Nonrandom	
X	Independent equivalent distribution		Quota
X	Simple random	X	Snowball
	Stratified		Purposive
	Cluster		Systematic
	Multi-stage		Arbitrary

2. You are conducting a survey on television viewing habits. Which sampling technique would you use for gathering data? *More than one answer may be correct.*

Random		Nonrandom	
	Independent equivalent distribution		Quota
X	Simple random	X	Snowball
	Stratified		Purposive
	Cluster	X	Systematic
	Multi-stage		Arbitrary

3. The representativeness of a sample is important for reasons of drawing general conclusions.

TRUE FALSE

4. A classifying table contains only classifying rows.

TRUE **FALSE**

5. EURO's rate of exchange is 1.35 EUR/USD. What type of ratio is this?

	Derived from classifying row		Derived from comparative row
	Distribution		Dynamic
	Coordination		Areal comparative
	Derived from descriptive row		Plan task
X	Intensity		Plan accomplishment

6. A company's revenues increased by 30% compared to last year. What type of ratio is this?

	Derived from classifying row		Derived from comparative row
	Distribution	X	Dynamic
	Coordination		Areal comparative
	Derived from descriptive row		Plan task
	Intensity		Plan accomplishment

7. 30% of a store's sales revenues come from selling trousers. What type of ratio is this?

	Derived from classifying row		Derived from comparative row
X	Distribution		Dynamic
	Coordination		Areal comparative
	Derived from descriptive row		Plan task
	Intensity		Plan accomplishment

8. Per capita average annual chocolate consumption in 2010 was 28 kg. What type of ratio is this?

	Derived from classifying row		Derived from comparative row
	Distribution		Dynamic
	Coordination		Areal comparative
	Derived from descriptive row		Plan task
X	Intensity		Plan accomplishment

9. The total human population can be divided into economically active and inactive groups from the perspective of the labor market.

TRUE **FALSE**

10. The employment rate represents the ratio of employed within the total population.

TRUE **FALSE**

11. When calculating population density, we compare the population of a country to the area of the country.

TRUE FALSE

12. P/E is an indicator which compares the earnings after tax to equity.

TRUE **FALSE**

13. 100 customers were observed on a particular day doing shopping in a hypermarket. Half of the customers spent over € 25 and half spent less than that. € 25 is the value of *which basic distribution descriptor?*

Means

	Mean
	Mode
X	Median
	Lower quartile
	Upper quartile

Dispersion and shape indices

	Range
	Standard deviation
	Relative standard devn.
	A index
	F index

14. 100 customers were observed on a particular day doing shopping in a hypermarket. Customers spent € 24 on average. € 25 is the value of *which basic distribution descriptor?*

Means

X	Mean
	Mode
	Median
	Lower quartile
	Upper quartile

Dispersion and shape indices

	Range
	Standard deviation
	Relative standard devn.
	A index
	F index

15. Mode is sensitive to outstandingly low or outstandingly high values in the data set.

TRUE **FALSE**

16. Quantiles are special data values that mark the boundaries in a data set.

TRUE FALSE

17. An overview of the wages of the employees of a company shows that men make HUF 30 000 more on average than women. We know that due to the difference in the distribution of professional qualifications among men and women, men would make only HUF 24 000 more than women. Due to the difference in average wages between men and women by professional qualification

– men would make HUF 6000 less than women

- men would make HUF 6000 more than women
- women would make HUF 6000 more than men

18. When calculating a temporally comparative index, the direction of the comparison can be chosen freely.

TRUE **FALSE**

19. The grand mean index is the power of the cluster mean and the comparative index.

TRUE FALSE

20. An overview of data on two resort areas shows that the average amount of time spent by visitors has dropped by 10% compared to last year. On average, the time spent by visitors from Austria, Germany, and Poland has increased by 2%, but the yearly distribution of nationalities has changed so much that it has caused an average drop of 11.76% in overall time spent by visitors.
    - the value of the grand mean index is 110.0
    - the value of the cluster mean index is 102.0
    - the value of the composition-effect index is 111.76

21. Which of these does not statistically belong in the category of value?

turnover **selling price** sales receipts gross production

22. If the value of the base weighted price index is 107.82, and the value of the chain weighted price index is 109.12, then the value of the Fisher index is:

117.65 106. 48 **108.47**

23. Fisher indices are computed as the geometric means of base and reference weighted indices.

TRUE FALSE

24. The elementary price index is the quotient of the elementary volume index and the elementary value index.

TRUE **FALSE**

25. Productivity shows
    - the number of workers per unit of products
    - the number of products per worker
    - the average amount of products produced

26. The value index of GDP is
    - the change in nominal GDP
    - the change in real GDP
    - the quotient of the nominal and real GDP

27. In computing the consumer price index used to measure inflation, the weight is
    - quantities of the reference year
    - quantities of the previous year
    - prices of the previous year
28. Terms of trade is
    - the quotient of the price indices of exported and imported goods
    - the quotient of the priced indices of imported and exported goods
    - the quotient of export and import
29. What index can you use to quantify the relationship between students' major subjects and their scores on a test?

    Yule's Tschuprov's **Standard deviation quotient**
30. An examination of the relationship between employees' qualifications and wages delivered the following value for the quotient of variance: 0.674. What does it mean?
    - there is a stronger than moderate relationship between the two attributes
    - qualification determines changes in wages to a degree of 67.4%
    - wages by qualification of employees deviate from the grand mean by an average of 0.674
31. The value of the H index between type of train and number of delays is 0.26. What inference can you draw from that?
    - a.type of train determines the number of delays to a degree of 0.26%
    - b.there is a weak relationship between the type of train and the number of delays
    - c.type of train determines the number of delays to a degree of 26%
32. The value of the Tschuprov index between the position and gender of employees is 0.28. How strong is the relationship?

    Weak Weaker than moderate Strong
33. Covariance may assume a value between -1 and 1, and its sign denotes the direction of the relationship.

    TRUE **FALSE**

34. Spearman's coefficient determines the strength of the relationship between attributes measurable on a ratio scale.

TRUE **FALSE**

35. r2 = 0.52 means that the explanatory variable affects changes in the result variable to an extent of 0.52%.

TRUE **FALSE**

36. A condition on multivariate regression models is that the explanatory variable and the result variable must be independent of each other.

TRUE **FALSE**

37. The equation of the linear trend that describes the annual turnover of a restaurant (HUF thousand) across 2000 and 2013 is y=620+12□t. You can draw the following inference on the basis of the trend:
    - the turnover of the restaurant decreased in the period examined
    - the turnover of the restaurant in 2000 was HUF 620 thousand
    - the turnover of the restaurant in the period examined increased annually by an average of HUF 12 thousand

38. The trend that fits best is the trend for which the value of residual standard deviation is the smallest.

TRUE FALSE

39. According to forecasts based on a linear trend and additive seasonality, a company is expected to sell 6500 products in the 2nd quarter of 2015. It would be 6200 products on the basis of the linear trend. How big is the seasonality of the 2nd quarters?
    - sales are 300 pcs lower on average in the 2nd quarters
    - sales are 300 pcs higher on average in the 2nd quarters
    - sales are higher on average by 4.82% in the 2nd quarters

40. In the case of a linear trend and additive seasonality, the raw and corrected season indices are equal.

TRUE **FALSE**

41. A bar chart or a line chart is best for depicting temporal comparison.

TRUE FALSE

42. Scatter plots are most commonly used in relationship examinations, as the dots depict a curve that best fits the pairs of associated values.

TRUE FALSE

43. A population pyramid is a type of bar chart used in population statistics.

TRUE **FALSE**

44. A carto-diagram uses different colors to represent the areal intensity of the phenomenon examined.

TRUE **FALSE**

MOCK EXAMINATION

1. You are conducting a survey on salary satisfaction of company employees. Which sampling technique would you use for gathering data? *More than one answer may be correct.*

Random		**Nonrandom**	
	Independent equivalent distribution	**X**	Quota
	Simple random		Snowball
X	Stratified		Purposive
	Cluster		Systematic
	Multi-stage		Arbitrary

2. The number of dimensions of a statistical table shows the number of attributes to which a particular value belongs.

TRUE FALSE

3. Thanks to exceptional achievements last year, the company expects a 10% increase in revenues next year. What type of ratio is this?

Derived from classifying row		**Derived from comparative row**	
	Distribution		Dynamic
	Coordination		Areal comparative
Derived from descriptive row		**X**	Plan task
	Intensity		Plan accomplishment

4. Per capita GDP in the USA is nearly five times that of Hungary. What type of ratio is this?

Derived from classifying row		Derived from comparative row	
	Distribution		Dynamic
	Coordination	X	Areal comparative
Derived from descriptive row			Plan task
	Intensity		Plan accomplishment

5. A person is economically inactive if they do not take part in doing socially organized work.

TRUE FALSE

6. In determining the healthcare coverage index, we compare the number of residents to the number of doctors.

TRUE **FALSE**

7. 100 customers were observed on a particular day doing shopping in a hypermarket. The amount of money spent by individual customers deviates from the average amount of money spent by € 4 on average. € 4 is the value of which basic distribution descriptor?

Means		Dispersion and shape indices	
	Mean		Range
	Mode	X	Standard deviation
	Median		Relative standard devn.
	Lower quartile		A index
	Upper quartile		F index

8. Standard deviation expresses the average deviation between data in terms of the unit of measurement of the base data, while relative standard deviation expresses the same in percentage.

TRUE **FALSE**

9. An overview of data on two resort areas shows that the average amount of time spent by visitors has dropped by 10% compared to last year. On average, the time spent by visitors from Austria, Germany, and Poland has increased by 2%, but the yearly distribution of nationalities has changed so much that it has caused an average drop of 11.76% in overall time spent by visitors.
   - the value of the grand mean index is 90.0
   - the value of the cluster mean index is 98.0
   - the value of the composition-effect index is 111.76

10. If the composition of two populations compared is identical, then the effect of the difference in composition by subpopulations is 1.

TRUE **FALSE**

11. The value of the elementary price index is 100, and the value of the elementary volume index is 110.
The value of the elementary value index is:
   - 110
   - 90.1
   - 100

12. For aggregate indices, the following relationship holds: the aggregate value index is the product of the inversely weighted price and volume indices.

TRUE FALSE

13. The specific use of materials is
   - the quotient of the total materials used and the products produced
   - the product of the total materials used and the products produced
   - the quotient of the products produced and the total materials used

14. Nominal GDP is
   - the product of quantities and prices of a particular year
   - the product of the prices of a particular year and a quantity typical of previous years
   - the product of the quantities of a particular year and the prices typical of previous years

15. What index can you use to quantify the relationship between students' success rate in an exam (pass or fail) and their place of residence (capital/small town/village)?
   - Yule
   - Tschuprov
   - Standard deviation quotient

16. An examination of the relationship between employees' qualifications and wages delivered the following value for the Standard deviation quotient: 0.82. What does it mean?
   - there is a strong relationship between the two attributes
   - qualification determines changes in wages to a degree of 82%
   - wages determine qualification to a degree of 82%

17. The regression function that fits the empirical values best is the one for which the value of the residual standard deviation is highest.

TRUE **FALSE**

18. For a power regression function, the elasticity coefficient is identical with the value of the *b* parameter of the function.

TRUE FALSE

19. On the basis of a linear trend, a restaurant expects HUF 1 400 thousand for its 2014 postseason. We also know that the turnover in the postseason is 10% lower on average than the estimated value. According to the forecast, the turnover for the high season in 2014, calculated by considering seasonality, is
    - HUF 1540 thousand
    - HUF 1260 thousand
    - HUF 1410 thousand

20. In the case of an exponential trend and multiplicative seasonality, the product of the raw and corrected season indices is 1.

TRUE FALSE

21. A histogram is a special sort of bar chart with no space between the bars, which is best for the representation of class interval frequencies.

TRUE FALSE

22. A cartogram represents changes in the phenomenon examined by diagrams in different colors.

TRUE **FALSE**

FINAL EXAMINATION

VARIANT A

1. You are conducting a survey on car purchasing experiences. Which sampling technique would you use for gathering data? *More than one answer may be correct.*

	Random		Nonrandom
	Independent equivalent distribution		Quota
X	Simple random		Snowball
	Stratified	X	Purposive
	Cluster		Systematic
	Multi-stage		Arbitrary

2. You can create comparative and classifying rows from different types of data.

TRUE **FALSE**

3. According to visitors' data of a museum there were two female visitors per one male visitor. What type of ratio is this?

	Derived from classifying row		Derived from comparative row
	Distribution		Dynamic
X	Coordination		Areal comparative
	Derived from descriptive row		Plan task
	Intensity		Plan accomplishment

4. A company suffered a 25% set-back compared to the plan, due to the crisis. What type of ratio is this?

	Derived from classifying row		Derived from comparative row
	Distribution		Dynamic
	Coordination		Areal comparative
	Derived from descriptive row		Plan task
	Intensity	X	Plan accomplishment

5. When calculating the unemployment rate, we can only consider registered unemployed.

TRUE FALSE

6. The ROA index compares the earnings after tax to the total assets.

TRUE FALSE

7. 100 customers were observed on a particular day doing shopping in a hypermarket. Most customers spent € 20. € 20 is the value of which basic distribution descriptor?

Means

	Mean
X	Mode
	Median
	Lower quartile
	Upper quartile

Dispersion and shape indices

	Range
	Standard deviation
	Relative standard devn.
	A index
	F index

8. The Lorenz curve can be used for the study of concentration.

TRUE FALSE

9. An overview of the wages of the employees of a company shows that men make HUF 30 000 more on average than women. We know that due to the difference in average wages by professional qualification between men and women, men would make only HUF 6 000 more than women. Due the difference in the distribution of men and women by professional qualification,
   - men would make HUF 6000 less than women
   - men would make HUF 6000 more than women
   - women would make HUF 6000 more than men

10. The way to determine the effect of the difference in cluster means is to take the composition of clusters unchanged.

TRUE FALSE

11. A greengrocer's income from the aggregate sale of all of its products has fallen by 10% compared to last year.
   - the elementary value index is 110.0
   - **the aggregate value index is 90.0**
   - the aggregate volume index is 90.0

12. Value is derived as the product of price and quantity.

TRUE FALSE

13. Real GDP is
   - the product of the quantities and prices of a particular
   - the product of the prices of a particular year and a quantity typical of previous years
   - the product of the quantities of a particular year and the prices typical of previous years

14. In computing the GDP deflator for the measurement of inflation, the weight is
    - quantities of the reference year
    - quantities of the previous year
    - prices of the reference year
15. The value of the Cramer index between two attributes is 0.64. Which could be the two attributes?
    - visitor's place of origin and residence
    - visitor's place of origin and type of accommodation
    - visitor's place of residence and the amount of money they have spent
16. The value of the quotient of the standard deviation between type of train and number of delays is 0.64. What inference can you draw from this?
    - type of train affects the number of delays to a degree of 40.96%
    - type of train has no effect on the number of delays
    - type of train affects the number of delays to a degree of 64%
17. The linear coefficient of determination represents the degree in % to which the explanatory variable determines changes in the result variable.

    TRUE FALSE
18. Based on the regression function y=325-12x, we can say that if the value of the explanatory variable is increased by one unit, then the value of the result variable will increase by 325 units.

    TRUE **FALSE**
19. The equation of the linear trend that describes the quarterly changes in the production of a company (thousand pcs) across 2009 and 2013 is y=1200-50t. You can draw the following conclusion on the basis of the trend:
    - the production of the company decreased annually by an average of 50 thousand pcs
    - the production of the company decreased quarterly by an average of 50 thousand pcs
    - the production of the company in 2013 was 1200 thousand pcs
20. In the case of a linear trend and additive seasonality, the sum of the raw and corrected seasonal deviations is 1.

    TRUE **FALSE**

21. A polar chart can represent comparison between data sets along several different dimensions.

TRUE FALSE

22. A cartogram represents differences in the areal intensity of the phenomenon examined by different shades of colors.

TRUE FALSE

Variant B

1. You are conducting a survey on high school students' plans for further education. Which sampling technique would you use for gathering data? *More than one answer may be correct.*

	Random		Nonrandom
	Independent equivalent distribution		Quota
	Simple random		Snowball
	Stratified	X	Purposive
X	Cluster		Systematic
X	Multi-stage		Arbitrary

2. A combination table contains different kinds of statistical rows.

TRUE **FALSE**

3. Amount of imported goods in 2013 increased by 20% compared to 2010. What type of ratio is this?

	Derived from classifying row		Derived from comparative row
	Distribution	X	Dynamic
	Coordination		Areal comparative
	Derived from descriptive row		Plan task
	Intensity		Plan accomplishment

4. There is a 50% discount on some products at the beginning of the year. What type of ratio is this?

	Derived from classifying row		Derived from comparative row
X	Distribution		Dynamic
	Coordination		Areal comparative
	Derived from descriptive row		Plan task
	Intensity		Plan accomplishment

5. The activity rate represents the ratio of economically active within the working-age population.

TRUE FALSE

6. The fertility rate is a raw intensity ratio, which represents the number of live births per women of maternal age.

TRUE **FALSE**

7. 100 customers were observed on a particular day doing shopping in a hypermarket. ¼ of the customers spent less than € 15 and ¾ spent more than that. € 15 is the value of which basic distribution descriptor?

Means

	Mean
	Mode
	Median
X	Lower quartile
	Upper quartile

Dispersion and shape indices

	Mean
	Mode
	Median
	Lower quartile
	Upper quartile

8. The F coefficient can be used for the numerical expression of curtosis.

TRUE **FALSE**

9. An overview of data on two resort areas shows that the average amount of time spent by visitors has dropped by 10% compared to last year. On average, the time spent by visitors from Austria, Germany, and Poland has increased by 2%, but the yearly distribution of nationalities has changed so much that it has caused an average drop of 11.76% in overall time spent by visitors.
   - the value of the grand mean index is 110.0
   - the value of the cluster mean index is 98.0
   - the value of the composition-effect index is 88.24

10. If the value of cluster means is unchanged from one period to another, then the value of the grand mean index is identical with value of the cluster mean index.

TRUE **FALSE**

11. In regard to a family's expenditure on food-stuffs, we know that the price index of fruits is 110 and that of vegetables is 90. From this we know that
   - **the price of fruits rose by 10%**
   - the price of vegetables rose by 10%
   - changes in the price of fruits and vegetables balance each other out, the aggregate price index is 100

12. In index computation, we compare the data of the earlier period to the data of the later period.

TRUE **FALSE**

13. The volume index of GDP is
   - the change in nominal GDP
   - the change in real GDP

- the quotient of the nominal and real GDP

14. The index used in external trade price statistics is the ……index

Fisher Pearson Yule

15. If, in a class of students, only those passed the statistics exam who had attended the lectures, then
    - the exam success rate and attendance at lectures are mutually independent
    - the relationship between exam success rate and attendance at lectures can be quantified by the standard deviation quotient
    - there is a function-like relationship between exam success rate and attendance at lectures
16. An examination of the relationship between employees' qualifications and wages delivered the following value for the H2 index: 0.325. What does that mean?
    - there is a weak relationship between the two attributes
    - qualification determines changes in wages to a degree of 32.5%
    - qualification affects changes in wages to a degree of 0.325%
17. r=0.894 means that there is a strong positive relationship between quantitative attributes.

TRUE FALSE

18. Based on the regression function y=325-12x, we can say that if the value of the explanatory variable is increased by one unit, then the value of the result variable will decrease by 12 units.

TRUE FALSE

19. We have the following information about the residual standard deviation of the trends on a company's revenue: the average pair by pair deviation of the actual and estimated values in the case of a linear trend is 625 pcs, in the case of the moving average it is 612 pcs, and in the case of an exponential trend it is 627 pcs. Which trend fits best?
    - linear trend
    - exponential trend
    - moving average
20. In the case of an exponential trend and multiplicative seasonality, the raw and corrected seasonal deviations are equal.

TRUE **FALSE**

21. The title, the source, and the legend are obligatory elements of diagrams.

TRUE **FALSE**

22. In diagrams based on a coordinate system, you can define the relative position of the features examined.

TRUE FALSE