Basic Statistics for Economists

Natalia Kovtun

Basic Statistics for Economists

Basic Statistics for Economists

By

Natalia Kovtun

Cambridge
Scholars
Publishing



Basic Statistics for Economists

By Natalia Kovtun

This book first published 2022

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Copyright © 2022 by Natalia Kovtun

All rights for this book reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the copyright owner.

ISBN (10): 1-5275-7583-7
ISBN (13): 978-1-5275-7583-7

TABLE OF CONTENTS

1. The Development of Statistics as a Science 1

2. The Methods and Main Concepts of Statistics: Statistical Observation 7
2.1. The concept, method, and main tasks of statistics 7
2.2. Statistical population. Principles of the formation of a statistical population 9
2.3. Statistical indicators. System of indicators 10
2.4. Statistical observation 12

3. Summary and Grouping of Statistical Data 16
3.1. Concept of summary and grouping 16
3.2. Methodology of grouping 17
3.3. Regrouping of statistical data 19
3.4. Rules for creating statistical tables 23
3.5. Statistical figures and types 25
Practice Exercises 28

4. Absolute, Relative, and Average Values 40
4.1. Concepts and types of absolute values 40
4.2. Types of relative values and methodology of their calculation 41
4.3. The concept of the average in statistics and kinds of average values 48
4.4. The properties of the arithmetic mean 54
4.5. The multivariate mean 55
Practice Exercises 60
Exercises for Interim Evaluation 72

5. Variation Indicators and Distribution Forms 80
5.1. The concept of distribution rows and their characteristics 80
5.2. Characteristics of the distribution center 84
5.3. Variation indicators 91
5.4. Calculation methodology of the characteristics of distribution form 95
5.5. The concept of the normal distribution 99
5.6. Statistical study of concentrations 105
Practice Exercises 109

6. Sample Observation 120
6.1. Sample observation and its application 120
6.2. Types and patterns of sampling 121
6.3. Sampling errors: Methodology of calculation 123
6.4. Determining the sample size 128
6.5. Features of small sampling 129
6.6. Use of agreement criteria 131
Practice Exercises 137

7. Statistical Methods for Measuring Interconnection 147
7.1. Types of interconnections between events 147
7.2. Stages of studying interconnections between events 149
7.3. Method of analytical grouping: Variance analysis 150
7.4. Correlation-regression analysis 155
7.5. Multifactorial correlation 163
7.6. Nonparametric methods for studying interconnections between events 164
7.7. Range correlation 173
Practice Exercises 178

8. Time Series 197
8.1. Principles and practices of time series statistical analysis 197
8.2. The concept and types of time series 199
8.3. Calculation of the average levels of a time series 203
8.4. Time series indicators 205
8.5. Statistical analysis of developmental tendencies 210
8.6. Prediction of the levels of a time series 226
8.7. The statistical study of seasonality 227
8.8. Harmonic fluctuations 231
Practice Exercises 239

9. Indices 257
9.1. The concept of indices: individual and summary indices 257
9.2. System of aggregated indices 258
9.3. Weighted average arithmetic and harmonic indices 262
9.4. Indices of average values 264
9.5. Spatial indices 270
9.6. System of interdependent indices: factor index analysis 274
9.7. Characteristics of applying the index method 279
Practice Exercises 288

Appendix: Tables 303

1. THE DEVELOPMENT OF STATISTICS AS A SCIENCE

In its modern understanding, the word "statistics" was first used by German scientist Gottfried Achenwall who borrowed this word from the Italian language. During the Renaissance, a special course on the knowledge of policy was given the name *ragione di stato* or *diciplina de statu* and widely used in Italy. The words *stato* and *statu* mean "state" and are the origin of the German word *Staat* and the English word *state*. People dealing with policy matters and who were experts on matters related to other countries were called *statista*. In the seventeenth century, the phrase *diciplina statistica* (*the discipline of statistics*) was quite well known in Germany. Achenwall, transformed the adjectival use of *statistica* into a noun form and introduced the word *Statistica*, which referred to knowledge needed by merchants, politicians, the military, and the intelligentsia.

The emergence of statistics can be traced to the formation of the state, as the existence of the state would not be possible without having the required amount of data about the activities that take place in a state and an understanding of their inherent potential. For example, it was imperative to have information about the population and its composition, the size and quality of the land, the number of livestock, and the volume of trade, etc. For instance, to count the army the Persian Emperor Darius (522-486 BC) ordered every warrior to bring a stone and put it in a certain place. According to the Greek historian Herodotus (484-420 BC), the Scythian tsar Ariant ordered every Scythian, under threat of execution, to bring a copper arrowhead so that he could know the number of his subjects. In ancient China (twenty-third century BC), information was gathered about:

- the population;
- classification of the population by gender and age;
- the profitability of land;
- and changes in trade.

In the book "Shu-King" by Confucius (551-479 BC), we find a description of the Chinese census in 2238 BC. In the Book of Numbers – the Fourth book of Moses in the Bible – we find a story about the number of the male

population capable of carrying weapons. In Athens, a well-organized record of natural movements of the population was undertaken, along with the land cadastres, including the characteristics of land property as well as counts of facilities, inventory, livestock, slaves, and the revenue generated. In Lydia, with the use of coins in the seventh century BC, a generalized monetary estimation of all the state's assets was made possible – before this time, only natural measuring instruments were used in accounting. Thus, the necessity to make an account of the inventory of a country gradually emerged.

The first known inventory of a state was by Aristotle (384-322 BC). It included 157 cities and states: "The size of a state" Aristotle wrote, "is measured by its population; however, attention is to be paid to the possibilities, but not to the number". He connected possibilities with the character of the people, and this latter element with the geographical environment: "The tribes who live in countries with a cold climate, namely in Europe, are courageous, but not very clever or skillful in crafts. Peoples who live in Asia have mental abilities and are capable at crafts, but they are not courageous enough, that is why they live in subordination and slavery".

In Ancient Rome, the statistical inventory of the state received a new and powerful impulse. In 550 BC, Servius Tullius created a prototype of the first statistics authority to carry out a "census" and assess the number and type of free citizens. Clerks of the census, who were called enumerators, recorded details about the head of a family – his name; the name of the tribe he belonged to (in Ancient Rome each territorial district paid a different amount of tax); his father's name and age or the name of a former landlord; and also the names, gender, and age of all his family members. In addition, the assets of the whole family were registered. Initially, (during the republic) censuses were conducted once every five years. Later (under the imperial regime), the census periods were increased to ten years (the last census was carried out in 72 BC). Based on the information in the census, the free population was classified into five categories (classes) in accordance with their assets:

- I class – property value of not less than 100 thousand assets;
- II class – not less than 75 thousand assets;
- III class – not less than 50 thousand assets;
- IV class – not less than 25 thousand assets;
- V class – not less than 12 thousand assets.

The evaluation included real estate with land and the inventory of possessions. Certainly, there was also the VI class – the poor – but they were deprived

of the right both to serve in the army and to vote. In addition to periodic censuses, observations of changes in population were made, which helped the lawyer Ulpian (170-228 AD) come to a conclusion about the possible life expectancies of different age groups.

The development of statistical accounting went backwards during the Middle Ages. The Domesday Book, a record of a general land census of England, is a unique statistical monument that has lasted up to the present day. The census lasted for four years (1083-1086), and captured a topographical inventory of the land and population in each place. The Domesday Book presents detailed information about royal, church, and feudal estates, covering about 240 thousand estates in total. Property was recorded in the first place, followed by details of people and families. In general, both in ancient and mediaeval times, the accuracy and reliability of the data were at a very low level. Besides, in most cases comparisons were of a qualitative nature and were made through individual judgments – more-less, better-worse, etc. Quantitative comparisons were very rarely used. As they were made using Roman numerals, calculations were also complicated and we find numerous mistakes in them.

The era of the Renaissance saw a new phase in the development of statistics. The intensive development of international trade had an impact on the formation of conventional and descriptive statistics. In particular, in the Venetian Republic (twelfth to seventeenth centuries), trade developed rapidly. In the twelfth century, consuls and ambassadors, on coming back to the homeland, were obliged to submit to the senate reports about the political, economic, and physical conditions of countries assigned to them. These were comprehensive and systematic inventories, although they rarely contained numerical information.

The spirit of humanism, an awareness of human greatness, and significant geographical discoveries all generated huge interest in other countries of the world. Merchants and politicians were no longer satisfied with fairytales and poetic fantasies about their journeys, but wanted to have reliable information from official documents or reports by travelers. This led to the development of a body of special surveys. The most notable of these was written by Francesco Sansovino (1521-1586), the son of the Tuscan sculptor and architect Andrea Sansovino. His “Del governo et amministrazione di diversi regni e republiche”, published in 1562, had descriptions of 22 countries. Sansovino gives the following information about each country: nature, residents, routine life, army, political system, culture, trade, and

religion. Sansovino's book was a great success and saw five editions in the first forty years alone, before its translation into foreign languages.

However, the description of the countries was narrative in style and did not contain numerical details. Similar documents were made by the followers of F. Sansovino: Botero (1589), d'Aviti (1614), Yanotti (1624) and others. It is interesting to note that none of them were state functionaries. They were either merchants or bankers as these were the people who were interested in developing this knowledge.

During this time, another important event took place. The Franciscan monk and mathematician Luca Pacioli (1445-1517) developed a fundamental encyclopedic work: the "Summary of arithmetic, geometry, proportions and proportionality" (1494). Pacioli's work is a landmark in the history of the development of probability theory – the science most closely connected to statistics.

In the fourteenth century, the first marine insurance partnerships emerged in Italy and the Netherlands. Insurance of overseas transport led to the insurance of goods transported via land, rivers, and lakes. Insurance partnerships evaluated the risks of carriage and the higher the risk, the higher the insurance premium charged. Insurance premiums were 12-15 % of a cargo's value (overseas transport) and 6-8 % of the freight value (overland and river carriage). By the sixteenth century, marine insurance had become widespread and was found in many countries. In the seventeenth century, other kinds of insurance began to appear. The activity of insurance companies relied on the generation of statistical data and supported the development of statistics and the theory of probability.

As such, in the seventeenth century certain conditions emerged in Western Europe that led to the development of modern statistics. Originally, it was not a science in the modern understanding, but rationalized opinions about the foundation of the state and theory was not independent from practice – they were considered inseparable. Overall, this marked a big leap forward.

The major precursors to the development of statistics were:

- The extensive development of primary accounting and the accumulation of large amounts of descriptive information of broad scope in the sphere of social events, which could be used for the generation of statistical data;
- The availability of members of society who could advance science, including social science;

- The increase in the need for quantitative measurement of events and for the regulation of public life in terms of practical necessities (political, economic, and administrative etc.) and also for the sciences, through which to study society;
- The development of the fundamental sciences (primarily philosophy, mathematics, and law) recognised the necessity of statistics as an instrument for the cognition of social events and processes so as to reveal their specificity and to comprehend relevant methodological principles;
- A change in human consciousness and vision of the world, leading to the formation of new ideas about the state and society.

All these factors appeared in the second half of the seventeenth century, which marks a frontier in the history of mankind and saw a great breakthrough in the history of the development of science and a period of extraordinary social change in Western Europe. In fact, during this very period, the scientific academies of England, France, and Germany were founded. The work of scientists like Galileo Galilei, Francis Bacon, René Descartes, Johannes Kepler, Baruch Spinoza, Gottfried Wilhelm Leibniz, and Isaac Newton established the foundations of the modern sciences and great achievements were made across disciplines ranging from mathematics and the social sciences to physics and astronomy. In the seventeenth century, the study of political economy and demography evolved. Political economy became the theoretical basis for statistics and demography developed alongside statistics and from statistics, although later it became its own field of study.

In the seventeenth century, significant changes took place in the sphere of socioeconomic relations. Commodity-money relations were intensively developed and foreign trade occupied a special place in the economy of European countries. Foreign and domestic markets expanded and trade capital strengthened. The local/home system of capital intensive industry began to develop and is associated with capital's entry into production. The first form of capital-intensive industry – the manufacturing industry – started to flourish. Mercantilism became a hallmark of the time with the active participation and development of investment in trade and industry by the state – the intervention of the state in the economy was a characteristic feature of this period. As a consequence, the demands of investment in industry and trade required quantitative knowledge about economic phenomena.

The development of statistics was inevitable under these conditions. It started simultaneously both in mainland Europe and England. But the forms and contents were different: in mainland Europe the focus was on state studies, while in England it was political arithmetic. Later, these became distinct trends in the development of statistical science. Works in the sphere of political arithmetic were devoted to socioeconomic problems. Political arithmetic sought to introduce some regulation into social and economic life. Measurement methods were acknowledged to be a mandatory condition for the research of mass recorded data. The term "political arithmetic" itself indicates the combination of mathematics with policy through measurement of the facts of socioeconomic life (the meaning of the word "policy" here was similar to the concept of "science about society").

A significant difference is seen in state studies. Despite the works of F. Sansovino and his followers providing the foundation of this trend, Germany, not Italy, was the real center of state studies – a country with strong traditions of cameral (budget) accounting. Statistics – state studies – was based on the concept that a state was the only source of observation and provided the only tools of observation – clerks, administrators, executives, and policemen. The representatives of this trend underestimated the capability of mathematical means of cognition. The measuring nature of statistics, at least at first, was not considered to be its primary attribute and quantitative estimates were considered to be a special case of general description. Another name for this trend – descriptive statistics – originated from this notion.

Thus, both political arithmetic and state studies were connected to the gradual development of enterprise records. Statistics and economic and political geography were derived from state studies, while political economy, statistics, and demography all came out of political arithmetic. The trends have common subjects: society (the "state" for state experts), but different methods – description and measurement.

2. The Methods and Main Concepts of Statistics: Statistical Observation

2.1. The concept, method, and main tasks of statistics

The word **"statistics"** is derived from the Latin **status**, meaning "a status" or "a state of existence"; it was also used with the meaning of "political state". The use of the term "Statistics" appeared in the scientific literature of the eighteenth century and, at first, referred to ***state studies.*** Currently, the term "statistics" is viewed from two perspectives: as a branch of knowledge and as an applied science specialty. Thus, statistics can be understood as:

1. A branch of knowledge (science), i.e., a scientific course.
2. A specialized branch of applied science aimed at collecting, processing, analyzing, and publishing mass data about the events and processes of social life.
3. A collection of digital information that characterizes any event or group of events.
4. A term used to refer to the functions of the results of observation.
5. A specific research method.

In Ukraine, legal relations in the branch of state statistics are regulated by the appropriate law. The law defines the rights and functions of the bodies of state statistics and the organizational principles of exercising state statistical activity the aim of which is to develop comprehensive and objective statistical information on the economic, social, demographic, and ecological situation, and on Ukraine's regions, so as to provide state and society with usable information.

Statistical research is the process of cognition (studying) of socioeconomic events with the help of statistical methods and the quantitative characteristics of indicator systems. Statistical research includes the following stages:

1. Statistical observation.
2. Aggregation, generalization, and processing of statistical data.

3. Analysis of statistical data and interpretation of the results.
4. Predication of the development of events over time.

Analysis, as a rule, is followed by the creation of statistical tables and figures. Attention should be paid to the quality of presentation of research results (in the form of an analytical report) and that they conform to an author's qualifications and culture. The appropriate chronological steps are general and the detailed contents of each stage depend on the purpose of the research and the nature of the data. In practice, one often needs to go back to previous stages and repeat some of them. In the specialized literature, the "stage of exploratory data analysis" is emphasized, the aim of which is to carefully study preliminary data and make an appropriate choice about substantive mathematical tools. Computers and special software, e.g. STATISTICA or SPSS, ensure a wide range of possibilities for statistical research. These can all improve the quality and speed of performance – from the comprehension of the primary block to the preparation of inferences including statistical tables and figures.

The concept of statistics can be comprehended in terms of studying a quantitative aspect of a mass event or process in its inseparable connection to respective qualitative contents under certain conditions of place and time, i.e. context. The regularity of the connection between a necessity, an individual random event, and the description of a phenomenon is called ***statistical regularity.***

The methodology of statistics includes a set of techniques, rules, and methods of statistical research into socioeconomic events. The typical characteristics of statistical methodology are:

1. A focus on quality analysis.
2. The separation of similar sets (types).
3. Modification of research techniques and methods due to a change in the essence and form of the events and processes under study.
4. The application of indicator systems.
5. Study of the data for analysis and its use in an indicator system.

General research methods see the development of specific methods for specific sciences. The dialectic interpretation of unity and divergence; necessity and randomness; or something of a general, partial, and single nature sees concrete expression in the contents of statistical categories and methods. It is practical to classify **statistical methods** into groups as:

- Methods of mass observation;

- Methods of consolidation and grouping;
- Methods for the identification of generalized and synthetic indicators (methods of average and relative values; analysis of distribution rows; measurement of connection, etc.).

2.2. Statistical population. Principles of the formation of a statistical population

A statistical population is made up of a plurality of elements united by common conditions and reasons. A population consists of several **units of population** that have common features or traits. **A feature** is a characteristic or property that demonstrates the essence, character, and attributes of a population. Thus, all the students of a group (major, course, institute, etc.) could be considered a statistical population. Each student is an element of a population defined by common features, i.e. gender, age, degree major, or examination grade, etc. However, every element of a population has its own individual value (level) in terms of a certain feature. Some principles of the construction of a statistical population include:

1. The availability of common conditions and reasons, i.e. the unity of aim and contents.
2. The availability of common features.
3. Within the whole population, one feature has to take different meanings. For example, "gender" can take on two meanings – "male" and "female" – while "examination grade" could have four meanings. The fluctuation of meaning of a feature in a population is called ***variation***.
4. A population has to be uniform in the important features that are studied. Namely, the meaning of a feature, i.e. **variation**, must not fluctuate beyond certain limits.

The level of a feature's meaning in some elements is measured on a scale covering a set of event properties and their corresponding meanings (numbers). There are three main scales:

Metric refers to the usual numerical scale, which is used to measure physical values or calculated results. All arithmetic operations can be used on this scale. This scale is used to measure an individual value such as the feature "age". A feature measured on this scale can be a minimum, a maximum, or an average value.

Nominal scale refers to a scale of names. Examples of attributive features measured on this scale are: “gender”, “degree major”, “education”, and “nationality”. Obviously, no arithmetic operations can be done using this scale, but we can think of using this scale for the values that occur most frequently in the population.

Ordinal (range) scale defines not only the similarity of elements, but also their succession by a type “more than”, “better than”, etc. Each dot of a scale receives a point (range). Thus, a “good” grade received in an examination, usually denoted as “4”, corresponds to a level of knowledge higher than “satisfactory”, but lower than the level “excellent”. However, it is a fact that two “2” grades do not make “4” on this scale. So, the calculation of “an average grade point” for students who have passed an examination is incorrect from the point of view of statistics. At the same time, we can calculate the average grade point received by a certain student in the examination session, as the addition of the points will characterize a general level of knowledge shown by that particular student during the examinations.

From the perspective of the above-mentioned scales, features can be classified as quantitative (variational) and qualitative (attributive). The level of a quantitative feature is measured with help of a metric or ordinal scale. For an attributive feature, measurement involves registration of the availability or lack of some properties, i.e. classification. Each class is marked with letters and numbers (coding is done). For example, the attributive feature “gender” has two classes, marked “m” and “f”, or “1”, “2”; the attributive feature “grade” has five classes, marked “1” to “5” or “fail” to “excellent”. ***In view of the interconnection of*** “cause and effect”, features are classified as factorial and effective (level of qualification and remuneration of labor; amount of fertilizer applied per hectare and harvest ratio, (centner or hundredweight (cwt)/hectare (ha)).

2.3. Statistical indicators. System of indicators

A statistical indicator is a number together with a set of features that characterize the circumstances to which they belong (i.e., what?, where?, when?, how? is it to be measured).

Statistical data comprise a set of indicators developed from statistical observation or data processing. From the point of view of state statistics, “statistical data is information received from conducted statistical observations, which has been processed and presented in a formalized and standard form in compliance with established principles and methodology.

Statistical data, which is the result of the collection and grouping of primary data, being impersonal, presents a summary of impersonal statistical information (data)". Furthermore, "statistical data is the data of banking and financial statistics, the data of the active balance sheet and alike, which are compiled based on the data received from the Central (Government) Bank and authorized bodies of state". Statistical data needs to meet specific requirements. It should be:

- Reliable (to correspond to the real state of things);
- Complete (to comprehensively reveal the essence of the event);
- Relevant in time;
- Comparable over time and space;
- Accessible.

Information concerns data required to solve a concrete task. If the data does not belong to/relate to the task or is not new for a researcher, it does not contain any information for the researcher. The reliability of an indicator is determined by its adequacy and the accuracy of its measurement.

Statistical information is a type of information that makes it possible to give a quantitative reflection of the events and processes taking place in various spheres of life. This is reliable and scientifically grounded information that objectively describes the state and norms of socioeconomic events and processes.

Depending on the type of calculation, there are primary and derivative (secondary) indicators. **Primary indicators** are generated by summarizing the data of statistical observation and take the form of an absolute value (e.g. output produced per quarter). **Derivative indicators** are calculated based on primary or secondary indicators. They come in the form of average and relative values (e.g. average salary; share of employees with university education).

According to the feature of time, the indicators are divided into "interval" and "moment" (e.g. average recorded number of employees per year; the size of the population on January 1). Interval indicators characterize events over a period of time and so the indicator has an economic connotation (e.g. to bring a residential area into operation within a year). Moment indicators characterize an event at a certain moment (e.g. the provision of housing facilities at the beginning of the year). The peculiarity of indicators of moment is that they cannot be added – their sum has no contents. Interval and moment indicators can be both primary and derivative.

The complexity and versatility of socioeconomic events require the use of **a system of statistical indicators** with which to engage in a comprehensive study of mass social and economic processes. This system results from the rationale of the research and ought to be in a hierarchical structure where indicators at higher levels are calculated on the basis of indicators of lower levels. An important factor for high quality and effective analysis is the extent of information support available to an indicator system.

Information support concerns a system of methods for the creation of information flows (a well-ordered population of data is required to solve a concrete problem) and the process of disseminating generated information. Information support is built on certain principles:

- A well-defined focus of information flows for the implementation of concrete tasks in a statistical analysis;
- The integrity of the information support;
- The dynamic, territorial, and methodological comparability of the main elements of the information system;
- The transparency of the information support of the system, as a whole, and its main elements, in particular;
- Accessible to a large number of users.

2.4. Statistical observation

Statistical observation is the planned and scientifically organized registration of mass data about socioeconomic events and processes. An illustration is given by the Law of the State "On state statistics": "statistical observation is a planned, scientifically organized process of data collection concerning mass events and processes concerning economic, social and other spheres of the life in the Country and its regions with registration of such events and processes using a special program based on statistical methodology".

The organized forms of observation used in statistical practice are: statistical reporting, specifically organized statistical observations, and registries. Specifically organized observations include such things as censuses, one-time surveys, special examinations, and interviews. Statistical observations can be of mass and non-mass nature.

Mass statistical observation covers all (without any exceptions and/or exclusions) units of the population that is under study.

Non-mass statistical observation is performed for separate units of the population under study.

Statistical observation starts with the development of a plan of observation. A statistical observation plan is a set of programmatic, methodological, and organizational issues. The programmatic and methodological issues of the plan define the purpose of observation; the object and units of observation; the sources and means of acquiring the data; the time (moment) of the observation; and the program of observation. The program of observation revolves around a list of questions the answers to which provide the desired results of observation. The program also includes: the development of statistical tools and the determination of the kinds and techniques of observation. The organizational issues of the plan center on the choice of bodies and personnel to carry out observation; the places of observation; logistics; control systems to ensure the accuracy of the results; and the time and period of observation. In statistical practice, the completeness of the data is first checked visually and its reliability is checked by means of logical and arithmetic control.

It is important to differentiate the concept of a unit of statistical observation from that of an element of a population. **A unit of observation** is a carrier of information, while **an element of a population** is a carrier of features. Thus, when doing an equipment inventory, for example, the unit of observation is an enterprise (factory or plant), while the element of observation is a tool or a mechanism. As was mentioned above, by population unit coverage, observations can be mass and non-mass. The latter has several types: observation of the main block, sampling, monographic, and questionnaires.

In sample observation, we study only a part of the population, chosen by particular methods, which can:

a) Provide equal possibilities for each unit of a population to be in the sample;
b) Determine a sufficient number of chosen units.

Accordance with these two conditions makes sampling **representative**, allowing us to extend the characteristics of sampling observation to the whole population, which is termed **general** (e.g. average value).

Observation of the main block involves registration of data about most of the population units that are significant in research into an event.

Monographic observation envisages a detailed inventory of a small number or separate units of a population that can be considered typical.

Questionnaire observation happens when not all the registration documents (questionnaires) are returned. This observation, for example, is applied in sociological research.

There are three ways to get statistical data: direct fact recording; document recording; and interviews. Direct recording is done by a statistical assistant through examination, evaluation, calculation, and measurement. Document recording involves fact generation from primary records through the analysis of documents. Observation through interviews can be done by self-registration, correspondence, or via a questionnaire.

A specific tool of statistical observation is monitoring. **Monitoring** is a kind of continuous observation. It follows a specially developed program of observation with variable frequency guided by the dynamics of an event under observation. Monitoring is organized to provide control over processes concerning nature, the economy, and society with the goal of more efficient management of the development of socioeconomic, natural, and other phenomena. Monitoring is used to study events statically and dynamically, and to identify the frequency and factors that define the pattern of occurrence of the events under study. The results of monitoring are used to guide interventions.

Discrepancies between observation and real data are called **observation errors**: there are two kinds of such errors. **Registration errors** result from the incorrect identification and registration of data. According to their nature, registration errors can be:

Random errors. These occur due to carelessness, negligence, insufficient qualifications of personnel and/or inadequate competence of a statistical assistants, or through respondent errors. The effect of random errors in mass observation balances out and does not affect the overall results.

Systematic errors result from one-side distortion and the aggregation of this effect leads to a shift in estimation. An example of this is a rounding off error. For instance, if five comes after an even number in a half-integer, then rounding off tends towards the smaller side; when five is comes after an odd number, then rounding off shifts to the larger side.

Errors of representativeness are typical only in sample observation and occur when the principles of the formation of a sample population have not been properly observed.

3. SUMMARY AND GROUPING OF STATISTICAL DATA

3.1. Concept of summary and grouping

Items that characterize individual population units are derived from statistical observation. Regular and necessary items underlie occasional and insignificant ones. Special processing of statistical data to construct a **summary** of the observations is necessary. The essence of a statistical **summary** is the classification and aggregation of statistical data.

A **summary** involves a set of operations aimed at the generalization of concrete individual data that identifies a population in terms of regular features and regularities typical for an event in general. A **summary** has the following stages:

1) Material grouping;

2) Development and measurement of the indicators that characterize typical groups and sub-groups;

3) Computation of group and general summaries;

4) Presentation of the results in the form of statistical tables and figures.

Such a **summary** relies on a grouping method. **Statistical grouping** sees the classification of a population into groups according to their significant features. Three main activities are involved in statistical grouping:

1) Distribution of a non-uniform population into qualitatively homogeneous groups – this task is performed by means of typological grouping (e.g., distributing students into groups by "gender").
2) Study of the structure and structural changes of qualitatively homogeneous populations and their distribution by scope of a variable feature – this task covers structural (variational) grouping. The distribution of employees in an organization into age groups (up to 20 years, 20-25 years, and so on.) is a routine example.
3) Identification and study of the interactions between features – this is termed **analytical grouping**. Here, at least two features are always

present when doing analytical grouping: one of them is factorial (factor) and the other one is the result. So, having grouped workers within a single profession by qualification level (e.g., category I, category II, etc.) and having calculated an average monthly wage for each group (by qualification), one can make an assumption that there is a statistical interconnection between the features "profession" and "wage" (an interconnectedness between the type of profession and a quantum of wage).

Such distinctive grouping, ***depending on the nature of the task***, is quite obvious, but, in practice, it is not always possible to draw a line between various types of groupings. Mostly, this is the case for typological and structural groupings. Grouping by a single feature is called ***simple*** *grouping*. When two or more features are taken together, it is called *combined grouping*. For instance, people under observation were grouped by hair color and then each of the groups, in turn, was grouped by eye color (Table 3.1). Tables from this type of grouping involve *cross tabulation* and can be used to study the interconnection between two qualifying features. In fact, the data in Table 3.1 gives grounds to assume that dark eye color is somehow connected to dark hair color and we can often see this match (unless it is dyed).

Table 3.1. Grouping by hair color and eye color

Eye color	Hair color			Total
	Fair	Brown	Black	
Blue	177	71	17	265
Grey	95	119	75	289
Hazel	12	24	43	79
Total	284	214	135	633

3.2. Methodology of grouping

It may seem that statistical grouping is rather simple and can hardly be called "scientific", but it occupies a special place in statistical research. A grouping method helps us single out uniform groups and defines the scope and feasibility of other research methods.

Preconditions for use of this method include: the comprehensive analysis of the essence of an event; the precise definition of significant features and intervals of grouping in such a way that constituted groups represent similar units of a population; and that individual groups differ from each other considerably. There are some peculiarities to grouping by qualitative and quantitative features. If grouping is done by an attributive feature, then the number of groups corresponds to the number of value types of that feature. If a feature is alternative, only two groups are possible.

When grouping is by a variational (quantitative) feature, two approaches are possible, depending on the type of feature variation. If a feature varies discretely and displays different values at small intervals, then the methodology of grouping is similar to grouping by an attributive feature. The presented example is that of grouping students by examination results. Here, four groups can be singled out according to the number of points received in an examination. In the case of grouping by a continuous feature or a discrete feature, which varies within large ranges, the grouping methodology relies on the choice of the number of groups and grouping intervals. When grouping employees of an organization by age, grouping intervals and, probably, the number of groups will differ from grouping full-time students by age. In the first case, an interval width could be five years and a group of employees between the age of 45 and 50 years could be considered to be a uniform one; however, in the second case, this approach would hardly be relevant. In any case, it is necessary to take into consideration how many units of a population can be put into each group.

If the intervals are too close and many small groups are formed, the grouping and the consequent picture of the population structure would be blurred, making it difficult to find out distinctive regularities. If we take very wide interval ranges, groups may become internally non-uniform, i.e. they will include units that differ by qualifying characteristics from each other within the group, violating the grouping principle of intergroup uniformity. The correct choice of intervals in the case of analytical grouping is of a great importance – a poor or biased approach can distort the real nature of the interconnection between events.

There are some formal rules for devising the number of groups, e.g. Sturges' Rule where the number of intervals is determined based on the scope of a population (n): $\mathrm{m} \approx 1 + 3.222 \log n$. However, it is important to acknowledge that the main role is played by the researcher's understanding of the events under study through her/his experience and intuition. To determine the number of groups, we can use the formula: $m \approx \log_2 n + 1$.

The size (width) of an interval is the difference between the maximum and minimum value of a feature in each group. The intervals can be equal or non-equal. Equal intervals are applied when a grouping feature is distributed more or less evenly in a population. The width of equal intervals is determined with the formula: $h = \frac{x_{max} - x_{min}}{m}$, where x_{max} and x_{min} refer to the maximum and minimum values of the characteristic of a population, respectively, and m is the number of groups. An example of grouping with non-equal intervals would be grouping some organizations by the number of employees: up to 10 people; 10-30 people; 30-100 people; 100-200 people; 200-500 people; and 500 people and more. With this approach, as a rule, the lower limit is included and the upper limit is excluded. This means that an organization with 30 people goes into the third group. The first and the last intervals in this example are ***open***, while the rest of the intervals are ***closed***; they have a lower limit and an upper limit. Sometimes, when the number of observations is small, the principle of equal frequencies is applied and, accordingly, population units are put in increasing order with each group carrying the same number of population units. This prevents the formation of small groups.

3.3. Regrouping of statistical data

If a variational row has unequal intervals, then the distribution density is calculated to achieve a correct reflection of the nature of the distribution. In some cases, data regrouping is carried out to form new groups on the basis of those available, if the existing groups do not meet the purposes of analysis. Regrouping is done based on absolute and relative distribution densities. Absolute distribution density is calculated with the formula:

$$\boldsymbol{\rho_j^f} = \frac{\boldsymbol{f_j}}{\boldsymbol{h_j}},$$

where $\boldsymbol{\rho_j^f}$ indicates the absolute distribution density in j-interval; $\boldsymbol{h_j}$ is the width of j-interval; and $\boldsymbol{f_j}$ is the frequency of j-interval.

We use frequencies to define relative density with the formula:

$$\boldsymbol{\rho_j^\omega} = \frac{\boldsymbol{\omega_j}}{\boldsymbol{h_j}},$$

where $\boldsymbol{\rho_j^\omega}$ is the relative distribution density in j-interval; $\boldsymbol{h_j}$ is the width of j-interval; and $\boldsymbol{\omega_j}$ is the relative frequency of j-interval.

These indicators are used to transform intervals with the aim of making a comparative evaluation of data that has been collected from various populations and processed in different ways. For instance, for two enterprises we can show the distribution of workers by percentage of output rate (Table 3.2).

Table 3.2. Distribution of workers by percentage of output rate

Enterprise №1		**Enterprise №2**	
Percentage of output rate	**Structure of workers, %**	**Percentage of output rate**	**Structure of workers, %**
Up to 90	2	Up to 100	8
90 – 100	3		
100 – 110	50	100 – 120	40
110 – 120	30	120 – 150	20
120 – 140	8	150 – 180	15
140 – 150	5	180 and more	17
150 – 160	2		
Total	100	Total	100

To regroup data so as to have comparable groups appropriate for analysis, we can use the aggregation of the interval (Table 3.3).

Table 3.3. Distribution of workers at two enterprises by percentage of output rate using the method of interval increase

Percentage of output rate	Structure of workers, %	
	Enterprise №1	Enterprise №2
Up to 100	2+3=5	8
100 – 120	50+30=80	40
120 – 150	8+5=13	20
150 and more	2	15+17=32
Total	100	100

However, another regrouping method could also be used: one can single out the groups by percentage of output rate (Table 3.4).

Table 3.4. Grouping of workers by percentage of output rate

Group	1	2	3	4	5	6
Percentage of output rate	Up to100	100 – 110	110 – 120	120 – 140	140 – 160	160 and more

For such a regrouping, it is necessary to split the intervals of Enterprise No. 2. With the information we have about relative distribution density (ω_j'), frequencies of an appropriate interval can be computed by multiplying distribution density (ρ_j^f) by interval size (h_j), as shown here: $\omega_j' = \rho_j^f \times h_j$. From the data in Table 3.2, we can define distribution density for the second, third, and fourth intervals of Enterprise No. 2 (Table 3.5).

Table 3.5. Calculation of the distribution density of workers at Enterprise № 2 by percentage of output rate

Percentage of output rate	Distribution frequencies, %	Distribution density	Percentage of output rate	New distribution frequencies, %
100 – 120	40	40/(120-100)=2.0	100 – 110	2×(110-100)=20
120 – 150	20	20/(150-120)=2/3	110 – 120	40-20=20
150 – 180	15	15/(180-150)=0.5	120 – 140	(2/3)×(140-120)=13
x	x	x	140 – 160	(2/3)×(150-140)+ 0.5×(160-150)=12

The results of regrouping are presented in Table 3.6.

Table 3.6. Distribution of workers at two enterprises by percentage of output rate after regrouping

Percentage of output rate	Structure of workers, %	
	Enterprise №1	Enterprise №2
Up to 100	2+3=5	8
100 – 110	50	20
110 – 120	30	20
120 – 140	8	13
140 – 160	5+2=7	12
160 and more	–	(20+15) - (13+12)+17=27
Total	100	100

3.4. Rules for creating statistical tables

Statistical tables are used to present the results of summarization and grouping in the best possible way. **A statistical table** presents the summary in a convenient form for understanding and analysis, i.e. this is a unified system of presentation of the results of statistical observation. By its logical construction, a statistical table is considered "an informational statistical sentence". Tables consist of a statistical subject and predicate. A subject is the object of statistical study (a statistical population) while a predicate is the system of statistical indicators that characterize a population.

A subject of a table denotes statistical populations characterized by different indicators, forming its **predicate**.

Let us consider a general model for a statistical table (Figure 3.1). Depending on the structure of a subject, statistical tables can be simple, group, or combined. For a predicate to be constructed, we can use simple predicate development and complex (combined) construction.

There are certain rules on how to make tables:

1) If the units of measurement are identical, the unit of measurement is presented in the column headings. Sometimes there is a separate column for similar units of measurement. A common measurement unit is mentioned above the table in brackets on the right or in the name of the table after a coma.
2) All the data in one column are given with the same accuracy.
3) Notes specific to some columns, rows, or cells can be placed below the table.
4) Objects of a subject and features of a predicate must follow a certain logical sequence.
5) It is expedient to supplement the absolute values of a predicate with relative and average values.
6) Columns of a predicate are numbered if a table spreads over several pages. It is practical to mark a subject column with a letter.
7) If the names of some columns (rows) are repeated or the columns have the same conditions or contents, it is convenient to combine them under a common heading.
8) Sometimes, the formula for the calculation of an indicator is given in the headings of a column. For example, we could use "col.2/col.1" in the name of column 3.

9) The information in summary rows (columns) of the table is marked as "total" if a summary row is the sum of previous ones or "overall" if a summary row contains both a general sum and relative and/or average values.
10) A table cannot have empty cells. If there is no information about the size of an event, it is filled with dots (...); the absence of an event is marked with a dash (–); the number 0.0 (0.00; 0.000 and others) is put in the case of a small value and/or when a figure in a column goes beyond the limit of accuracy defined for the table. The sign 'x' is put in the case of a column that is not filled in. An example is presented in Figure 3.3 below.

Table №

HEADING

(Contents, place, time, unit of measurement, if there is one for the whole table)

Predicate Subject			**Total**	**Including**		
				total	**including**	
A	1	2	3	4	5	6
...						
...						
Total						

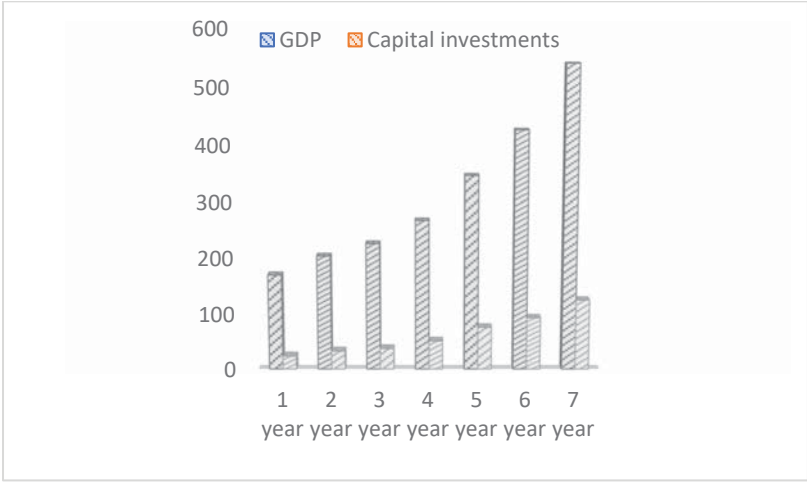
Kind of table:
Simple (list of units); simple complex
Group (unit grouping);
Combinational (grouping by several features).
Predicate development

Figure 3.1. Model of a statistical table

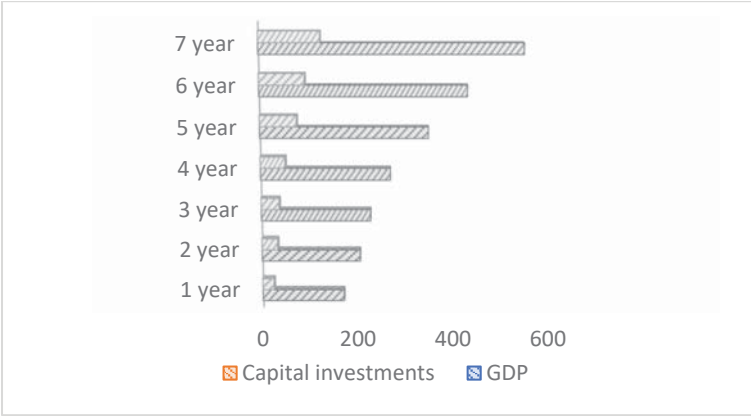
3.5. Statistical figures and types

A statistical figure is a nominal expression of numerical values and their relationship in the form of geometrical figures, lines, and other graphical forms of presentation. The main purpose of a figure is to provide clarity. Figures are classified by:

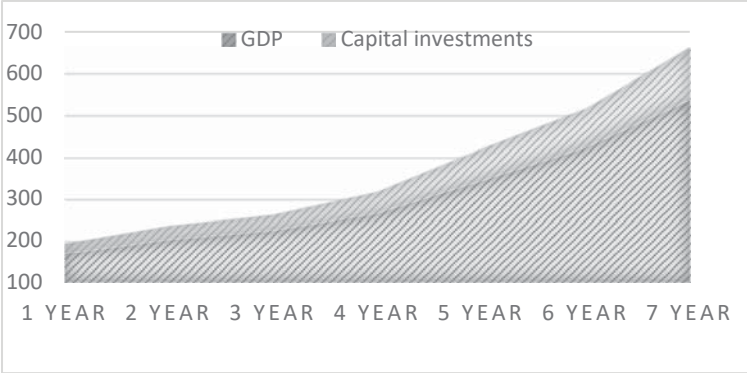
1) functional purpose;
2) field form;
3) form of graphical image;
4) coordinate system.



a) Column diagram



b) Band diagram



c) Line diagram

Figure 3.2. Dynamics of GDP and capital investments in a country for a period of 7 years (in national currency)

For ***functional purposes***, we can use figures of grouping, distribution rows, dynamics rows, interconnection, and comparison. For ***field form***, we use diagrams, cartograms, and carto-diagrams. Diagrams are widely used, which is why the term "diagram" is often identified with "statistical figure". Examples of column, band, and line diagrams, which are used to show dynamics, are presented in Figure 3.2. A pie graph is shown in Figure 3.3. Pie graphs show the structure of an event, with the slice of the sectors proportional to corresponding size (quantum). If a population has more than

5-6 structural elements, or there is no significant difference between the constituent parts, there is no need to use a pie graph and such diagrams lose their relevance.

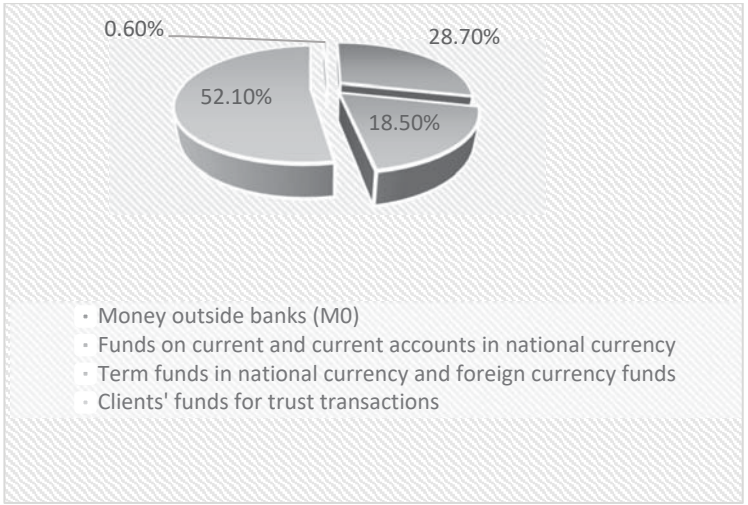


Figure 3.3. Structure of money aggregate M3

According to the ***form of graphical image***, we use linear, plane, sizeable, and shaped figures. Of the plane forms, square and round figures are most frequently used. The area of the figures is proportional to the size of the values they represent.

According to the ***type of coordinate system***, figures are built with rectangular and polar systems of coordinates; there are figures with even and uneven scales, according to the type of scale used.

Practice Exercises

Exercise 3.1

Analyze the data in Table 1.1. Identify a subject and predicate for each table, the type of table according to the subject and predicate, and the kind of grouping according to the nature of the defined task.

Table 1.1. Distribution of trade enterprises by type of ownership

Trade enterprises	Number of enterprises
State-run	10
Cooperative	15
Rental	5
Total	30

Table 1.2. Distribution of farm enterprises by level of automation of work

Level of automation, %	Number of enterprises	Cost of 1t of sugar beet, euro
40–60	10	36.0
60–80	25	30.0
80–100	15	26.0
Total	50	x

Table 1.3. Distribution of respondents characterizing risk-aversion or risk-taking attitudes to protect income level

Level of income	Propensity to take risks		Total
	Avoids taking risks	Likes to take risks	
High	14	9	23
Average	37	27	64
Low	100	28	128
Very low	151	30	181
Total	302	94	396

Table 1.4. Distribution of enterprise workers by level of remuneration

Wage, euro	Up to 1000	1000–1200	1200–1500	1500–2000	Over 2000	Total
Number of workers	10	25	50	20	15	120

Table 1.5. Distribution of doctors of science by age group

Age, years	% summary
younger than 40	18
40–50	15
50–60	37
older than 60	30
Total	100

Table 1.6. Characteristics of people's deposits in saving and commercial banks in two regions

Type of a bank	Average size of deposit, euro		Number of deposits by region, %		
	Region A	Region B	Region A	Region B	Average in two regions
Saving	150	240	95	85	90
Commercial	2850	3700	5	15	10

Exercise 3.2

Make models of the tables to characterize:

(1) the composition of people in the region by age (children; parents; grandparents), gender, and type of a settlement (cities; rural areas);
(2) the scope of services and the share of the services provided for the population by type of economic activities over two years;
(3) the structure of total costs of the population (consumer and non-consumer) depending on the type of settlement;
(4) the distribution of residents by place of residence and age group (too young to work; working age; and too old to work);
(5) the correlation between income level (low; sufficient; high) and income dynamics (decreased; did not change; increased).

Exercise 3.3

Below, we have some data on the living conditions of some workers.

№	Family size	Number of children	Number of rooms	Total area, м^2	Living area, м^2	Security available[1]	Location[2]
A	1	2	3	4	5	6	7
1	1	0	1	35	15	1	2
2	3	1	1	35	17	1	2
3	3	1	1	35	16	1	2
4	1	0	1	40	22	0	1
5	1	0	1	45	20	1	1
6	3	1	2	45	28	1	2
7	3	1	2	45	25	1	2
8	3	1	2	50	35	0	2
9	4	2	2	70	40	1	1
10	4	2	3	70	50	1	2
11	2	0	3	75	52	0	1
12	3	1	3	75	55	1	3
13	3	1	3	80	55	1	1
14	3	2	3	85	60	1	1
15	3	1	4	85	45	1	2
16	3	1	3	90	62	1	3
17	4	2	2	60	42	0	2
18	2	1	2	65	48	1	3

19	5	3	2	65	45	1	3
20	4	2	3	65	45	1	2
21	5	3	3	65	40	1	2
22	5	3	3	65	40	1	2
23	4	2	2	70	40	1	1
24	4	2	3	70	50	1	2
25	2	0	3	75	52	0	1
26	3	1	3	75	55	1	3
27	3	1	3	80	55	1	1
28	3	2	3	85	60	1	1
29	3	1	4	85	45	1	2
30	3	1	3	95	60	1	3
31	3	1	3	100	65	1	1
32	5	3	4	120	75	1	1
33	1	0	1	35	15	1	2
34	4	2	3	65	45	1	2
35	5	3	3	65	40	1	2
36	5	3	3	65	40	1	2
37	4	2	2	70	40	1	1
38	4	2	3	70	50	1	2
39	2	0	3	75	52	0	1
40	4	2	3	70	50	1	2
41	2	0	3	75	52	0	1
42	3	1	3	75	55	1	3

43	3	1	3	80	55	1	1
44	3	2	3	85	60	1	1
45	3	1	4	85	45	1	2
46	3	1	3	90	62	1	3
47	4	2	2	60	42	0	2
48	2	1	2	65	48	1	3
49	1	0	1	45	20	1	1
50	3	1	2	45	28	1	2

(1) 1 – security available; 0 – security not available; (2) 1 – new bedroom, suburbs; 2 – old districts; 0 – downtown.

From the given data, build:

(1) combined grouping by characteristics such as: family size – housing; family size – number of rooms; location – total area of apartment;
(2) analytical grouping to characterize dependence of location and total housing area; number of rooms and total area of apartment; and level of housing and number of children in the household.

Exercise 3.4

Using the data in the table, build an analytical grouping to characterize the correlation between volume of output and cost of fixed assets, forming four groups with equal intervals. Make a conclusion about the connection between the specified features.

№ enterprise	Average annual cost of fixed assets, million euro	Output for an accounting period, million euro
1	3.1	3.6
2	5.5	9.4
3	3.9	4.2
4	7.0	12.9
5	4.7	3.5
6	6.6	10.3
7	3.3	4.3
8	3.1	2.5
9	2.7	2.5
10	2.0	2.5
11	1.0	1.6
12	4.0	2.8
13	3.1	3.0
14	4.9	4.4
15	4.5	5.7
16	2.8	3.0
17	3.0	1.4
18	3.5	2.5

19	3.3	6.4
20	5.6	8.9
21	4.4	7.9
22	6.1	9.6
23	3.0	3.2
24	2.0	1.5

Exercise 3.5

Using the data in the table build an analytical grouping to characterize the correlation between bank revenue and the amount of capital of a company, forming four groups with equal intervals. Make a conclusion about the connection between the mentioned features.

№ bank	Capital of a company, million euro	Revenue per year, million euro
1	64.3	10.3
2	37.4	5.2
3	26.4	9.5
4	50.2	8.9
5	28.3	3.4
6	40.6	7.8
7	50.5	9.5
8	24.3	4.6
9	32.0	5.4
10	62.1	9.2
11	29.4	9.1

12	36.3	8.4
13	47.9	10.2
14	55.6	11.8
15	35.7	6.3
16	42.8	5.2
17	60.6	20.4
18	50.4	8.9
19	54.4	10.2
20	38.5	6.1

Exercise 3.6

Using the given data, make a statistical table. Identify a subject and a predicate; the kind of table by the subject and predicate; and the kind of grouping by the nature of the defined task. Give a name to the table and analyze the data.

In 2020, fixed assets put into operation in a country amounted to 1141 billion euro, including: agriculture, 75 billion euro; industry, 420 billion euro; and transport and communication, 168 billion euro. The volume of investments in fixed capital for the same period was: fully in a country, 75.7 billion euro; in agriculture, 3.4 billion euro; in industry, 28.2 billion euro; and in transport and communication, 15.0 billion euro.

Exercise 3.7

Using the given data, make a statistical table. Identify a subject and a predicate; the kind of table according to the subject and predicate; and the kind of grouping according to the nature of the set task. Give a name to the table. Analyze the data.

At the beginning of 2020, the cost of fixed assets in a country amounted to 1026 billion euro, including: private property, 44.9 %, and state-run and state-cooperative, 29.9 %. At the end of the year, this amounted to 1141

billion euro, including private property, 515.8 billion euro, and state-run and state-corporative, 359.1 billion euro.

Exercise 3.8

Using the given data, make a statistical table. Identify a subject and a predicate; the kind of table according to the subject and predicate; and the kind of grouping according to the nature of the set task. Give a name to the table. Analyze the data.

According to the volume of assets, banks can be classified into three groups: small, medium, and large. Their share of the general volume of assets is: 25 %; 45 %; and 30 %. The average profitability level of the assets is: 2.0 %; 2.5 %; and 4.0 %, respectively.

Exercise 3.9

Using the given data, group commercial banks by the profitability level of their own capital, depending on the volume of assets. Identify a subject and a predicate; the kind of table according to the subject and predicate; and the kind of grouping according to the nature of the set task. Give a name to the table.

№	Volume of assets, million euro	Profitability of own capital, %	№	Volume of assets, million euro	Profitability of own capital, %
1	90	9	8	200	18
2	60	7	9	250	20
3	110	10	10	300	22
4	150	11	11	65	7
5	130	15	12	220	18
6	70	8	13	160	12
7	100	11	14	180	14

Exercise 3.10

According to the given data, make an analytical grouping to characterize the effect of rates of profitability on loan operations, forming three groups with equal intervals. Make conclusions about the connection between the mentioned features.

№ bank	Average interest rate on credits, %	Profitability of loan operation, %
1	60	32
2	65	36
3	54	20
4	69	34
5	65	30
6	61	26
7	52	15
8	59	23
9	58	20
10	68	33
11	51	15
12	64	34
13	62	30
14	57	21
15	56	19
16	60	22
17	55	18
18	64	28

19	52	16
20	61	24

4. ABSOLUTE, RELATIVE, AND AVERAGE VALUES

4.1. Concepts and types of absolute values

Absolute statistical values express the sizes, volumes, and levels of events and processes. They can be classified into individual and summary values.

Individual absolute values express sizes of quantitative features for different population units (e.g., the volume of output produced by a construction worker per month). These are acquired directly in the process of collecting statistical data.

Summary absolute values characterize the size of a certain feature for all population units or for different groups; such values are derived from the generalization (usually summation or product) of absolute sizes of a feature for different population units (e.g., the volume of output produced by workshops per month). As a rule, these values are the results of calculation.

Absolute statistical values are always denominate numbers and are expressed in certain measurement units (natural units; conditionally natural units; labor units; cost (money); and time).

Physical (natural) units are used to characterize sizes and levels (e.g., centimetre (cm), meter (m), and square meter (m^2)); and volume (e.g., litre (l), (dm^3), scope (units), kilogram (kg), and ton (t)).

Nominal physical (conditionally natural) units are used if there is a need to combine types of data that have one common property. Recalculation into nominal physical units of measurement is done with the help of special coefficients. For example, an author's sheet, a unit of a literary work volume, is used to count print production in the publishing business and is equal to 40 thousand printed characters or 3000 cm^2 of printed material.

Labor units of measurement are used to determine time costs for a collection of people. For example, labor hours are measured in man-hours or man-days.

Cost units of measurement enable the combination of data types that have different properties. For example, sale volumes in a store, in cost terms, is a product of the unit price and the quantity of physically counted sold commodities.

Time units of measurement are used to determine time cost or duration (e.g., seconds, hours, days, and years). For example, the payback term of an investment project is measured in years.

For the presentation form of a measurement unit, absolute values can be simple (meter, kg) or complex (ton-km). For example, the scope of cargo transportation is determined as the cargo weight multiplied by the distance travelled.

4.2. Types of relative values and methodology of their calculation

Comparison and contrast are an **indispensable** part of any statistical analysis. The complex use of absolute and relative values is an important condition of statistical analysis.

Relative values characterize a quantitative correlation of two comparable values. All relative values are the result of the division of similar and dissimilar indicators. A comparable value is a numerator and the base of a comparison is a denominator. Relative values can be measured in times, percentages (%), and also per mille ($^0/_{00}$), per decimille ($^0/_{000}$), per centimille ($^0/_{0000}$).

By analytical function, there are several types of relative values:

- structures;
- coordination;
- comparisons – characterizing the correlation of similar indicators that relate to different objects or territories (for example, the amount of sulfuric anhydride in the air of city A is 1.5 times higher than the air of city B);
- dynamics;
- implementation of a plan;
- a planned task;
- intensities that show the extent of an event spread across a specified area (for example, level of meat production per capita).

The relative value of a structure characterizes the structure of a population or object. It is calculated as the relation of the size (scope) of a component to the total, which is a single object or a whole population: $d_j = \frac{f_j}{\sum f_j}$, where d_j indicates the part; f_j is the scope (size) representing a part of the population (object); and $\sum f_j$ is the whole (the total). Thus, the value achieved is called *a percentage*. The sum of all *percentages* is equal to one or 100 %. A comparative analysis of the difference between populations and the estimation of structural shifts over time are based on *percentages*. The difference between *percentages* is measured in percent points (p.p.). Coefficients of structural shifts are defined so as to evaluate structural shifts over time. They are linear or mean-quadratic.

A linear coefficient is calculated with the formula:

$$l_d = \frac{\sum_{j=1}^{m} |d_{j1} - d_{j0}|}{m},$$

while a mean quadratic coefficient is calculated with the formula:

$$\sigma_d = \sqrt{\frac{\sum_{j=1}^{m} (d_{j1} - d_{j0})^2}{m}},$$

where d_{j1} and d_{j0} are *percentages* in the current period and in the base period respectively; and m represents the number of structural components.

Example 4.1

Using the data on the structure of investments in nonfinancial assets for two years (Table. 4.1), make a conclusion about structural shifts in sources of financing.

Table 4.1. Structure of investments in nonfinancial assets by sources of financing in percent (%) of the total

Sources of financing	**1st year** d_{j0}	**2nd year** d_{j1}	$\lvert d_{j1} - d_{j0} \rvert$	$(d_{j1} - d_{j0})^2$
State budget funds	4.5	7.0	/7-4.5/=2.5	2.5^2
Local budget funds	3.5	4.1	/4.1-3.5/=0.6	0.6^2
Funds from enterprises and organizations	70.9	61.4	/61.4-70.9/=9.5	9.5^2
Foreign investor funds	4.8	5.5	/5.5-4.8/=0.7	0.7^2
Savings	3.4	3.6	/3.6-3.4/=0.2	0.2^2
Bank credits	4.4	8.2	/8.2-4.4/=3.8	3.8^2
Other sources of financing	8.5	10.2	/10.2-8.5/=1.7	1.7^2
Total	100.0	100.0	19.0	114.72

Analysis of the structure of the sources of financing leads us to make a conclusion about the poor development of capital market – more than two thirds of all capital investments are made up with their own funds. This means that enterprises deal with the financing of nonfinancial assets and this in turn influences the level of investment activity, which is estimated to be very low.

Evaluating structural shifts, one can make a conclusion that, in the second year, on average the specific weight of each source of funding changed by 4 p.p., leading to a slight improvement in the structure of financing sources.

$$l_d = \frac{\sum_{j=1}^{m} |d_{j1} - d_{j0}|}{m} = \frac{19}{7} = 2.71 \text{ p. p.} \text{ or } \sigma_d = \sqrt{\frac{\sum_{j=1}^{m} (d_{j1} - d_{j0})^2}{m}} = \sqrt{\frac{114.72}{7}} = 4.05 \text{ p. p.}$$

When a large number of structural elements, for example for the analysis of structural changes in terms of regional and branch diversification, are

available, it is advisable to use chain indicators on the degree of structural shift, which can be evaluated as a half-sum of the individual (regional) chain deviations of specific weights by module $\boldsymbol{KS} = \frac{\sum_{j=1}^{n}|d_{jt} - d_{jt-1}|}{2}$, where $\boldsymbol{d_{jt}}$ is the *percentage* of $\boldsymbol{j}$ region (economic activity) in period $\boldsymbol{t}$ and $\boldsymbol{d_{jt-1}}$ is the *percentage* of $\boldsymbol{j}$ region (economic activity) in the previous period.

Using data about the structure of investments in fixed assets according to the kind of economic activity in a country for four years (Table 4.2), let us estimate the intensity of structural shifts in the dynamics (Table 4.3).

Table 4.2. Structure of investments in fixed assets according to kind of economic activity in a country, in percent (%) of the total

Type of activity	1st year	2nd year	3rd year	4th year
Agriculture, hunting	4.96	5.19	4.20	4.47
Fisheries	0.06	0.09	0.08	0.05
Industry	41.91	40.65	38.67	37.23
Construction	3.40	4.90	4.90	6.17
Wholesale and retail	3.94	5.43	6.42	7.03
Hotels and restaurants	1.44	1.20	1.33	1.42
Transport and communication	22.8	18.84	20.05	19.83
Financial activity	1.63	2.40	1.92	1.64
Operations with real estate	13.95	14.93	15.13	14.84
Public administration	1.01	1.21	1.55	1.53
Education	1.25	1.56	1.28	1.26
Healthcare and social care	1.46	1.55	2.06	1.94
Collective, public and individual service	2.11	2.05	2.41	2.58
Total	100.00	100.00	100.00	100.00

Looking at the results of the analysis of coefficients on the extent of a structural shift according to type of economic activity, one can make the conclusion that there has been a change (decrease) in the structure of investments in fixed assets according to type of economic activity (Table 4.3).

Table 4.3. Calculation of the degree of structural shift in investments in fixed assets according to kind of economic activity in a country, in p.p.

Type of activity	2nd year/ 1st year	3rd year/ 2nd year	4th year/ 3rd year
Agriculture, hunting	5.19-0.96=0.23	|4.2 – 5.19|=0.99	4.47-4.2=0.27
Fisheries	0.03	0.01	0.03
Industry	1.26	1.98	1.44
Construction	1.50	0.00	1.27
Wholesale and retail	1.49	0.99	0.61
Hotels and restaurants	0.24	0.13	0.09
Transport and communication	4.04	1.21	0.22
Finance activity	0.77	0.48	0.28
Operations with real estate	0.98	0.20	0.29
Public administration	0.20	0.34	0.02
Education	0.31	0.28	0.02
Healthcare and social care	0.09	0.51	0.12
Collective, public, and individual service	0.06	0.36	0.17
Total	11.20	7.48	4.83
Coefficients of structural shift	11.2/2=5.60	7.48/2=3.74	4.83/2=2.42

If a dimensional estimate or an estimate of an object by two features is undertaken, a coefficient of a structural similarity can be calculated with the formula:

$$K_{sim} = 1 - \frac{\sum_{j=1}^{m} \left| d_j^1 - d_j^2 \right|}{2},$$

where d_j^1 and d_j^2 refer to the specific weight of the *j*-component for the 1st and 2nd objects respectively, or for two features. If the structures are similar, then the coefficient of similarity approaches one. As such, the larger the difference between structures, the lower the coefficient of similarity.

The relative value of coordination characterizes the ratio of different parts of the whole. One of them is taken as a base for comparison

$$K_{coor} = \frac{f_j}{f_k},$$

where the volume (size) of f_j characterizes a comparable part of a population (object) and the volume (size) of f_k characterizes the comparison base of a population (object).

For example, if the population of a city is 350 thousand people, including 192.5 thousand women and 157.5 thousand men. The share of women is 192.5/350 = 0.55 (or 55 %), while the share of men is 157.5/350 = 0.45 (or 45 %). The correlation of women to men is 192.5/157.5 = 1.(2). As such, there are 122 women per 100 men. A comparison base is chosen randomly, but, if the components display a serious difference in size, then it is better to put the bigger value in the position of the numerator when undertaking calculation and further interpretation.

The relative value of comparison is used in territorial (spatial) comparisons and is defined as a ratio of similar indicators that characterize different objects (populations) or territories (regions) and belong to a particular period or moment in time $K_{comp} = \frac{a}{b}$, where *a* is the reference indicator and *b* is an indicator that characterizes a comparison base. The comparison base is chosen randomly. The interpretation of a computed indicator depends on the choice of a comparison base. For example, in a country the number of

visits to theaters per year is 6.2 million, while 17.5 million visits are made to museums. If the number of theater visits is taken as a comparison base, then we can see that the level of museum visits is 2.8 times higher than the level of theater visits (17.5/6.2 = 2.8). If the number of museum visits is the comparison base, then the level of theater visits achieves only a third of museum visits (6.2/17.5 = 0.35). Additionally, a normative or standard value can be used as a comparison base. For example, the average daily consumption rate for a man is 3500 Kcal. If actual consumption is 3000 Kcal, then consumption is 14.3 % less than the average (3000/3500 = 0.857).

The relative value of dynamics characterizes the change in an event over a period of time in terms of a direction and intensity: $= \frac{y_1}{y_0}$, where t is the relative value ***of the dynamics***; y_1 is the current value of an indicator; and y_0 is the value of an indicator in the period chosen as a comparison base. A comparison base can be based on a previous level or on a level that is more distant in time. For example, in Ukraine as of 01/01/89, the population was 51.7 million people; on 01/01/04, the population was 47.6 million people; and by 01/01/05, the population had become 46.9 million. As we can see, the population of Ukraine decreased by 1.47 % (46.9/47.6 = 0.9853) over a year. However, when compared to 1989, the decrease in population is 9.28% (46.9/51.7 = 0.9072).

The relative dynamics value can be presented in the form of a product of two values: the relative value of a planned task, $\frac{y_{pl}}{y_0}$, and *the relative value of a plan implementation*, $\frac{y_1}{y_{pl}}$. Then, the correlation is

$$t = \frac{y_1}{y_0} = \frac{y_{pl}}{y_0} \times \frac{y_1}{y_{pl}} \cdot$$

This means that the relative dynamics value can be calculated without absolute values. For example, a plan was made to decrease production by 4% (i.e. to the level of 96 %); the plan was performed and 98 % was achieved. In fact, the inputs decreased by 5.98 % (0.96 × 0.98 = 0.9408) so that, the overall decrease was 1.98 p.p. (4 % – 5.98 % = -1.98).

The relative intensity value is calculated as the correlation of two values that have a certain interconnection. The relative intensity value is a named

value for which the measurement units of a numerator and a denominator are combined to form a new measurement unit and a new indicator. Relative intensity values can be divided into two groups: quantitative values that characterize a quantitative correlation of two indicators (e.g., GDP per capita, birth rate per 1000 people, and sickness rate per 100 thousand people) and analytical relative intensity values, which, in addition to correlations, have analytical load [e.g., harvest ratio (centner (cwt)/hectare (ha)) = gross harvest/cultivation area and the profitability of assets (euro/euro) = revenue/asset cost].

4.3 The concept of the average in statistics and kinds of average values

In statistics, **the average** is an abstract, generalized value, which shows the level of a variable feature in a homogeneous population. The fluctuation of individual values of a feature due to the effect of various factors is balanced out in an average value.

The average values used in statistics belong to a degree class and, in a generalized form, they look like,

$$\bar{x} = \sqrt[k]{\frac{\sum x_i^k}{n}},$$

where x_i refers to the individual values of a variable feature; k is the indicator of the degree of the average; and n is the number of feature values (sample size). *A definite form of the average depends on its degree. The main types of degrees of averages are given in Table 4.4.*

Table 4.4. Formula for degrees of averages

k	General form of function	Name of average	Calculation formula	
			simple	weighted
1	$f = \frac{1}{x}$	Arithmetic mean	$\frac{\sum x}{n}$	$\frac{\sum x \times f}{\sum f}$
-1	$f = \log x$	Harmonic mean	$\frac{n}{\sum \frac{1}{x}}$	$\frac{\sum z}{\sum \frac{z}{x}}$
0	$f = x$	Geometric mean	$\sqrt[n]{\prod \langle x \rangle}$	$\frac{\sum ln\, x \times f}{\sum f}$
2	$f = x^2$	Mean square	$\sqrt{\frac{\sum x^2}{n}}$	$\frac{\sum x^2 \times f}{\sum f}$

The arithmetic mean is used when distribution regularities are studied; the mean square is used when variations are studied; and the geometric mean is used when dynamics is studied. In mathematical statistics, different kinds of average are calculated from the same data to give different values. The correlation between them is called *a rule of majority* and is expressed as $\bar{x}_{square} > \bar{x}_{arithm} > \bar{x}_{geom} > \bar{x}_{harmon}$. As such, the harmonic mean is smaller than the geometric mean, which, in turn, is smaller than the arithmetic mean, and the latter is smaller than the square of the mean. Other forms of average (third, fourth degree, etc.) also exist in the theory of statistics.

The higher the degree of an indicator, the larger the average value. In socioeconomic statistics, the calculation of various averages for the same population is not advisable, so the choice of the type of average in every specific case of research is of importance. A formula for the average value is chosen for a definite purpose:

1) to identify the nature and specific aspects of the event under study;
2) to define the purpose of the calculation of the average value and its indicator;
3) to find a formal form for the indicator – a determinative function;

4) to make an equation for the average, replacing values of individual features for each unit of a population with the average value of a determinative function;
5) to identify a specific formula for the average value, which is calculated from an equation of the average. The components, which constitute an equation of the average, must be connected to the contents in a way that helps us get the dimension of the indicator.

Let us examine the conditions with a few examples for the calculation of averages.

The arithmetic mean is one of the most widely used in cases where the variable feature represents the whole population and displays the sum of individual values. The simple form of the arithmetic mean is calculated for non-grouped data, while the weighted form of the arithmetic mean is calculated for grouped data. For example, if we have a list of workers of a building team with data about individual wages per month, it is likely easier not to look for those people who earned the same amount of money for any period, but rather to sum all the wages and divide by the number of team workers: $\bar{x} = \frac{\sum x_i}{n}$, where x_i is individual wages.

In another example, if a department has various groups of employees, say professors, associate professors, laboratory assistants, and so on, the average salary of these departmental employees is calculated as $\bar{x} = \frac{\sum x_j \times f_j}{\sum f_j}$, where f_j is the number of employees who hold a similar position. In this case, frequency plays the role of *weight* and so the average is described as *weighted*. In both cases, the results of calculation will be equal.

If fraction (w) is used as the weight, then the formula is:

$$\bar{x} = \frac{\sum x_j \times w_j}{\sum w_j} = \frac{\sum x_j \times w_j}{100},$$

where w_j is given in percent, and $\bar{x} = \sum x_j \times w_j$, where w_j is the coefficient.

If the average is calculated for an interval distribution row, then individual values are calculated as the half-sum of two interval limits. The width of an open interval is based on the width of a nearby interval. The calculation of the average of the nominal values (average percentage and average specific weight) has a peculiarity. Weight refers to denominators of those correlations that are used to calculate individual relative indicators.

Example 4.2

Using the given data, calculate the average interest of the implementation plan for two teams (Table 4.5). There is a logical assumption that both teams have implemented the plan, on average, to the value of 103 %. But the average indicator will move closer to the team that has a higher specific weight in the scope of a general plan, i.e. to Team No. 1.

Table 4.5. Implementation of the output plan by the teams in the workshop

Team	Plan implementation, %	Plan output, items
№ 1	101	600
№ 2	105	160

Thus, $\bar{x} = \frac{1.01 \times 600 + 1.05 \times 160}{600+160} \times 100\ \% = 101.8\ \%$.

When calculating the weighted arithmetic mean, a fraction rather than the frequency is taken as the weight. ***The simple form of the harmonic mean*** is used for inverse values.

Example 4.3

For example, we have some data about the hours that three workers spent on one element of production: ½ hour, 1/3 hour, and 1/7 hour. To calculate the average amount of time spent on this production element, the formula is $\bar{x} = \frac{1+1+1}{\frac{1}{1/2}+\frac{1}{1/3}+\frac{1}{1/7}} = \frac{1}{4}$ hour.

Now, we illustrate the weighted formula of the harmonic mean (example 4.4).

Example 4.4

We have some data about the average output per worker and the volume of output for two teams per month (Table 4.6). We can calculate the general average output per month for two teams.

Table 4.6

Team	Actual production, thousand euro (Q)	Productivity per worker, thousand euro (w)	Number of team members
Specialized	720	6	8
Complex	900	5	12
Total	1620	x	20

To solve this exercise, it is practical to start with the economic essence of the indicator, i.e. the average output per worker is equal to $= \frac{Q}{T}$. If no data for the number of workers are available (T), i.e. the frequencies are not known (f), then this is computed with the formula for each team as $T = \frac{Q}{w}$. So, in our sample, we use the following formula for the weighted harmonic mean: $\bar{x} = \frac{\sum z}{\sum \frac{z}{x}} = \frac{720+900}{\frac{720}{6}+\frac{900}{5}} = 5.4$, where x is the average output per worker in each team and z is the actual volume of the output.

In this example, the average output per worker per month for two teams was 5.4 thousand euro.

Example 4.5

We have some data about production and production cost for two subdivisions of an enterprise for two quarters of the year (Table 4.7). We can determine the average production cost in each quarter.

To choose the appropriate form of the average, we first define the nature of the initial data, i.e. what is x and what is f (frequency).

Table 4.7

№ sub-division	Quarter - I		Quarter - II		Quarter - I $x \times f$	Quarter - II $\frac{z}{x}$
	Production cost (x), euro	Output volume (f), thousand items	Production cost (x), euro	Production expenses (z), thousand euro		
A	1	2	3	4	5	6
1	7.0	6	6.5	26.0	42	4
2	11.0	4	10.8	64.8	44	6
Total	x	10	x	90.8	86	10

The production cost is x and f is the output volume. Thus, in the first quarter of the year, the average production cost can be calculated as a weighted arithmetic mean (column 5, Table 4.7):

$$\bar{x} = \frac{86}{10} = 8.6 \text{ euro.}$$

Having looked at the data in the second quarter carefully, we can easily notice the absence of frequency value (f). Instead, we find the output volume in value terms, which is the product of production cost and output volume ($x \times f$), defined as z. So, in this case, it is realistic to use the formula for the weighted harmonic mean (column 6, Table 4.7)

$$\bar{x} = \frac{90.8}{\frac{26}{6.5} + \frac{64.8}{10.8}} = \frac{90.8}{4+6} = 9.08 \text{ euro.}$$

Hence, during the second quarter, in two sub-divisions production costs increased by 0.48 euro on average, regardless of the decrease in production costs in each sub-division. This increase was caused by the change in the output structure. The increased share in total output of the second sub-division resulted in a production cost increase.

In the literature, one can find a number of recommendations related to the identification of averages with the characteristics of ordinal and nominal scales. The authors believe that, when the range of an ordinal scale presents almost equal distances between different qualities of the event, the average range can be calculated in the same way that the characteristics of a metric

scale are measured. An average level of qualification (category) and an average attestation point and others are given as an example. In our opinion, "distance similarity" in the given examples is rather disputable. Therefore, it is understood that, in some cases, ranges can be positive and negative numbers. It is suggested that we capture the levels of worker satisfaction by profession – "satisfied", "indifferent", or "unsatisfied" marked 1, 0, and -1 respectively – to determine the arithmetic mean for the whole team. However, the results from these procedures can be rather nominal and we should be very cautious in their use.

4.4. The properties of the arithmetic mean

The properties of the arithmetic mean are:

1. The algebraic sum of deviations in all individual values from the average is equal to null: $\sum(\boldsymbol{x_i} - \overline{\boldsymbol{x}}) = \boldsymbol{0}$;
2. With the increase or decrease of each individual value by any constant value, the average changes by the same value: $\frac{\sum(\boldsymbol{x_j} \pm \boldsymbol{a})\times \boldsymbol{f_j}}{\sum \boldsymbol{f_j}} = \overline{\boldsymbol{x}} \pm \boldsymbol{a}$;
3. With the division or multiplication of each individual value by any number, the average decreases or increases by the factor of that number: $\frac{\sum(\boldsymbol{x_j} \times [\div]\, \boldsymbol{a})\times \boldsymbol{f_j}}{\sum \boldsymbol{f_j}} = \overline{\boldsymbol{x}} \times [\div]\boldsymbol{a}$;
4. The average does not change by increasing or decreasing the frequencies of all individual values by the same factor: $\frac{\sum \boldsymbol{x_j} \times (\boldsymbol{f_j} \times [\div]\, \mathbf{a})}{\sum (\boldsymbol{f_j} \times [\div]\, \mathbf{a})} = \overline{\boldsymbol{x}}$;
5. The sum of the square deviation of the individual values from the average is smaller than for any other value: $\sum(\boldsymbol{x_i} - \overline{\boldsymbol{x}})^{\boldsymbol{2}} \rightarrow \boldsymbol{min}$.

Therefore, a conclusion can be made from the formula that the change in a population structure influences the average. An example of this is presented using data about the salaries and number of employees in two departments with professors and lab-assistants (Table 4.8).

Example 4.6

Table 4.8. Remuneration of departmental staff for two periods

Position	Department I		Department II	
	Remuneration, euro	Number of employees	Remuneration, euro	Number of employees
Professor	2000	4	2000	1
Lab-assistant	500	1	500	4
Total	x	5	x	5

We can calculate an average salary of the first department using the formula of the weighted arithmetic mean

$$\bar{x} = \frac{2000 \times 4 + 500 \times 1}{4 + 1} = 1700 \text{ euro},$$

and for the second department it will be equal to

$$\bar{x} = \frac{2000 \times 1 + 500 \times 4}{1 + 4} = 800 \text{ euro}.$$

So, under similar conditions of labor remuneration and department size, the average saw a smaller value due to the structural change in the composition of departmental personnel.

4.5. The multivariate mean

The multivariate mean is a complex characteristic that is used to study complicated events and processes and is presented with a series of indicators. Each indicator of the series is independent, while being part of the generalized characteristics. Since the indicators in the series are expressed, as a rule, by different dimensions, we can undertake a valuation of all the initial indicators. This valuation involves adjusting the indicators into one form to achieve *standardization*. In this standardization, the individual values of different indicators are substituted for relative values

or ranges (points). The question remains as to what we take to be the norm for standardization. The valuation can be the maximum level achieved, the minimum level achieved, or the average level achieved, and can be expressed with the formulas:

$$x'_{ij} = \frac{x_{ij}}{\overline{x_i}};\ x'_{ij} = \frac{x_{ij}}{x_i^{\max(\min)}};\ x'_{ij} = \frac{x_{ij}}{x_i^{\text{standard}}},$$

where x'_{ij} is the standardized value of the indicator i for the unit of population j; x_i is the average value; x_{ij} is the value of indicator i for unit of population j; $x_i^{\max(\min)}$ is the maximum or minimum value of indicator i; and x_i^{standard} is the standard value of indicator i.

In making a valuation, we need to take into consideration that all indicators are divided into stimulants and de-stimulators. **Stimulants** are indicators whose increase symbolizes positive dynamics; a standardized value of more than one corresponds to a high level of an indicator for different units of a population (e.g., labor productivity, profitability, and profit). **De-stimulators** are indicators whose increase is characterized by negative changes and a high level of an indicator for different units of a population is indicated by the value of a standardized indicator that is less than one (e.g. production cost, labor input, or expenses). To adjust for unique interpretations in the valuation of indicators, de-stimulation is done with the inverse formula $\frac{1}{x'_{ij}}$.

The multivariate mean can be calculated as a balanced one with the formula of the arithmetic mean, simplified from the standardized estimation $\overline{x'_j} = \frac{\sum x'_{ij}}{m}$, where m is the number of indicators. If the indicators have different weights (different effects), then each indicator is given a definite weight d_i ($0<d_i<1$) and the multivariate mean is calculated as

$$x'_j = \sum x'_{ij} \times d_i.$$

Based on the maximal and standard values, the value of the multivariate mean lies within 0 and 1. If valuation is done on the average level, the multivariate mean is larger than one, implying that the occurrence of the event under study for j units of a population is higher than the average level; if the multivariate mean is smaller than one, then the occurrence is lower than the average level.

Example 4.7

Let us make a comparative analysis of the indicators of socioeconomic development for several regions in a country over a period of two years (Table 4.9) using the multivariate mean.

Table 4.9. Main indicators of socioeconomic development of regions of a country

Region	Unemployment rate according to International Labor Organization, %		Indices of industrial produce, in % of previous year		Investments in fixed assets, billion euro		Real income per capita, thousand euro		Wages of hired workers, euro	
	Year - 1	Year - 2	Year - 1	Year - 2	Year - 1	Year - 2	Year - 1	Year - 2	Year - 1	Year - 2
Country	11.1	9.1	114	116	32.6	51.0	2.4	3.4	311	462
Region A	6.6	6.7	109	119	1.2	1.9	2.4	3.4	301	433
Region B	8.7	7.4	109	110	2.8	4.2	2.1	2.7	370	526
Region C	12.7	10.5	143	114	1.4	2.6	2.2	3.0	272	419
Region D	10.8	10.4	102	121	1.3	2.2	2.7	3.7	379	541
Region E	17.1	13	119	135	0.3	0.5	1.9	2.4	190	304
Region F	11.,8	9.6	117	111	2.0	3.6	2.4	3.3	310	455

To do an analysis using the multivariate mean, we expect to engage in valuation of all indicators. In this example, valuation is done at the level of a country (Table 4.10).

Table 4.10. Standardized indicator values

Region	Unemployment rate		Indices of industrial produce		Investments in fixed assets		Real income per capita		Wages of hired workers	
	Year - 1	Year - 2	Year - 1	Year - 2	Year - 1	Year - 2	Year - 1	Year - 2	Year - 1	Year - 2
Region A	1.6818	1.6567	0.9561	1.0259	1.0005	1.0128	1.0192	0.9859	0.9678	0.9372
Region B	1.2759	1.5000	0.9561	0.9483	2.3174	2.2479	0.8642	0.7976	1.1897	1.1385
Region C	0.8740	1.0571	1.2544	0.9828	1.1239	1.3548	0.9167	0.8824	0.8746	0.9069
Region D	1.0278	1.0673	0.8947	1.0431	1.0817	1.1382	1.1329	1.0765	1.2186	1.1710
Region E	0.6491	0.8538	1.0439	1.1638	0.2269	0.2414	0.7713	0.7203	0.6109	0.6580
Region F	0.9407	1.1563	1.0263	0.9569	1.6192	1.8815	1.0104	0.9735	0.9968	0.9848

For the indicator “unemployment rate” (as an indicator of de-stimulation), valuation is done in reverse (the level of a country is divided by the level of the region). For the indicator “investments into fixed assets”, the level of a country is divided by the number of regions to find the average level of investments for each region and then the standardization of valuation is undertaken.

Let us calculate the multivariate mean as a simple arithmetic mean from standardized values and create its rating on this basis (Table 4.11). An integral estimate for a country is 1.0.

Table 4.11. Rating of the regions of a country based on the multivariate mean

Regions	Integral estimate		Rating	
	Year - 1	Year - 2	Year - 1	Year - 2
Region A	1.125	1.124	2	3
Region B	1.321	1.326	1	1
Region C	1.009	1.037	5	5
Region D	1.071	1.099	4	4
Region E	0.660	0.727	6	6
Region F	1.119	1.191	3	2

The results of analysis using the multivariate mean substantiate the conclusion that the regions under study (except for region E) are at a higher level of socioeconomic development than the average level for the country. Region B stands out particularly against the general background. The increase in the multivariate mean in the second year (except for the Region A where the estimate shows a marginal decrease) confirms the positive dynamics of the indicators characterizing socioeconomic development.

Practice Exercises

Exercise 4.1

A company planned to increase the volume of its output by 10 % in the fourth quarter of the year. In reality, the company increased output by 7 %. Calculate how well the company achieved its plan to increase output volume.

Exercise 4.2

A company planned to decrease production costs in the fourth quarter of the year by 5.5 %. In reality, the company decreased production costs by 7 %. Calculate how well the company achieved its plan to decrease production costs.

Exercise 4.3

A company planned to decrease labor inputs in the third quarter of the year by 9 %. This plan was achieved to the level of 110 %. Calculate the real decrease in labor input.

Exercise 4.4

A company planned to increase labor productivity in the fourth quarter of the year by 5.5 %. This plan was 93 % achieved. Calculate the real increase in labor productivity.

Exercise 4.5

With a price increase of 25 %, demand decreased by 40 %. Calculate the coefficient of price elasticity. Make a conclusion about demand elasticity.

Exercise 4.6

A company planned to decrease production costs in the fourth quarter of the year by 5.5 %. This plan was 107 % achieved. Calculate the real decrease of production costs (as a percentage value) and, also, by how much production costs decreased if, in the third quarter of the year, they were 156 thousand euro.

Exercise 4.7

In the third quarter of the year, production volume increased by 10 % compared to the first half of the year. In the fourth quarter of the year the company planned to increase production volume by 7 %. Calculate how much the company achieved through its plan to increase production volume, as well as how much production increased in the second half of the year, if, in the first half of the year, the volume of production was 156 thousand euro.

Exercise 4.8

In the first quarter of the year, production volume increased by 5 % compared to the fourth quarter of the previous year. An increase of 8 % was planned for the second quarter. This plan was 99 % achieved. How much did production volume increase in the first half of the year, if it was 254 thousand euro in the fourth quarter of last year?

Exercise 4.9

With an increase in supply of 25 %, prices dropped by 15 %. Calculate the coefficient of price elasticity. Make conclusions about price elasticity.

Exercise 4.10

A wage of 1500 euro was received with an income tax rate of 1 %. How much money was charged and what amount of tax was withheld?

Exercise 4.11

A worker received 2000 euro, the amount of compulsory payments was 2.5 % and the income tax rate was 17 %. What quantity of compulsory payments and taxes were deducted all together and separately for each deduction?

Exercise 4.12

The price of a commodity with Goods and Service Tax (GST) is 15 euro; the GST rate is 19 %. Calculate the commodity price without GST and the GST amount.

Exercise 4.13

In the first quarter of the year, prices of dairy products increased by 10 %, and then by 5 % more in the second quarter. Prices decreased by 8 % on

average in the third quarter of the year. By how much did these prices change?

Exercise 4.14

The price of commodity A decreased by 15 % and then by 10 % more. The price of commodity B decreased by 10 % and then by 15 % more. The price of which commodity decreased more intensively? Support your conclusions with calculations.

Exercise 4.15

Identify the absolute and relative indicators:

(1) an increase in real gross domestic product by 2.6 %;
(2) an increase in average nominal monthly wage by 36.7 %;
(3) if the number of registered unemployed people at the end of the year was 881.5 thousand;
(4) an unemployment rate of 7.2 %;
(5) the natural reduction of the population is 356 thousand;
(6) budget revenue is 134183 million euro;
(7) fixed assets in real prices were 1249 billion euro at the end of the year;
(8) the index of consumer prices for December of the previous year was 110.3 %;
(9) the share of investments in fixed assets is 84 %;
(10) the number of retirees per 1000 people is 301;
(11) the birth rate per 1000 people is 9;
(12) life expectancy at birth among men is 12 years shorter than for women;
(13) there are 55 divorces per 100 marriages;
(14) 106 boys were born for 100 girls.

Exercise 4.16

The life expectancy of men and women at birth is characterized as follows.

Birth year	All population	Men	Women
1985–1986	70.5	65.9	74.5
2004–2005	68.0	62.2	74.0

Make a comparative analysis of life expectancy. What type of relative values is calculated?

Exercise 4.17

Using the results of bond bidding below, make a comparative analysis of bond quotations. Make some conclusions. What type of relative values is calculated?

Bond issuer	Bond cost, euro		
	Nominal	Market	
		1st day	2nd day
A	1000	900	910
B	5000	5050	5075

Exercise 4.18

Using the data, determine:

(1) the dynamics of cast iron production in a metallurgical plant;
(2) the structure of cast iron production by type in each year;
(3) structural changes.

Make a conclusion.

Type of cast iron	Production, thousand items		Recalculation coefficient on pig (steel making) iron
	Last year	Current year	
Foundry	125	75	1.15
Chrome nickel	50	70	1.50
Ferro manganese	75	120	2.50
Ferro phosphoric	35	50	4.00

Exercise 4.19

From the given data about production and consumption of primary energy carriers (million tons of nominal fuel) in different countries, determine the relative values for each country, reflecting:

(1) the dynamics of consumption and production of energy carriers;
(2) the extent of self-sufficiency of a country in power resources;
(3) the dynamics of the dependence of a country on primary resource imports.

Make a conclusion.

Country	Consumption of power resources		Production of power resources	
	Previous period	Current period	Previous period	Current period
A	880	1074	760	1060
B	450	510	440	500
C	607	550	730	845
D	210	290	260	310

Exercise 4.20

From the data, determine:

(1) the dynamics of preserved food production in physical form at a cannery;
(2) the structure of preserved food production in each year;
(3) structural changes in preserved food production.

Make conclusions.

Volume of jar/can, liters	Production, thousand Jars/cans		Conversion coefficient into nominal jars/cans
	Previous year	Current year	
0.5	45	48	1.305
1.0	42	43	2.611
2.0	37	37	5.222
3.0	28	25	7.833

Exercise 4.21

From the results of a selective check of product quality, the following data were received.

Type of goods	Returned for correction, items	Share of returned goods, %
Sewing products	300	6.5
Knitwear	150	5.0
Total	450	x

Determine the average share of the goods returned for correction. Explain the use of the form of the average.

Exercise 4.22

Data on the distribution of farm enterprises by harvest ratio (centner (cwt)/hectare (ha)) are presented below.

Harvest ratio, cwt/ha	Number of farm enterprises	Cultivation area, in % of total
15–17	32	5
17–19	43	25
19–21	95	50
21–23	20	20
Total	200	100

Determine the average harvest ratio. Explain the use of the form of the average.

Exercise 4.23

Below we have some data on the stores in a region.

Type of store	Average sale amount per salesperson, thousand euro	Total turnover of goods per year, million euro
State-run	50	150
Rental	52	120
Private	60	180
Total	x	450

Determine the average sales per salesperson in all types of store. Explain the use of the form of the average.

Exercise 4.24

Data on the rejection rate of a selective check of manufactured cloth, are presented below.

Type of cloth	Checked cloth amount, million meter	Rejection rate, %
Cotton	116.0	10.0
Woolen	9.1	6.6
Silk	26.8	7.1
Total	151.9	x

Determine the average rejection rate of all checked cloth. Explain the use of the form of the average.

Exercise 4.25

Below are some data about the scope of services per capita and the total scope of services in two cities.

City	Scope of provided services per capita, euro	Total scope of provided services, million euro
A	1318	2686
B	364	114
Total	x	2800

Determine the average scope of provided services per capita. Explain the use of the form of the average.

Exercise 4.26

From the per month data below, determine the average share of high-quality output and average labor productivity. Explain which type of average should be used and why.

Nature of team	Number of teams	Volume of output produced per month, million euro	Share of high quality output, %	Labor productivity of one worker, thousand euro
Specialized	8	720	96.0	6000
Complex	12	900	92.6	5000
Total	20	1620	x	x

Exercise 4.27

Use the method of the multivariate mean and the tabulated data below to build a rating of some countries according to their demographic situation. Use the best indicators for your valuation.

Countries	Demographic indicators			
	Death coefficient, ‰	Child death coefficient, ‰	Senescence coefficient, %	Expected life expectancy, years
A	13.0	12.0	24	73
B	16.5	13.5	27	69
C	14.0	11.5	25	74
D	17.0	13.9	30	68
F	16.0	12.9	28	71

Exercise 4.28

Use the method of the multivariate mean and the tabulated data below to build a rating of banks according to profitability indicators. Take the lowest indicators for valuation.

Banks	Indicators			
	Profitability coefficient of assets, %	Profitability coefficient of own capital, %	Net operational margin, %	Income spread, %
A	5.0	25	30	5.5
B	4.0	22	32	6.5
C	3.5	24	29	5.0
D	1.5	10	25	4.2
F	2.0	15	24	4.0

Exercise 4.29

Two years of data on the "credit portfolio" of a commercial bank are presented in the following.

Type of credit	Previous year	Current year
All allotted credits, including:	7.61	10.77
Interbank	1.07	1.98
Short-term	5.52	6.03
Long-term	1.02	2.76

Using the data, analyze the structure and structural changes in the correlation of separate components of the "credit portfolio" and the dynamics of the indicators. Make conclusions.

Exercise 4.30

Using the data below, determine the average share of citizens' deposits in funds raised by city banks. Explain the choice of the form of the average.

Banks	Citizens' deposits, million euro	Share of raised funds, %
A	16.6	24.9
B	8.6	23.8
C	2.8	16.4

Exercise 4.31

Two years of data on the "credit portfolio" of a commercial bank are presented in the following (million euro).

Type of support	Previous year	Current year
All allotted credits including:	7.92	10.77
Guaranteed by other banks	0.51	0.55
Overdue	0.62	0.66
Prolonged	6.58	9.20
Doubtful to be returned	0.21	0.36

Use the data to analyze the structure and structural changes; the coordination of separate components of the "credit portfolio"; and the dynamics of the indicators. Make conclusions.

Exercise 4.32

Using the following data about the amount of investments in the national economy of the country (billion euro), determine the relative values of the structure, coordination, and dynamics. Make appropriate conclusions.

Type of investments	Period	
	II half-year of the past year	II half-year of the accounting year
Direct	0.7	2.8
Portfolio	2.0	13.0
Others	8.9	7.8
Total	11.6	23.6

Exercise 4.33

The annual income from the shares of two different companies must be 18 and 20 %. Determine the income of each investor in a securities portfolio if the investors were to distribute their deposits in the shares of companies in the proportions presented here.

Companies	Deposits (%) by Investors			
	A	B	C	D
1	40	30	60	50
2	60	70	40	50

Exercise 4.34

Using the data about incomes and expenses of a bank, determine the relative value of the dynamics, structure, and intensity. Make appropriate conclusions.

Indicators	By 1.01 of the last year	By 1.01 of the current year
Incomes – total including:	2234.08	2340.10
Received interests	1781.12	1826.05
Expenses – total including:	2015.12	2034.24
Paid interests	1674.02	1626.51

Exercises for Interim Evaluation

Exercise 4.35

From the data for a quarter of the year about the number of media clips highlighting the performance of the four largest banks in the country, determine the average share of information that is positive in tone. Explain the choice of the form of the average.

Banks	Number of fragments with positive information	In percentage, %
A	110	50.7
B	37	32.8
C	55	56.7
D	48	55.8

Exercise 4.36

A price of a good increased by 20 euro over a period and demand decreased by 100 pieces (pcs.) during the same period. Determine the relative change in price and demand. Also determine the coefficient of elasticity if the price was 220 euro and the sales volume was 1100 pcs. Make conclusions.

Exercise 4.37

The supply of a good to market over a period increased by 10 thousand pcs. and the price decreased by 0.5 euro during the same period. Determine the relative change in price and supply to market. Determine also the coefficient of elasticity if, at the beginning of the period, the price was 25 euro and the sales volume was 200 thousand pcs. Make conclusions.

Exercise 4.38

Prices of some meat products in the first quarter of the year increased by 15 %; in the second quarter, they increased by 10 % more. Prices decreased by 5 % on average in the third quarter. How did the prices of the meat products change?

Exercise 4.39

A worker received 1500 euro. The amount of compulsory payments was 3.5 % and the income tax rate was 15 %. What amount of compulsory payments and taxes were deducted all together and separately? Make conclusions.

Exercise 4.40

A company planned to increase labor productivity by 7 %. The plan was achieved to 102 %. What was the real increase in labor productivity? Make conclusions.

Exercise 4.41

In the first quarter of the year, prices increased by 8 %; in the second quarter, they increased by 10 % more. Prices decreased by 5 % on average in the third quarter. How did prices change in the fourth quarter if the overall price increase was 20 % over the year?

Exercise 4.42

Prices of commodity A decreased by 5 %; later, they increased by 10 %. Prices of commodity B increased by 20 % and then decreased by 15 %. The price of which commodity changed more intensively? Support your conclusions with calculations.

Exercise 4.43

A selective check of product quality gave the following data.

Type of commodity	Returned for correction, items	Share of goods returned for correction from all checked products, %
Sewing goods	40	1.5
Knitwear	55	5.0
Total	95	x

Determine the average share of the goods that were returned. Explain the use of the form of the average.

Exercise 4.44

The following data were received from a selective check of product quality.

Type of commodity	Checked goods, thousand items	Percentage of spoiled goods
Sewing goods	500	5.5
Knitwear	450	1.0
Total	950	x

Determine the average share of the goods returned for correction.

Explain the use of the form of the average.

Exercise 4.45

Below are some data about the share of services provided for people in the total services of two cities.

City	Share of services provided for population from total services, %	Total services, million euro
A	34.6	2686
B	35.8	114
Total	x	2800

Determine the average share of provided services for people in the general scope of provided services. Explain the use of the form of the average.

Exercise 4.46

Below are data some about stores in the region.

Type of store	Average annual sales volume per salesperson, thousand euro	Number of salespeople, thousand people
State-run	50	1.5
Rental	52	1.2
Private	60	1.8
Total	x	4.5

Determine the average annual sales volume per salesperson for all types of store. Explain the use of the form of the average.

Exercise 4.47

Below are data some on the production plan of some teams in a workshop.

Team №	Real output volume, thousand items	Plan achievement, %
1	500	105.5
2	450	100.0
3	100	98.0
Total	1050	x

Determine the average percentage of the plan's general achievement. Explain the use of the form of the average.

Exercise 4.48

From the tabulated data and using the method of the multivariate mean, build a ranking of countries according to demographic situation. Use the lowest indicator for valuation.

Countries	Demographic indicators			
	Death coefficient, ‰	Child death coefficient, ‰	Senescence coefficient, %	Expected life expectancy, years
A	13.5	12.0	24	73
B	16.5	13.5	27	69
C	14.0	11.5	25	74
D	17.0	13.9	30	68
F	16.0	12.9	28	71

Exercise 4.49

The "credit portfolio" of a commercial bank over two years is presented with the following data, million euro.

Type of support	Last year	Current year
All allotted credits, including:	7.71	10.41
Long-term	0.51	0.55
Medium-term	0.62	0.66
Short-term	6.58	9.20

From the data, analyze the structure and structural changes; the correlation of separate components of the "credit portfolio"; and the dynamics of the indicators. Make conclusions.

Exercise 4.50

From the given data about the results of privatization, determine the average share in monetary terms in the general scope of privatized entities. Explain the use of the form of the average.

Scope of privatized entities, billion euro	Share in monetary terms (the rest – in property certificates), %
3.5	10.2
4.7	12.4
5.1	11.1

Exercise 4.51

Using the data below about the distribution of saving institutions of a region by location (at the end of the year), determine relative values of structure, coordination, and dynamics. Make conclusions.

Indicator	Last year	Current year
All saving institutions of the region, including:	755	709
in cities	496	478
in rural area	259	231

Exercise 4.52

From the data below about the enrollment of students in one of the economic educational institutions of a city, determine the average share of students who finished their course at a specialized school. Explain the use of the form of the average.

Form of education	Total number of all enrolled students who finished magnet school, thousand people	Specific weight of students who finished magnet school, %
Full-time	0.8	60
Evening time	0.7	30
By correspondence	0.6	10

Exercise 4.53

An average monthly wage of a highly-qualified workers in the past year was 1950 euro; that of less-qualified workers was 1150 euro. In the current year, while the overall number of workers decreased by 10 %, the share of less-qualified workers decreased by 15 %. How will the average wage of all workers change in the current year if the average monthly wage for each group remains unchanged?

5. VARIATION INDICATORS AND DISTRIBUTION FORMS

5.1. The concept of distribution rows and their characteristics

As mentioned in the preceding chapter, in the tabulation and grouping of statistical data the distribution rows are the result of grouping. A **distribution row** is an orderly sequence of paired elements – category and frequency. ***Category*** constitutes a distinct form of a grouping feature, while a ***frequency (count)*** refers to the number of elements with a certain value (level) of a feature in a group. A good example of a distribution row is given below on the results of an examination for a group of students. The students' results are distributed in the following manner:

Grade	Number of students
"fail"	4
"satisfactory"	8
"good"	10
"excellent"	3
Total	25

Depending on the feature, the distribution rows can be ***attributive (categorical)***, as in the given example, or ***variational***, e.g., the distribution of a group of workers by level of remuneration.

Variational rows can be both discrete and interval. ***Discrete rows*** are used for primary or discrete features. A ***discrete feature*** is a feature that takes whole values at defined intervals; other values cannot be placed between them (such as the number of children in a family). ***Interval rows*** are used, as a rule, for continuous features, which can take any value according to certain limits and have an approximate expression (e.g., the height of a person).

An interval row can be built for a discrete value as well if it varies within wide limits (e.g., the distribution of all legal entities of a city by number of employees). In addition, variants are grouped into intervals; frequencies do not belong to the distinct values of a feature, as in discrete rows, but belong to a whole interval.

In analyzing distribution rows, it is more convenient to use a ***percentage***, which is expressed as a coefficient or percent rather than a frequency value. A percentage enables the study of a population's structure. Another characteristic of a distribution row is cumulative frequency (count) or cumulative percentage. The cumulative frequency is calculated from data on frequency distribution, while the cumulative percentage is calculated from data on a population's structure using the appropriate formulas,

$$\sum f_1 = f_1$$

$$\sum f_2 = f_1 + f_2$$

$$\sum f_m = f_1 + f_2 + K + f_m$$

$$\sum f_m = f_1 + f_2 + \cdots + f_m$$

Example 5.1

Data are given below on the work experience (years) and wage (euro per hour) of 30 workers. We can build a distribution row by work experience with equal intervals and determine its characteristics.

Table 5.1

№	Work experience	Wage	№	Work experience	Wage
1	1.0	200	16	10.5	276
2	1.0	202	17	1.0	234
3	3.0	205	18	9.0	270
4	6.5	290	19	9.0	264
5	9.2	298	20	6.5	252
6	4.4	250	21	5.0	241
7	6.9	280	22	6.0	256
8	2.5	230	23	10.1	262
9	2.7	223	24	5.5	245
10	16.0	310	25	2.5	240
11	13.2	284	26	5.0	244
12	14.0	320	27	5.3	252
13	11.0	295	28	7.5	253
14	12.0	279	29	7.0	252
15	4.5	222	30	8.0	262

First, let us calculate a maximal number of groups with Sturges' formula: $\boldsymbol{m = 3.322 \times \log n + 1 = 3.322 \times \log 30 + 1 = 5.9}$. This is the maximal number of groups by which a population can be divided. Taking into consideration the difference between the maximal and minimal values of the feature (equal to 15) and the peculiarity of the feature "work experience" (discrete), let us use five groups so that the width of an interval will be a whole number. Let us determine the width of an interval for a grouping value

$$h = \frac{x_{max} - x_{min}}{m} = \frac{16-1}{5} = 3 \text{ years.}$$

Thus, gradually adding the width of an interval to an upper limit of a previous interval, we have the following groups for work experience: 1–4(1+3); 4–7(4+3); 7–10(7+3); 10–13(10+3); and 13–16(13+3). The resulting distribution row will look like Table 5.2.

Table 5.2. Distribution row of workers by work experience

Groups by work experience (x_j), years	Number of workers (f_j)	Percentage (ω_j), %	Cumulative Frequency, ($\sum f_j$)	Cumulative Percentage ($\sum \omega_j$), %
1–4	7	(7/30) ×100%=23	7	23
4–7	10	33	7+10=17	23+33=56
7–10	6	20	17+6=23	56+20=76
10–13	4	13	23+4=27	76+13=89
13–16	3	11	27+3=30	89+11=100
Total	30	100	x	x

Distribution rows can be shown graphically. Discrete rows are shown with a distribution histogram (see Figure 5.1). A graphical representation of an interval row is given in the histogram in Figure 5.2.

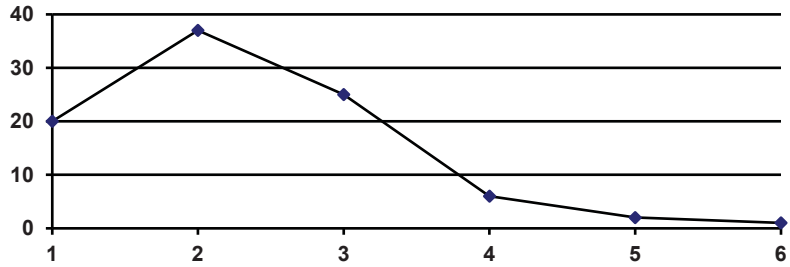


Figure 5.1. A histogram showing the distribution of families by number of family members (using data from a sociological observation).

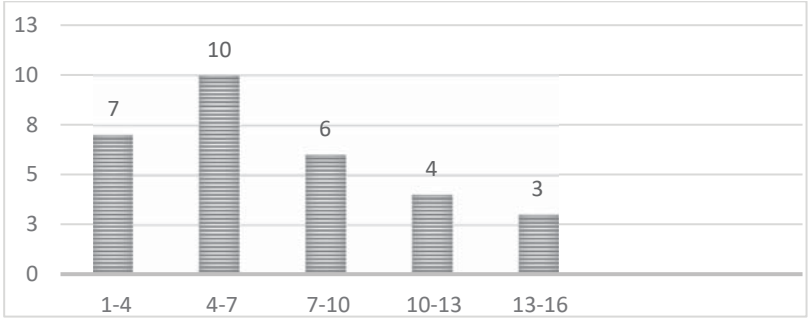


Figure 5.2. A histogram showing the distribution of workers by work experience (from the example data).

As the scope of a population increases and the interval width decreases, the histogram approaches a curve.

A distribution row can be characterized by a system of indicators (statistical estimates), among them we find indicators of the distribution center, indicators of variation, and indicators of the distribution form.

5.2. Characteristics of the distribution center

The characteristics of a distribution center include the mean, mode, and median. That is, in addition to the arithmetic mean, mode, and median, we also use the characteristic of the distribution center. This characteristic is otherwise known as the ordinal mean, which is considered in combination with quartiles, deciles, and percentiles.

Mode (Mo) is the value that is repeated most frequently in a distribution row. The mode can be easily identified visually in a discrete row. Only a modal interval can be found in an interval row. The approximate value of a mode is calculated with the formula

$$M_0 = x_{M_0} + h_{M_0} \frac{(f_{M_0} - f_{M_{0-1}})}{(f_{M_0} - f_{M_{0-1}}) + (f_{M_0} - f_{M_{0+1}})},$$

where x_{M_0} is the lower limit of the modal interval; h_{M_0} is the width of the modal interval; f_{M_0} is the frequency of the modal interval; $f_{M_{0-1}}$ is the frequency of the interval that comes before the modal interval; and $f_{M_{0+1}}$ is the frequency of the interval that comes after the modal interval.

It should be noted that the calculation of a precise value for the mode of a continuous variable in an interval distribution row is not practical, because not all units of a population may have a calculated mode value and, in turn, cannot be interpreted correctly. For an interval row, it is more efficient to use the ordinal mean in addition to a modal interval value.

Median (Me) is the value that divides a ranged row into two numerically equal parts. So, if a distribution row represents workers by age, *Me* = 34 years old, this means that half of them are younger than this age and the other half is older. When a row has an even number of row members, then the median is equal to the average of the two values situated in the middle of the row.

To determine the median in a discrete row, we first calculate the half-sum of frequencies and then, based on the cumulative frequency, we determine which variant accounts for it.

For an interval row, the first median interval is determined using the principle of finding the median in a discrete distribution row. Then, the median is calculated with the formula

$$M_e = x_{Me} + h_{Me} \frac{\frac{\sum f}{2} - S_{Me-1}}{f_{Me}},$$

where x_{Me} is the lower limit of the median interval; h_{Me} is the width of the median interval; $\frac{\sum f}{2}$ is the half-sum of frequencies; S_{Me-1} is the sum of accumulated frequencies before the median interval; and f_{Me} is the frequency of the median interval.

Example 5.2

Below we have some data about the distribution of families by number of children in a city (Table 5.3). Determine the mode and median of the distribution. Give a proper interpretation.

Table 5.3. Distribution of families by number of family members in a city (data from sociological research), in percent (%)

Family size	Percentage, %	Cumulative percentage, %
1	9.4	9.4
2	20.3	29.7
3	36.6	66.3
4	24.7	91.0
5	6.2	97.2
6	2.2	99.4
7 and more	0.6	100.0
Total	100.0	x

In this distribution row, Mo = 3 people and Me = 3 people. This value matches the maximal percentage (36.6%) and the larger half of the cumulative percentage (66.3%), i.e. more than half of the families in a city consist of 1–3 people.

Example 5.3

Determine the mode and median of distribution by the age structure of the adult population of a city (Table 5.4).

Table 5.4. Age structure of adult population in a city (data from sociological observation), in percent (%)

Groups by age	Percentage, %	Cumulative percentage, %
16–25	20.8	20.8
26–35	18.7	39.5
36–45	21.5	61.0
46–55	17.2	78.2
56–65	14.4	92.6
66–75	5.6	98.2
Older than 75	1.8	100.0
Total	100.0	x

In this example, the modal interval is *Mo* (36–45) and

$$M_o = 36 + 9\,\frac{21.5-18.7}{(21.5-18.7)+(21.5-17.2)} = 39.6 \text{ years.}$$

As mentioned earlier, the value of a mode is not interpreted and is presented here only to illustrate the method of calculation for a modal value.

The median interval is *Me* = (36–45), using the formula

$$Me = 36 + 9\,\frac{\frac{100}{2} - 39.5}{21.5} = 40.4 \text{ years.}$$

That is, half of the population is up to 40 years old, while the other half is older than 40. Each of the two parts, by which the median divides a population by a certain attribute, can, in their turn, be divided into ***quartiles (Q)***.

Thus, the first quartile Q_1 separates a quarter of a population; the second Q_2, the median itself, separates half; the third Q_3, accounts for three quarters. A formal representation of the first and third quartiles is given in the following

$$Q_1 = x_{Q_1} + h_{Q_1} \frac{\frac{\sum f}{4} - S_{Q_1-1}}{f_{Q_1}}, Q_3 = x_{Q_3} + h_{Q_3} \frac{\frac{3\sum f}{4} - S_{Q_3-1}}{f_{Q_3}},$$

where x_{Q_1} and x_{Q_3} are the lower limits of the quartile intervals; h_{Q_1} and h_{Q_3} is the width of the quartile interval; S_{Q_1-1} and S_{Q_3-1} are the sum of cumulative frequencies before the quartile interval; and f_{Q_1} and f_{Q_3} are the frequencies of the quartile interval.

In Example 4.3, the first quartile lies in the interval of 26 to 35 years old and the third quartile – 46 to 55 years old. We can use the formula

$$Q_1 = 26 + 9 \frac{\frac{100}{4} - 20.8}{18.7} = 28 \text{ years}, Q_3 = 46 + 9 \frac{\frac{3 \times 100}{4} - 61.0}{17.2} = 53 \text{ years}.$$

Thus, 25 % of the population of a city is younger than 28 and older than 53. Half of the population in a city are people in the age range of 28–53 years old.

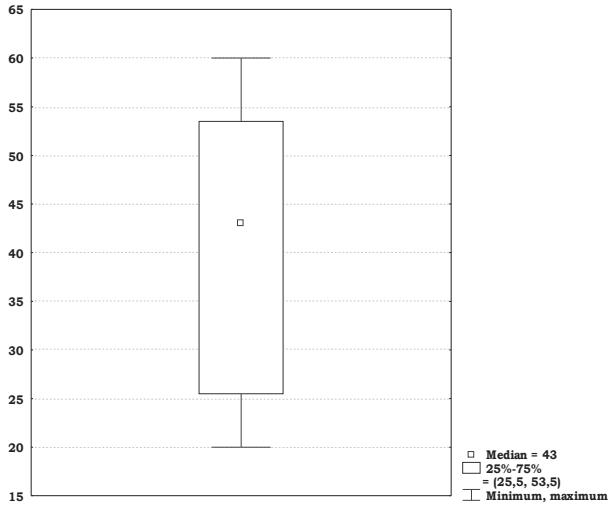


Figure 5.3. Distribution of workers by age

Median and quartiles may be graphically represented with the help of a "box and whisker plot" (Figure 5.3).

Also, the ***deciles and percentiles*** are calculated. The *k*-percentile is a number, the smaller of which % units of a population receive the value *k.*

So, the 25th percentile is the first quartile and the 10th percentile is the first decile. Sometimes Q_1 and Q_3 are called the lower and upper quartiles, respectively.

The extent of variant scatter can be characterized by the value ($Me - Q_1$) or ($Q_3 - Me$), or even better – using their average value – average quartile deviation, which is calculated with the formula

$$Q = (Q_3 - Q_1)/2.$$

We state that half of all the variants lie in the interval ($Me \pm Q$). The mode and median do not depend on the values of all the variants of a population unit and so they do not use the average as a generalized value, but rather supplement it. This has some advantage over the arithmetic mean in some cases. The values of all three characteristics coincide only in the case of symmetry in a distribution row.

If,

- $x > Me > Mo$: the distribution has left-sided asymmetry (Figure 5.4);

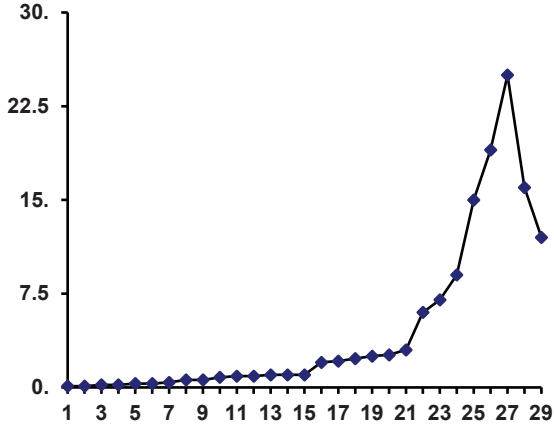


Figure 5.4. Left-sided asymmetry

- $x = Mo = Me$: symmetric distribution (Figure 5.5);

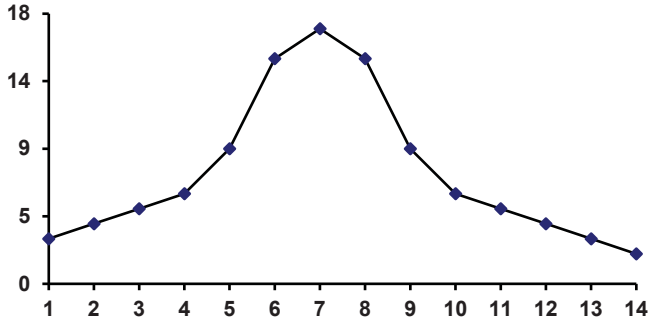


Figure 5.5. Symmetric distribution

- $x < Mo < Me$: the distribution has right-sided asymmetry (Figure 5.6).

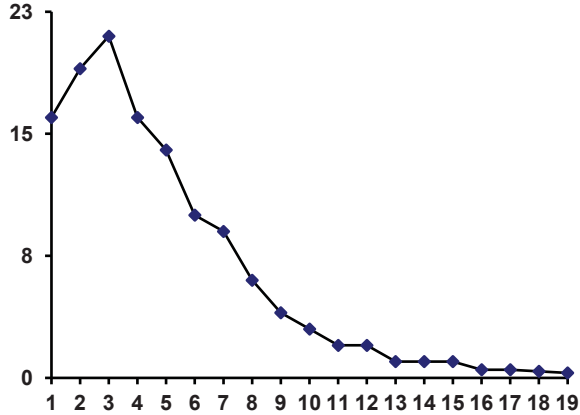


Figure 5.6. Right-sided asymmetry

The characteristics of the distribution center, generalizing the individual, give a general characterization, but they do not show the degree and regularities of the deviation of the individual from the general, i.e. the degree of variation and form of distribution.

5.3. Variation indicators

A feature variation is an attribute of a statistical population that is caused by the effect of numerous interconnected reasons, divided by basic (first cause) and minor (secondary). Basic ones form a distribution center; secondary ones form a variation of values, their cumulative effect, and distribution form. The smaller the variation, the more reliable and typical the attributes of the center are and the first of these is the average.

To characterize a variation, the following system of estimates is used.

Range of a variation is the difference between the largest and the smallest values of feature $R = x_{max} - x_{min}$.

In the interval distribution, row R is determined as the difference between the upper and lower limits of the last interval and a lower limit of the first one, or as the difference between the average values of these intervals.

A measure of the variation R cannot always be a reliable estimate as it depends on two extreme values, which are rarely typical for a population or only have an occasional nature. These are "outliers". In practical statistical research, extreme values have to be processed or carefully considered. Usually these are errors of coding or registration, but sometimes they are of an accidental nature. Quite frequently they are eliminated and in this way the range of a variation is narrowed, leading to the higher uniformity of a population. Additionally, ***a quartile range*** reduces the effect of occasional reasons. It is calculated as $R_Q = Q_3 - Q_1$.

When eliminating extreme values, it is worth remembering that sometimes something interesting or relevant phenomena may be connected to them. In the literature, in addition to the simple elimination of "outliers", some calculation procedures of distribution estimates are suggested; those ones that are not sensitive to the data structure are known as ***robust***. The distribution estimates that result from the use of such methods are called robust estimates. Statistical programs often propose the calculation of estimates using the methods of Hampel, Andrews, and Tukey. For example, Tukey proposed a method for robust estimation, in particular **winsorized** estimates, the essence of which lies in the fact that extreme values are not eliminated, but are replaced. If we have an orderly row of values, $X_1, X_2, \ldots X_n$, then X_1 is given the value X_2 and X_n is given the value X_{n-1}. If this operation does not give the expected results, i.e. the population does not become uniform, the procedure is repeated (e.g., up to five times). Under double winsorization, X_1 and X_2 are given the value X_3, and the last two in

a row are given the value X_{n-2}. The estimates calculated for such "changed" populations (average, mode, median, etc.) are called ***winsorized estimates***.

It is important to recognize that any statistical analysis where the work relies on the success of the whole analysis depends on diligence in preparing the material. As to "cleaning" or preliminary processing procedures applied to the data, there is an ethical side in addition to a professional one: a researcher has to strive for an objective and scientifically grounded result, even if it is not the expected one.

Another form of variation is ***average deviation***, which can be calculated in two ways:

1. **Average linear deviation**

a) unweighted

$$\bar{l} = \frac{\sum|x_i - \bar{x}|}{n},$$

b) weighted

$$\bar{l} = \frac{\sum|x_j - \bar{x}|f_j}{\sum f_j}.$$

2. **Average quadratic deviation**

a) unweighted

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

b) weighted

$$\sigma = \sqrt{\frac{\sum(x_j - \bar{x})^2 f_j}{\sum f_j}}.$$

The variation characteristic σ^2 is called ***dispersion***. It characterizes the extent of scatter of feature values around the mean. Dispersion has no units of measurement and is not interpreted:

a) unweighted

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

b) weighted

$$\sigma^2 = \sqrt{\frac{\sum(x_j - \bar{x})^2 f_j}{\sum f_j}}.$$

In practice, a simplified formula of the dispersion calculation is used for non-grouped data as the difference between the square of the mean and the mean in the square

$$\sigma^2 = \overline{x^2} - \bar{x}^2 .$$

The smaller the average deviation, the more typical the mean is, and thus the more uniform a population is. An average quadratic deviation is always bigger than a linear one. In symmetric and moderate asymmetric distributions

$\sigma = 1.25 \times \bar{l}$. Attributes $\bar{l}$, σ, and R are named values that have measurement units of a feature.

To compare the degree of variation of one feature in various populations, we use the average quadratic ***coefficient of variation***

$V_\sigma = \frac{\sigma}{\bar{x}} \times 100\%$, $V_{\bar{l}} = \frac{\bar{l}}{\bar{x}} \times 100\%$, or the linear variation coefficient.

A population's uniformity can be estimated with the ***coefficient of variation***. A population is considered to be uniform when $V_s < 33\%$. From the data in example 5.1, let us analyze the variation level of work experience. We can construct a table for this purpose.

Table 5.5. Computed table of variation indicators

Groups by work experience (x_j), years	**Number of workers (f_j)**	x_j'	$x_j' f_j$	$x_j' - \bar{x}$	$(x_j' - \bar{x})^2$	$(x_j' - \bar{x})^2 f_j$	$\lvert x_j - \bar{x} \rvert f_j$
1	2	3	4	5	6	7	8
1–4	7	2.5	17.5	-4.6	21.16	148.12	32.2
4–7	10	5.5	55.0	-1.6	2.56	25.60	16
7–10	6	8.5	51.0	1.4	1.96	11.76	8.4
10–13	4	11.5	46.0	4.4	19.36	77.44	17.6
13–16	3	14.5	42.5	7.4	54.76	164.28	22.2
Total	30	x	213.0	x	x	427.20	96.4

Average work experience is calculated with the formula of the arithmetic mean weighted for grouped data. It is necessary to find the middle of the interval x_j' (Table 5.5).

Then, we find $\bar{x} = \frac{\sum x_j' \times f_j}{\sum f_j} = \frac{2{,}5\times7+5{,}5\times10+8{,}5\times6+11{,}5\times4+14{,}5\times3}{7+10+6+4+3} = 7.1$ years.

Modal interval: (4–7).

Median interval: (4–7).

The median is equal to $Me = 4 + 3\frac{15-7}{10} = 6.4$ years.

To calculate the average quadratic deviation, we calculate columns (5–7), then, $\sigma = \sqrt{\frac{427.2}{30}} = \sqrt{14.27} = 3.77$ years and dispersion is equal to 14.27.

Average linear deviation is (Table 5.5, column 8) $\bar{l} = \frac{96.4}{30} = 3.2$ years.

The variation coefficient, based on the average quadratic deviation, is $V_\sigma = \frac{3.77}{7.1} \times 100 = 53.1\%$, which is more than 33 %; based on the average linear deviation, we have $V_{\bar{l}} = \frac{3.2}{7.1} \times 100 = 45.1\%$, which is a more extreme value of 26.4 %.

Thus, the population of workers is not uniform in terms of work experience and 7 years of work experience is not typical for the whole population of workers. As such, it needs to be regrouped into two uniform populations.

Let us consider the calculation peculiarities of the form of the mean and dispersion for an alternative feature. An **alternative feature** is one that takes two opposite values, e.g. "gender": "male" and "female". However, any attributive feature can be presented in the form of an alternative one. For example, "level of education/educational attainment": "availability of university education" and "lack of university education". To calculate the attributes of an alternative feature, let us introduce a system of denotation. The availability of a modal value is denoted with 1, while its lack is denoted with 0. The share of units that have a modal value is denoted with p, and those that do not have it are denoted with q. We state that a sum of probabilities p and q is equal to one. Then, the average is calculated as the weighted arithmetic mean

$$\overline{x} = \frac{1 \times p + 0 \times q}{p + q} = p \,,$$

Dispersion is the product of two probabilities

$$\sigma_p^2 = \frac{(1-p)^2 \times p + (0-p)^2 \times q}{p+q} = \frac{q^2 \times p + p^2 \times q}{p+q} = \frac{p \times q \times (p+q)}{p+q} = p \times q$$

or $\sigma_p^2 = p \times (1-p)$.

Then, the average quadratic deviation is calculated with the formula

$$\sigma_p = \sqrt{p \times q}\,.$$

Example 5.3

For example, in the quality control data for 1000 finished products, 15 spoiled goods were recorded. Let us build a distribution row denoting quality goods with 1 and spoiled goods with 0 (Table 5.6).

Table 5.6. Distribution row of goods by quality

Feature value	Number of produce units	Specific weight
1	1000-15=985	p =985/1000=0.985
0	15	q =1-0.985=0.015
Total	1000	$p + q$ =1.000

Then, $\bar{x} = \frac{1 \times 985 + 0 \times 15}{1000} = 0.985 = p$. That is, the share of appropriate quality produce is 98.5 %. The dispersion of a share is $\sigma_p^2 = 0.985 \times 0.015 = 0.014775$ and the average quadratic deviation is equal to $\sigma_p = \sqrt{\sigma_p^2} = \sqrt{0.014775} = 0.1216$ (percentage points).

Obviously, if there is no variation $\sigma^2 = 0$, the maximal value of dispersion is 0.25 when $p = q = 0.5$. If an attributive feature gets more than two values, the estimate of its variation is equal to the product of the shares $\sigma^2 = \omega_1 \times \omega_2 \times \ldots \times \omega_n$.

5.4. Calculation methodology of the characteristics of distribution form

The analysis of a statistical population can be made more complete to express regularities of the correlation of variants and frequencies in a certain function. This is called ***a theoretical curve***. This is a general model of a real

event. If a curve is built using statistical observation data, this is known as an ***empirical curve***.

Distribution curves are divided into ***symmetric*** and ***asymmetric*** according to their form. If, in an asymmetric distribution, a peak is shifted to the left, then it is right-sided asymmetry ("right tail") and vice versa. There are one, two, and multi-apical curves. A multi-peaked curve proves the non-uniformity of a population. Depending on the form of the peak, curves can be peaky or flat-topped. The degrees of asymmetry and "peakiness" are measured with ***coefficients of asymmetry and excess (kurtosis)***, which are denoted as A and E, respectively. In a normal distribution, $E = 3$; in a peaky one, $E > 3$; and with a flat-topped peak, $E < 3$ (figures 5.7–5.8). As such, one may discuss these curves in terms of larger or smaller peaks, and flat-topped peaks.

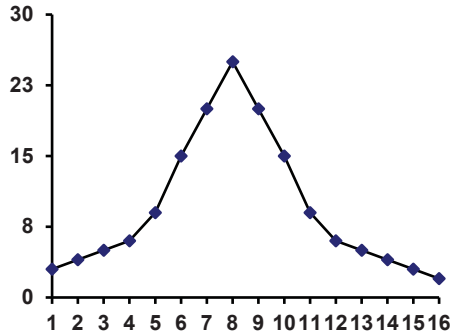


Figure 5.7. Peaky distribution

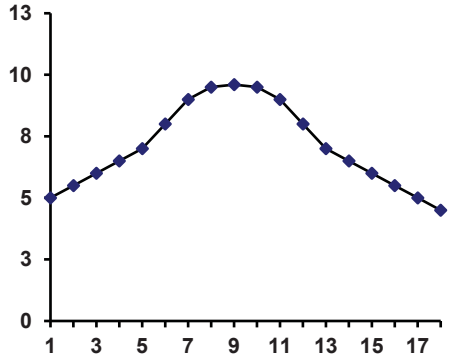


Figure 5.8. Flat-topped peak distribution

The asymmetry coefficient can be calculated with a simplified formula

$$A = \frac{\bar{x} - Mo}{\sigma} \text{ or } A = \frac{\bar{x} - Me}{\sigma}.$$

Value A can be positive or negative. In a symmetric distribution, $A = 0$; in right-sided asymmetry, $A > 0$; and in left-sided asymmetry, $A < 0$. The greater the deviation from 0, in either direction, the larger or smaller the deviation is.

The characteristic of distribution form is based on its distribution moments. A ***distribution moment*** is the average of k-extent of deviations $\bar{x}$, a. Depending on size, a moments are classified into primary ($a = 0$), central ($a = \bar{x}$), and nominal (a = constant). The degree of k characterizes the order of a moment. In the work of Ukrainian and former Soviet authors, we find the assumption that

$$A = \frac{\mu_3}{\sigma^3}, \text{ where } \mu_3 = \frac{\sum (x_j - \bar{x})^3 f_j}{\sum f_j}, \text{ with the kurtosis coefficient,}$$

$$E = \frac{\mu_4}{\sigma^4}, \text{ where } \mu_4 = \frac{\sum (x_j - \bar{x})^4 f_j}{\sum f_j}.$$

Therefore, for the calculation of coefficients of asymmetry and kurtosis (excess), first we need to calculate the moments of the third and fourth orders. As an example, let us study distribution form 5.1 in terms of its asymmetry and kurtosis (excess) coefficients (see Table 5.7).

Table 5.7. Computed table of indicators of distribution form

Groups by work experience (x_j), years	Number of workers (f_j)	x'_j	$x'_j - \bar{x}$	$(x'_j - \bar{x})^3 f_j$	$(x'_j - \bar{x})^4 f_j$
1	2	3	4	5	6
1–4	7	2.5	-4.6	-681.4	3134.2
4–7	10	5.5	-1.6	-41.0	65.5
7–10	6	8.5	1.4	16.5	23.0
10–13	4	11.5	4.4	340.7	1499.2
13–16	3	14.,5	7.4	1215.7	8996.0
Total	30	x	x	850.6	13718.0

First, we calculate the asymmetry coefficient using a simplified formula

$$A = \frac{\bar{x} - Mo(Me)}{\sigma} = \frac{7.1 - 5.3}{3.77} = 0.48.$$

Preliminary calculations confirm that the distribution shows a small right-sided asymmetry.

To calculate the precise values of the asymmetry and kurtosis coefficients, we construct columns 5 and 6 in Table 5.7. The third order moment is equal to

- $\mu_3 = \frac{850,6}{30} = 28.35,$

the asymmetry coefficient is $A = \frac{28,35}{3,77^3} = 0.53$, and the fourth order moment is given by

- $\mu_4 = \frac{13718}{30} = 457.3.$

The excess is equal to $E = \frac{457,3}{14.27^2} = 2.25.$

Thus, the distribution has a flat-topped peak, E < 3, showing the absence of a particular concentration of workers with work experience deviating from normal around the mode and distribution.

5.5. The concept of the normal distribution

Among distribution curves, a special place is taken by the ***normal curve***, which shows a normal or ***Gaussian distribution***. Estimates of the relevance of indicators of asymmetry and kurtosis, as illustrated in the previous example, allow us to make conclusions about the possibility of including the empirical distribution of this type of normal curve at a preliminary stage of research. This is due to the impact of an unlimited number of mutually independent factors that quite often occur in nature. A great number of statistical methods are based on the concept of the normal distribution.

A continuous random value in a normal distribution has the distribution density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\bar{x})^2}{2\sigma^2}}.$$

In a typical, normal distribution curve:

1. The curve is symmetrical relative to the maximal ordinate. The maximal ordinate corresponds to the value of $\bar{x} = M_e = M_o$ and its value is equal to $\frac{1}{\sqrt{2\pi\sigma}}$.
2. The curve asymmetrically approaches the X-axis to infinity.
3. The curve has two crossing-points lying at a distance of ±s from the average value.
4. The curve has two crossing-points at a distance $\pm\sigma$ from the average value.
5. Under the constant value of the average and the increase of σ, the curve becomes more flat. When the average value changes and the σ value is constant, the curve does not change its form and only shifts to the right or left on the X-axis.
6. 68.3 % of all values lie in the interval $\bar{x} \pm \sigma$ (*t=1*). The interval $\bar{x} \pm 2\sigma$ (*t = 2*) covers 95.4 % of all values. 99.7 % of all feature values are located in the interval $\bar{x} \pm 3\sigma$ at (*t = 3*).

A normal distribution only occurs when a great number of random reasons influence a feature value. The influence of these values is independent and none of them have an advantage over the others.

To check relevance, i.e. "not randomness", of the difference between an empirical distribution and normal, consent criteria are used.

One of the most widespread consent criteria is *chi*-square χ^2, as proposed by Pearson

$$\chi^2 = \sum \frac{(f_j - f'_j)^2}{f'_j}$$

where f_j and f'_j are, respectively, frequencies of empirical and theoretical distribution in a given interval.

Theoretical frequencies are calculated with the formula

$$f'_j = p_j \times n \,,$$

where p_j is the probability of a random value in a given interval. This is calculated with the formula

$$p_j = \Phi(t_L) - \Phi(t_U),$$

where $\Phi(t_L)$; $\Phi(t_U)$ are Laplace integral function values in the lower and upper limits of corresponding intervals, which are determined by a table of normal distribution functions (appendix, Table 10). The larger the difference between the empirical and theoretical frequencies, the larger the value of Pearson χ^2. To answer a question about a normal distribution, a real, calculated criterion value is compared with a table (critical) value for a definite number of degrees of freedom and relevance (at the level 0.05 or 0.01). If $\chi^2 > \chi^2_{table}$, i.e., χ^2 is in a critical area, then the difference between the empirical and theoretical frequencies is significant and, in turn, this confirms that it cannot be explained by a random fluctuation of the sample data. In this case, the zero hypothesis on the normality of the distribution is eliminated. If $\chi^2 < \chi^2_{table}$, i.e. χ^2, the calculated difference does not exceed the maximal possible value of differences between the empirical and theoretical frequencies. This can occur as a result of random fluctuations in

the sample data. To calculate the number of degrees of freedom, the difference between the number of groups (*k*) and the number of limits (*l*) is used.

For calculating Pearson's criterion, it is necessary to meet the following requirements:

1) there should be at least 50 observations;
2) there should be at least 5 frequencies in each interval.

Using the value of χ^2, V.I. Romanovskyi suggested the evaluation of the correspondence of an empirical distribution to the normal curve with the formula

$$K_R = \frac{\chi^2 - (k-3)}{\sqrt{2(k-3)}},$$

where *k* is the number of groups and value *(k – 3)* is the number of degrees of freedom. If the correlation is more than 3, the differences between the frequencies of the empirical and normal distributions are not random and the empirical distribution cannot be considered to be close to normal. Otherwise, the differences are random in nature and the distribution can be classified as normal.

Another criterion used to check correspondence to the normal distribution is the Kolmogorov-Smirnov criterion. This is based on the calculation of the maximal difference (*d*) between relative frequencies in empirical and theoretical distributions in terms of absolute values, and then finding the number for $\frac{\lambda}{\sqrt{n}}$ and comparing it to the calculated one at a given level of relevance $K(\lambda) = 0.09505$.

If $d > \frac{\lambda}{\sqrt{n}}$, then the difference between empirical and theoretical frequencies is significant and cannot be explained by random fluctuations in the sample data. In this case, the zero hypothesis on the normality of the distribution can be eliminated.

If $d < \frac{\lambda}{\sqrt{n}}$, then the calculated difference does not exceed the maximal possible value of differences between the empirical and theoretical

frequencies, which can occur as a result of random fluctuations in the sample data. Let us consider the estimation methods of correspondence to a normal distribution using example data characterizing the distribution of workers by work experience (Table 5.8). It is worth remembering that the average is equal to 7.1 and the average quadratic deviation is 3.77.

Table 5.8. Computed table of frequencies of normal distribution

Work experience (x_j), years	f_j	$\frac{x_L - \bar{x}}{\sigma}$	$\frac{x_U - \bar{x}}{\sigma}$	(Φt_L)	(Φt_U)	$(\Phi t_L) - (\Phi t_U)$	f'_j
1	2	3	4	5	6	7	8
1–4	7	-1.6180	-0.8223	-0.4474	-0.2939	0.1535	4.605
4–7	10	-0.8223	-0.0265	-0.2939	-0.0080	0.2859	8.577
7–10	6	-0.0265	0.7692	-0.0080	0.2764	0.2844	8.532
10–13	4	0.7692	1.5650	0.2764	0.4406	0.1642	4.926
13–16	3	1.5650	2.3607	0.4406	0.4909	0.0503	1.509
Total	30	x	x	x	x	x	28.149

Thus, having rounded off the obtained values, we can determine the theoretical frequencies and calculate the criterion $\chi^2 = 2.61$ (Table 5.9). Using tables of critical values, we can determine a theoretical value for the criterion for degrees of freedom (5 – 3 = 2) and level of relevance 0.05, which is equal to 5.99. As $\chi^2 < \chi^2_{table}$, then differences between the theoretical and actual frequencies are random and the distribution can be classified as normal.

Using the criterion of V.I. Romanovskyi, we get $K_R = \frac{\chi^2 - (k-3)}{\sqrt{2(k-3)}} = \frac{2{,}61-2}{\sqrt{2\times 2}} = 0.305$, which also confirms to our previous conclusion.

We can use the Kolmogorov-Smirnov criterion to check the distribution for normality. We calculate the cumulative percentages of the empirical and theoretical distributions and determine the maximal absolute difference between them (Table 5.9).

Table 5.9. Computed table of frequency correspondence to normal distribution

Work experience (x_j), years	f_j	f'_j	$\frac{(f_j - f'_j)^2}{f'_j}$	$\sum f_j$	$\sum f'_j$	$\sum \omega_j$	$\sum \omega'_j$	$\lvert\omega_j - \omega'_j\rvert$
1–4	7	5	0.80	7	5	0.233	0.167	0.066
4–7	10	9	0.11	17	14	0.567	0.467	0.100
7–10	6	9	1.00	23	23	0.767	0.767	0.000
10–13	4	5	0.20	27	28	0.900	0.933	0.033
13–16	3	2	0.50	30	30	1.000	1.000	0.000
Total	30	30	2.61	x	x	x	x	x

From the results of the calculation of f, we can see that the maximal deviation is *d = 0.1*, which corresponds to the second interval. Using the tables of critical values (appendix, Table 8) to give $\lambda = 1.36$, then *n* = 30 $\frac{\lambda}{\sqrt{n}} = \frac{1,36}{\sqrt{30}} = 0.248$. This is larger than an actual value of 0.1, which allows us to consider this distribution to be closer to normal. It is possible to come to a logical conclusion using the maximal difference of the absolute cumulative frequencies and then $D = \lambda\sqrt{n}$. The maximal difference of absolute cumulative frequencies is equal to 3 = 17–14, at a probability level of 0.9505 and $\lambda = 1.36$.

$D = 1.36\sqrt{30} = 7.45$ and this is also larger than the maximal calculated difference.

5.6. Statistical study of concentrations

Research into changing tendencies in the spatial structure of economic events has significant influence on the development of local and regional economies, and the national economy more generally. These tendencies are used to predict economic development and such processes of economic concentration have a serious impact on the structure of economic processes. **Concentration** here refers to the centralization of an event in separate groups. We may consider such processes of concentration to cover elements like incomes, resources, and investments etc. at the regional or local level.

The methodological approach to the estimation of the degree of concentration relies on the Herfindahl-Hirschman index of concentration. The main purpose of this approach is to analyze the extent of production to estimate the prevalent tendencies of concentration and their outcomes. The index of concentration envisages the calculation of sums of fraction squares of each region and is calculated by

$$H = \sum_{j=1}^{n} d_j^2,$$

where d_j is the fraction of j region and n is the number of regions. To simplify perception and interpretation, one can use the quadratic mean. The quadratic mean is used to characterize the direct connection between a coefficient and the degree of concentration. For a convenient interpretation, it can be estimated as a percentage

$$K_{conc} = \sqrt{\frac{\sum_{j=1}^{n} d_j^2}{n}} \times 100\%.$$

Table 5.10. Variation limits/boundaries of quadratic coefficients of concentration depending on the number of structural elements

Number of elements	Variation limits, %		Variation range *R*, percentage points	K_{min}/R	$\Delta = \left\lvert \frac{K_{min}}{R} - 1 \right\rvert$
	Minimum, K_{min}	Maximum, K_{max}			
2	50.0	70.7	+20.7	2.42	1.42
3	33.3	57.7	+24.4	1.36	0.36
4	25.0	50.0	+25.0	1.00	0.00
5	20.0	44.7	+24.7	0.81	0.19
6	16.7	40.8	+24.1	0.69	0.31
7	14.3	37.8	+23.5	0.61	0.39
13*	7.7	27.7	+20.0	0.39	0.61
27**	3.7	19.2	+15.5	0.24	0.76

*) analysis by kind of activity;
**) regional analysis.

As we can see in Table 5.10, the more diffuse the population, the more difficult its analysis – the concentration coefficient will be less sensitive making the interpretation more difficult. In our opinion, best results are achieved when there are 4-5 structural elements: in this case $R = max$ and deviation $\Delta = 0$.

To avoid the diversification of elements in a concentration coefficient and have it in a comparable form for the identification of the statistical relevance of concentration, it is expedient to compute a concentration coefficient.

Computation can be done in different ways, but its essence lies in one thing: the higher the value of the coefficient, the more relevant the degree of concentration.

1. Computation based on comparison with a minimal value $K_k^I = \left(\frac{K_k}{K_{\min}} - 1 \right) \times 100\%$, where K_{mir} is the minimal coefficient value with different numbers of structural elements.
2. Computation based on comparison with a value range $K_k^{II} = \frac{K_k - K_{\min}}{R} \times 100\%$.
3. Computation based on comparison with a maximal value $K_k^{III} = \frac{K_k}{K_{\max}} \times 100\%$.

In the distribution of rows (Table 5.10), the coefficients of localization and concentration are used to analyze non-uniformities in the distribution of feature values among separate components of a population. These are effective tools to estimate the level of differentiation of a population from the data in distribution rows with uneven intervals, as well as for distribution rows by feature quality.

The estimate of a distribution of non-uniformity is based on the comparison of structures of two distributions according to the number of population units (d_j) and the size of the feature values (D_j).

The **coefficient of localization** characterizes in which groups an event is concentrated and how significant the level of localization is. A localization coefficient is calculated for each component of a population with the formula

$$L_j = \frac{D_j}{d_j}.$$

Under conditions of uniform distribution, then $L_j = 1$. If there is a certain concentration of feature values in a group, then $L_j > 1$; if there is no concentration in a group, then $L_j < 1$.

A concentration coefficient makes it possible to estimate the distribution of non-uniformity in general and is calculated with the formula

$$K = \frac{\sum_{j=1}^{m} \left| D_j - d_j \right|}{2}.$$

A concentration coefficient value ranges from 0 to 100 %. In a uniform distribution, it is equal to null. In a non-uniform distribution, the value is greater than 0. Thus, the higher the concentration level, the higher the coefficient value.

Example 5.4

From the given data, determine the localization coefficient and concentration coefficient of electrical energy production (Table 5.11). Come to some conclusions. From the computed data in Table 5.9, the concentration coefficient is

$$K = \frac{\sum_{j=1}^{m} \left| D_j - d_j \right|}{2} = \frac{90}{2} = 45 \text{ percentage points.}$$

This confirms a sufficiently high concentration of electrical energy production at large power plants with capacities of 1000 mW and higher (the localization coefficient for these electrical power plants is 3 and 5, respectively).

Table 5.11. Distribution of power plants by capacity and electrical power production

Capacity of power plant, mW	Specific weight of power plants, % (d_j)	Electrical power production, in % to total (D_j)	$L_j = \frac{D_j}{d_j}$	$\lvert D_j - d_j \rvert$
Less than 50	20	2	2/20=0.10	|2 – 20|=18
50–100	13	3	3/13=0.23	|3 – 13|=10
100–200	37	20	20/37=0.54	|20 – 37|=17
200–400	11	15	15/11=1.36	|15 – 11|=4
400–1000	12	33	33/12=2.75	|33 – 12|=21
1000–3000	4	12	12/4=3.00	|12 – 4|=8
3000 and more	3	15	15/3=5.00	|15 – 3|=12
Total	100	100	x	90

Practice Exercises

Exercise 5.1

Using the given distribution data, make a distribution row, calculate the variation coefficient, and make a conclusion about the uniformity of the population of workers of a company in terms of their level of remuneration. Present your calculation in a table.

Wage, euro	1250	2190	1280	1880	1295	1600	Total
Number of workers	10	25	30	20	15	10	110

Exercise 5.2

Below are some data on the distribution of doctors of science by age.

Age in years	Number
younger than 40	2000
40–50	1700
50–60	4100
older than 60	3400
Total	11200

Make a conclusion about the typical nature of the average age of a doctor of science based on the coefficient of variation. Check the normality of the distribution to a relevance level of 0.05 using the *chi*-square criterion.

Exercise 5.3

The distribution of some families by level of average income per capita is characterized by the following data.

Average income per capita, euro	Family structure, %
less than 100	10
100–200	65
200–250	15
over 250	10
Total	100

With a population of 1200 families in total, determine the size of each group, and the mode and median of the distribution row. Check the normality of the distribution to a 0.05 level of relevance using the Kolmogorov-Smirnov criterion.

Exercise 5.4

From the data about age structure below, calculate the variation indicators and make some conclusions.

Age in years	Structure, %
16–18	10
18–24	27
24–30	18
older than 30	45
Total	100

Exercise 5.5

From the data on the age structure of some workers, determine the distribution form and make some conclusions.

Age in years	Structure of workers, %
younger than 30	15
30–50	25
older than 50	60
Total	100

Exercise 5.6

From the data about the education structure of some workers, make some conclusions about the distribution in terms of the mode and the median.

Education	Structure of workers, %
Secondary	15
Technical secondary	25
Higher	60
Total	100

Exercise 5.7

Below are some data on the wages (euro) of workers in two teams. Make a conclusion about the level of variation in each team. Can we consider an average wage to be typical for the group?

Team № 1	Team № 2
1500	1500
1600	2000
1600	1900
1700	2500
1800	3000
1500	1500
1600	2000
1700	3000

Exercise 5.8

From the data about the wage (euro) of workers of two teams, make conclusions about a distribution form in each team.

Team № 1	Team № 2
1450	1450
1460	1500
1460	1490
1470	1550
1480	1600
1450	1450
1460	1500
1470	1600

Exercise 5.9

From the data, calculate the education dispersion of a group of company employees.

Education	Share of workers, %
Secondary	45
Specialized secondary	15
Higher	40
Total	100

Exercise 5.10

Present the form of a distribution row by "remuneration".

Worker's №	1	2	3	4	5	6
remuneration, euro	750	900	780	820	930	860

Exercise 5.11

From the data in Exercise 3.3 (Chapter 3), group workers by characteristic "level of housing supply" (first, calculate the correlation between the size of housing area and family size) and build an interval distribution row for three or four groups. Arrange your results in the form of a table. Make a figure to present the distribution. Calculate the following features for a distribution row:

(1) indicators of distribution center: average, mode, and median;
(2) distribution quartile, construct a box and whisker plot;
(3) variation indicators: variation range, average quadratic deviation, dispersion, and variation coefficient;
(4) coefficients of asymmetry and excess (kurtosis).

Make some conclusions.

Exercise 5.12

The deposited funds of clients (million euro) in thirty banks of the region are found below.

33.7	17.8	24.,8	27.6	23.8	16.4
11.5	10.6	15.6	19.8	10.4	3.7
14.6	16.6	20.3	20.4	26.9	19.8
23.2	17.3	3.9	9.5	23.6	17.2
5.8	9.8	7.9	29.9	14.7	30.1

Build a variation row with three groups and equal intervals. Specify the distribution elements.

Determine the average, mode, median, variation range, average linear deviation, and linear variation coefficient. Explain the contents of the calculated variation characteristics/attributes. Calculate an indicator of distribution form. Make some conclusions.

Exercise 5.13

At the beginning of the current year, the profitability of the joint stock of 30 commercial banks in the region (10 % mechanical sample) was as follows (%).

16.2	19.9	33.3	24.2	6.3	18.2
12.6	18.3	11.6	16.7	11.9	10.8
6.9	12.9	19.7	16.2	19.9	7.9
11.5	27.9	12.9	18.6	16.1	12.9
22.9	9.8	22.8	15.6	10.3	17.7

Build an interval distribution row with equal intervals. Specify the elements of the distribution row. Determine the average, mode, median, variation range, average linear deviation, and linear variation coefficient. Explain the contents of the calculated variation characteristics/attributes. Learn a distribution type. Make some conclusions.

Exercise 5.14

From the data about the distribution of some households by level of average income per capita (percentage of the total number of households), determine: the quartile coefficient of differentiation of people's expenses; the correlation of expenses made by 20 % of the most well-off and 20 % of the least well-off; and the share of people with average costs per capita that are lower than the cost of living (in 2019 this was 423 euro; in 2020 this was 472 euro).

Level of average income per capita	2019			2020		
	Total	Cities	Rural area	Total	Cities	Rural area
less than 180.0	13.4	7.3	25.6	8.2	3.5	17.9
180.1–240.0	12.7	10.6	17.1	10.1	7.2	16.0
240.1–300.0	13.7	12.9	15.3	11.1	9.2	15.1
300.1–360.0	12.7	13.3	11.5	12.2	12.0	12.5
360.1–420.0	10.9	12.0	8.8	10.4	10.9	9.4
420.1–480.0	8.4	9.5	6.1	8.6	9.3	7.0
480.1–540.0	6.4	7.5	4.3	8.7	10.0	5.7
540.1–600.0	4.8	5.7	3.0	5.9	7.0	3.8
600.1–660.0	3.6	4.5	2.0	5.2	6.3	3.2
660.1–720.0	2.8	3.3	1.6	3.8	4.5	2.2
more than 720.0	10.6	13.4	4.7	15.8	20.1	7.2
Total	100.0	100.0	100.0	100.0	100.0	100.0

Exercise 5.15

Using the data in Exercise 4.57 (Chapter 4), determine the median and variation indicators for each distribution row: average quadratic deviation; dispersion; variation coefficient; and coefficients of asymmetry and kurtosis. Make some conclusions.

Exercise 5.16

Below are some data on the distribution of food stores in a city by scale of turnover.

Turnover, thousand euro	Structure of stores, %
less than1000	10
1000–3000	20
3000–5000	50
more than 5000	20
Total	100

Determine: 1) the modal and median interval as well as the median of the distribution row; 2) variation indicators: dispersion and average quadratic deviation; and 3) distribution form in terms of coefficients of asymmetry and kurtosis. Make some conclusions.

Exercise 5.17

Below are some data about the distribution of manufacturing plants by level of output.

Output, %	Number of workers, people
less than 100	20
100–110	50
110–150	20
more than 150	10
Total	100

Determine: (1) the modal and median interval, as well as the median of the distribution row; (2) variation indicators: dispersion, average quadratic deviation, and variation coefficient; and (3) the distribution form in terms of coefficients of asymmetry and kurtosis. Make some conclusions.

Exercise 5.18

Below are some data about the distribution of workers by level of remuneration (euro).

Remuneration	Structure of workers, %
less than1000	25
1000–1400	30
1400–2000	20
2000–2500	15
more than 2500	10
Total	100

Determine: (1) the modal and median interval, as well as the median of the distribution row; (2) variation indicators: dispersion, average quadratic deviation, and variation coefficient; and (3) the distribution form in terms of coefficients of asymmetry and kurtosis. Make and conclusions.

Exercise 5.19

Below are some data about the distribution of food stores in a city by volume of turnover.

Level of turnover volume, %	Number of stores
fewer than 5	5
5–7	10
7–10	28
more than 10	7
Total	50

Determine: (1) the modal and median interval, as well as the median of the distribution row; (2) variation indicators: dispersion, average quadratic deviation, and variation coefficient; and (3) the distribution form in terms of coefficients of asymmetry and kurtosis. Make some conclusions.

Exercise 5.20

The duration of use of short-term credit (days) by 50 companies in the region is shown below.

74	80	76	45	55	25	24	31	58	73
53	60	48	75	39	73	71	55	75	42
40	33	66	79	90	58	92	26	58	64
60	23	60	66	72	70	78	50	82	40
27	37	41	72	62	20	70	43	38	54

Build a variation row making four groups with equal intervals. Specify the distribution elements. Determine: average, mode, median, variation range, average linear deviation, and linear variation coefficient. Explain the contents of the calculated variation characteristics/attributes. Present the distribution form. Make some conclusions.

Exercise 5.21

The number of employees of 40 commercial banks of a city is characterized by the following data (people).

124	98	123	87	145	87	97	130	150	72
140	82	140	89	120	98	88	100	128	89
146	70	118	99	112	96	92	104	116	76
139	85	126	76	108	91	99	118	126	98

Build a variation row with equal intervals. Specify the distribution elements. Graphically represent the distribution row. Calculate the indicators of the distribution center. Make some conclusions as to the uniformity of the population and the distribution form.

Exercise 5.22

From the data in Exercise 5.13, determine the indicators of the distribution center, analyze the variation level of the indicator of average income per capita and present the distribution form. Graphically represent the distribution row. Make a comparative analysis of the variation level in terms of its dynamics and also in terms of the difference between urban and rural areas.

6. SAMPLE OBSERVATION

6.1. Sample observation and its application

Statistical studies mostly deal with mass events and processes and, in some cases, statistical research can be quite labor intensive. In addition, some methods of data control or testing require the destruction of the samples studied. The issue of substituting a total observation for a sample one becomes a necessary one. Theory and practice prove the feasibility and practical convenience of this substitution.

The intense growth in the scale of human activity and the expansion of spheres of human interest drive the need to develop accurate, high-quality, and complete information about the state and the dynamics of the development of socioeconomic processes and changing external conditions. From this perspective, first, total observations cannot cover all possible directions of research and second, such an approach is not applicable to the study of new processes.

It is simpler, faster, and cheaper to collect data about only a part of a population (not the whole population), but the sampled part has to represent the whole with a certain degree of accuracy, which can only occur under certain conditions. Experience of how to organize and conduct sample observations has grown across the world. Methodological principles and ways of forming sample populations, which are smaller than the events under study, as well as the rules and techniques needed to estimate the parameters of the object of observation, have been developed. All these methods ensure the accurate representation of the results of observation.

The technique of ***sample observation*** is a scientifically grounded way of dealing with only a part of a population. This part is chosen according to certain rules that ensure the results can characterize the whole population being researched.

The population from which the research units are chosen is described as the **general** population and the chosen part is called the ***sample***. There are certain principles on how to extract a sample population:

1. The principle of randomness – the guarantee of an equal chance for each population unit falling into the sample.
2. The principle of representation – the scope of sampling has to be sufficient to ensure the reliability of the results of the sample observation.

The ***characteristics*** of a sample population rely on estimates of the corresponding parameters of a general population. However, a sample does not necessarily reproduce the structure of a general population accurately and such estimates do not necessarily confirm the validity of the parameters. Such differences are called ***representation errors***. These differences can be either systematic or random (accidental).

Systematic errors can occur when the principle of random selection, giving an equal chance for all elements of a general population to be part of the sample, is not followed. Systematic errors for all elements of a population are unidirectional and are called bias errors. Errors that inevitably occur when the principle of random selection is practised, but that do not follow a broader tendency, are called ***random (accidental)*** errors and do not lead to a shift in estimation.

When a sample observation is undertaken, it is important to avoid systematic errors; however, random errors are both typical and inevitable. We can use statistical theory to identify appropriate margins with a certain reliability. A sample is representative when each unit has an equal chance of entering the sample population, which has to be of sufficient scope as well.

6.2. Types and patterns of sampling

When sample observations are used, the design mechanism of the sample population receives a particular meaning. The design of a sample population is the foundation of making a sample observation. There is an ordered process that follows certain rules, mechanisms, and stages. Various kinds of sampling are used in statistical practice.

There are different types and patterns of sampling. Their peculiarities influence the error size and the methods of its calculation. The sampling types are:

- random;
- systematic;

- typical;
- serial.

Simple random sampling is done using lots or a table of random (accidental) numbers. Such sampling requires serious preparation. An example of this would be the winning numbers in the Lottery.

Systematic (mechanical) sampling demands the presentation of a whole population in the form of a list compiled by some neutral feature. Element choice is done with equal intervals. As an example, to perform a 10 % sample observation of some students, a list of their names is compiled in alphabetical order and each tenth student is automatically chosen. An initial element is sampled as a random number from the first interval with the help of lots, for example "6". The following ordinal numbers of the elements are thus: 16, 26, 36… One can infer that this method is a variant of the previous one, but it is easier to organize.

Typical (stratified) sampling is based on the representation of the typical groups of a general population in a sample. Alongside this, the whole population is divided (stratified) into duplicate uniform groups of the same type. Then, a number of units, proportional to a specific weight for each group in the general population, are sampled from each group using any one of the methods specified above.

Serial sampling involves sampling from whole groups (series, cells), rather than from the different units of a population using a random or mechanical method. In each of these groups, a solid observation is identified and the results are applied to a whole population. For instance, this type of sampling is applied in the quality control of produce.

The application of either of the methods of sample population design depends on the purpose of the sample observation, as well as the conditions of its organization and conduct. The most common practice uses combined sampling. In addition to the types of sample population design, there are different patterns of sampling: repeated and non-repeated sampling.

Repeated sampling is characterized by the condition that each sampled unit returns to a general population and can fall into the sample again (for example, checking quality on a conveyer belt).

In ***irretrievable*** sampling, no sampled unit returns to the general population, e.g. playing the lottery.

In practice, the ***moment*** of an observation is widely used. Here, all the elements of a population (solid observations) are subject to observation at a certain moment in time. The concept of general and sample populations is connected to the time of the observation, rather than to the elements (units) of a population. Focus on the moment is a common practice when the structure of labor-hour expenditures is studied.

6.3 Sampling errors: Methodology of calculation

Below, general notation is suggested. The number of units of a general population is denoted with *N* and that of a sampling population with *n*. The generalized characteristics of a general population – average, variance, and proportion – are described as ***general*** and are denoted with $\bar{x}$, s^2, and *p*, respectively. Where *p* is the correlation of the number of *M* units of a population, which has the modal value of a feature and the total number of the general population (*N*), then $p = M/N$.

The generalized characteristics of a sampled population describe the ***sample*** and are denoted with $\tilde{\boldsymbol{x}}$, $\boldsymbol{\sigma_x^2}$, and *w*, respectively, where $w = m/n$. The theory of calculation of random errors can be found in the works of the famous scientists Jacob Bernoulli, S. Poisson, P. L. Chebyshev, A. A. Markov, and A.M. Liapunov.

The ***law of large numbers*** is a general principle according to which the accumulated effect of a large number of independent factors leads to a result that almost does not depend on randomness (occasion). In socioeconomic statistics, this can be understood as a matter of quantitative regularities, which are typical for mass events and expressed only in a large number of observations.

In every distinct sampling and among all possible errors, the random error of sampling $|\bar{x} - \tilde{x}|$ can have equal value. In the case of a large number of observations, the distribution of random errors of an average value approaches a normal distribution. Thus, we discuss the ***average (standard)*** error of sampling. It has been established that, for simple random sampling done by repeated sampling, the standard error is $\approx \sqrt{\frac{\sigma_x^2}{n}}$. If a sample observation is applied to a binary variable, then the average error of the proportion is calculated with the formula $\approx \sqrt{\frac{w(1-w)}{n}}$.

Using a function of the normal distribution, we can calculate the probability of a marginal error of a certain size. For instance, the probability an error in a defined sample does not exceed 2μ, i.e. 0.954, or 3μ, i.e. 0.997.

In the given formulas, p and q stand for the characteristics of a general population, which are unknown to the sample observation. For non-repeated sampling, the standard error is equal to $\approx \sqrt{\frac{\sigma_x^2}{n}\left(1-\frac{n}{N}\right)}$ and the proportional error is $\mu \approx \sqrt{\frac{w(1-w)}{n}\left(1-\frac{n}{N}\right)}$.

To estimate an acceptable level of a standard error, a relative sampling error is calculated: $\mu_{\%} = \frac{\mu}{\tilde{x}}$ x 100%. The relative error of sampling should be within 5 %. To solve practical problems, it is not enough to limit the calculation of the standard error of sampling and so a marginal error size of a certain probability is determined: $\Delta = t \times \mu$, where t is the quantile of a normal distribution, called the ***confidence coefficient (confidence number)***. Using a few examples, let us examine the determination of a marginal error for the average and the proportion.

Example 6.1

In a flock of 1000 sheep (N), a sample control shearing was done on 100 (n). The average wool sheared off amounted to 4.2 ($\tilde{x}$) kg per sheep with a standard deviation of 1.5 (σ)x kg. Determine the limits when the average wool from 1000 sheep has a probability of 0.954 (t = 2). In this sample, we use simple random sampling, which is non-repeated.

Let us put the data into the formula $\mu = \sqrt{\frac{(1.5)^2}{100}\left(1-\frac{100}{1000}\right)}$ = 0.142 kg.

$$\Delta = 2 \times 0.142 = 0.284 \text{ kg.}$$

One of the possible values within the range of which the average quantity of wool lies can be calculated as $\bar{x} = \tilde{x} \pm t\mu$. In a general form, this can be written as follows: $\bar{x}$ = 4.2 ± 0.284. This is equal to 3.92 kg ≤ $\bar{x}$ ≤ 4.48 kg. Thus, from the conducted sample observation, we can guarantee that the average quantity of wool will range from 3.9 to 4.4 kg per sheep in 954 cases out of 1000.

Example 6.2

To investigate product quality, 500 out of 10000 units were chosen. 50 items of the third category were recorded. Determine the marginal error of the proportion with a probability of 0.997.

The mode refers to items of high quality (first and second category). Therefore, the probability is *w* = (500-50)/500 = 0.9. The proportion of the items of the third category is *(1–w)* = 50/500 = 0.1. Let us put the data into the formula for simple random non-repeated sampling $\boldsymbol{\mu} = \sqrt{\frac{0.1 \times 0.9}{500}\left(1 - \frac{500}{10000}\right)}$ = 0.0131 = 1.31 **percentage points**. Therefore, the confidence limit of the proportion of third-category items is q = 0.1 ± 3×0.0131.

Thus, in the conducted observation, we have established that a proportion of third-category items will be within 10 % ± 3.9 percentage points.

For the probability of 997 cases out of 1000, we can state that a proportion of third-category items in the whole consignment will range from 6.1 to 13.9 %, counting 610 units and 1390 units, respectively.

The specified formulas for the average and marginal error of sampling are applied when random and mechanical samplings are done. For a typical sampling, a marginal error is calculated with the formulas given in Table 6.1.

Table 6.1. Marginal errors of sampling for a typical sampling procedure

Sampling pattern	Marginal error of sampling
Repeated	$\Delta_x = t \times \sqrt{\frac{\widetilde{\sigma^2}}{n}}$
Non-repeated	$\Delta_x = t \times \sqrt{\frac{\widetilde{\sigma^2}}{n}\left(1 - \frac{n}{N}\right)}$

For comparison with the formulas for random sampling, rather than of variance and proportion, which are determined in a sample population, in

general, it is necessary to calculate the averages from group variances and the proportion determined for each group for a typical sampling procedure.

Example 6.3

A typical sample with a 10 % proportion of the total number of workers' groups was undertaken. Find the average percentage of each group in terms of worker output with the general probability limit of 0.954. The sampling is not repeated.

Table 6.2. Sampling characteristics of workers

Groups of workers by specialty	Number, people	Average output rate, %	Standard deviation, %
T	40	98	2
S	50	108	3
F	60	104	5

We can calculate the average percentage of output of workers in the sample as follows: $\widetilde{x} = \frac{98 \times 40+108 \times 50+104 \times 60}{150} = 103.7$ %.

Now, we can determine the average group variance:

$$\widetilde{\sigma^2} = \frac{4 \times 40+9 \times 50+25 \times 60}{150}.$$

Then, the marginal error of the sample average for a typical sample: $\Delta_{\bar{x}} = 2\sqrt{\frac{14.1}{150}\left(1 - \frac{150}{1500}\right)} = 0.581$ percentage points (N = 1500 indicating a sampling rate of 10 %).

Thus, we can state that the average percentage of the output by workers in a manufacturing plant is within $\bar{x} = 103.7\ \% \pm 0.581$ percentage points with a probability of 0.954 and this is equal to $103.1\ \% \leq \bar{x} \leq 104.3\ \%$.

For serial sampling with an equally-large series, a marginal error should be determined with the formulas in Table 6.3, where s is the total number of series. Here, each series is a unit of a population and inter-series sample variance is a measure of fluctuation:

$$\delta^2 = \frac{\Sigma(\bar{x_j} - \tilde{x})}{s},$$

where $\bar{x_j}$ is the average for each series; $\tilde{x}$ is the general sample average; and s is the number of sampled series.

Table 6.3. Formulas for marginal errors in serial sampling

Sampling pattern	Marginal error of sampling
Repeated	$\Delta_x = t \times \sqrt{\frac{\delta^2}{s}}$
Non-repeated	$\Delta_x = t \times \sqrt{\frac{\delta^2}{s}\left(1 - \frac{s}{S}\right)}$

As such, for comparison with the results of random sampling, the calculation of inter-group variance of a serial sampling is necessary, rather than of general variance.

Example 6.4

To calculate the average harvest ratio of sugar beets, a 20 % serial sampling was taken for a region (5 out of 25 districts were included in the sample). The average harvest ratio for each district was 250, 260, 275, 280, and 300 centner/hectare per cultivation area of 800, 1000, 1200, 1200, and 2800 hectares, respectively. Below, we discover the range of average yield capacity of sugar beets within a probability of 0.954 in the region.

First, we find the general yield capacity

$$\tilde{x} = \frac{\sum x_j \times f_j}{\sum f_j} = \frac{250 \times 800 + 260 \times 1000 + 270 \times 1200 + 280 \times 1200 + 300 \times 2800}{800 + 1000 + 1200 + 1200 + 2800} = 280$$

centner/hectare.

Then, we determine the inter-series variance

$$\delta^2 = \frac{\sum(\bar{x}_j - \bar{x})s_j}{\sum s_j} = \frac{(250-280)^2\,800+(260-280)^2\,1000+(270-280)^2\,1200+(280-280)^2\,1200+(300-280)^2\,1200}{800+1000+1200+1200+2800} = 337.$$

Finally, we calculate the marginal error for serial non-repeated sampling

$$\Delta_x = t \times \sqrt{\frac{\delta^2}{s}\left(1 - \frac{s}{S}\right)} = 2 \times \sqrt{\frac{337}{5}\left(1 - \frac{5}{25}\right)} = \pm 7.34 \text{ centner/hectare.}$$

We can state that the average yield capacity of sugar beets in the region ranges from 272.66 cwt/ha to 287.34 cwt/ha with a probability of 0.954. In this way, the standard and marginal error for a proportion is calculated.

6.4. Determining the sample size

Before conducting a sample observation, it is imperative that we determine the appropriate sample population size, i.e. the volume and characteristics of a sample population that will ensure the accuracy of statistical observation.

The required size of sample *n* is determined with the formula for a marginal error. Formulas for random and mechanical samplings are given in Table 6.4.

Table 6.4. Formula for the determination of sample population size/coverage

Sampling pattern	Scope of sampling	
	average	proportion
Repeated	$\boldsymbol{n_x = \frac{t^2\sigma_x^2}{\Delta_x^2}}$	$\boldsymbol{n_w = \frac{t^2 w(1-w)}{\Delta_w^2}}$
Non-repeated	$\boldsymbol{n_x = \frac{N \times t^2 \times \sigma_x^2}{N \times \Delta_x^2 + t^2 \times \sigma_x^2}}$	$\boldsymbol{n_w = \frac{N \times t^2 \times w\,(1-w)}{N \times \Delta_w^2 + t^2 \times w\,(1-w)}}$

Example 6.5

There are 2500 cows in a district. In determining the random sample size for repeated and non-repeated sampling, the marginal error of average milk yield does not exceed 20 kg with a probability of 0.954 when σ_x = 300 kg.

For repeated sampling: $n_x = \frac{4 \times 300^2}{20^2}$ = 900 cows.

For non-repeated sampling: $n_x = \frac{2500 \times 4 \times 300^2}{2500 \times 4 + 4 \times 300^2}$ = 662 cows.

Example 6.6

Determine the sample size, within a proportional error of 3 percentage points, with the specified weight of a cow at 80 % and a probability of 0.954.

For repeated sampling: $n_w = \frac{4 \times 0.16}{0.03^2}$ = 711 cows.

For non-repeated sampling: $n_w = \frac{2500 \times 4 \times 0.16}{2500 \times 0.03^2 + 4 \times 0.16}$ = 554 cows.

When we determine the sample size in cases of unknown variance, these values can be determined approximately based on similar or test research. When a feature is alternative, it is assumed that w = 0.5, and with variance of $w \times (1\text{-}w) = 0.5 \times 0.5 = 0.25$.

6.5. Features of small sampling

A sample is considered small when its size ranges from 5 to 30 units. Small sampling is the only method of research available in those cases where the organization of a continuous or sample observation is not possible. The method of small sampling is most frequently used when the quality of an industrial product is studied or the output rate is determined. However, caution has to be exercised when using small sampling.

It is known from the theory of sample observation that representation in sampling, to a great extent, depends on the sample's volume. Random errors of sampling have a normal distribution when the size is rather large. Alongside this, a condition of equality of general and sample variance is assumed.

In the case of small sampling, these assumptions cannot be used. This peculiarity lies in the fact that random errors of small sampling are not

subject to the law of normal distributions. Student's *t*-test is used to estimate the results of small sampling and the possible limits of its random error $t = \frac{\bar{x}-\bar{x}}{\mu_s}$, where μ_s is the standard error of small sampling, calculated with the formula, $\mu_s = \frac{\sigma}{\sqrt{n-1}}$.

We can see from the formula that *n* is not used as the denominator, as is usually the case in sampling; rather we use *n-1*. It is important to note this when calculating the error for small sampling. The marginal error of small sampling is calculated using a standard technique with the formula $\Delta = t \times \mu_m$, where *t* refers to the quantiles of Student's *t*-distribution. The value of *t* depends on Student's distribution (and is true only for sampling taken from a general population with a normal distribution). To determine the probability *P*(*t*), we calculate the values of *P*(*t*) for the given values *t* and *k* = *n-1* (*k* is the number of degrees of freedom), which are taken from special tables (see Table 6.5).

Table 6.5. Probability *P(t)* of distribution *t* [$P_k(t)\times 1000$]

k / *t*	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**15**	**¥**
2.0	884	898	908	914	919	923	927	936	954
2.5	933	946	953	959	963	966	969	976	988
3.0	960	970	976	980	983	985	987	991	997

A two-sided criterion can be determined using the data in this table, i.e. the probability of the real value of *t* for various random reasons will not be bigger than in the table in terms of any absolute value.

Example 6.7

The small sampling procedure is used to check the quality of bulbs/lamps at a lamp manufacturing plant (Table 6.6). 10 lamps are sampled by a random non-repeated method. We can determine the marginal sampling error and find the confidence interval for the average.

Table 6.6. Calculation of a marginal error of sampling

Burning duration, hours (x_j)	**Number of bulbs/ lamps, pcs. (f_j)**	$x_j f_j$	$x_j - \tilde{x}$	$(x_j - \tilde{x})^2$	$(x_j - \tilde{x})^2 f_j$
1480	2	2960	26	676	1352
1500	4	6000	6	36	144
1520	3	4560	14	196	588
1540	1	1540	34	1156	1156
Total	10	15060	x	x	3240

The sampling average is equal to $\bar{x} = \frac{15060}{10}$ = 1506 hours. The sampling variance is given by $\sigma_s^2 = \frac{3240}{10}$ = 324 and the standard error is $\mu_s = \sqrt{\frac{324}{9}} = 6$ hours.

In the conditions of small sampling, when $k = n - 1 = 9$ and $t = 2.5$, with probability $P_k(t)$ equal to 0.966, the marginal error, as an absolute value, will not exceed $\Delta_s = 2.5 \times 6 = 15$ hours. The probability that this statement is not true and the error may exceed the limits of 15 hours is equal to 1 – 0.966 = 0.034. We can find the confidence interval for a general average based on the calculated characteristics as $1491 \text{ hours} \leq \bar{x} \leq 1521 \text{ hours}$.

6.6. Use of agreement criteria

During the course of statistical analysis, there is always a need to compare distributions. If we have an empirical distribution and choose a theoretical curve expressing a regular distribution with which to compare it, this process is called *simulation.* We ask to what extent a sampled theoretical curve matches an empirical one – its extent of similarity.

In other cases, we can compare two empirical curves. It is important to find out how similar they are; more exactly, whether they differ considerably. In

other words, we visualize whether the populations differ significantly in terms of their distributions. We examine this in examples 6.8–6.12.

Example 6.8

Data on the mass of animals placed in control and experimental groups where the latter received certain feed additives is given below.

Group	Mass (kg)						
	3	4	5	6	7	8	9
Experimental-Group	1	1	6	11	8	4	1
Control Group	1	4	9	7	2	x	x

Does the feed additive increase the mass of the animals?

Example 6.9

Data is given below on monthly wages (euro) of workers in similar jobs at two small enterprises.

	Month					
	1	2	3	4	5	6
Enterprise A	1000	1000	1000	1000	1000	1000
Enterprise B	0	2000	0	2000	0	2000

Do the workers at these two enterprises receive the same wages over the course of six months?

Example 6.10

In a questionnaire, workers were asked the question “Are you satisfied with your working conditions?” The answers of men and women were grouped as follows.

Options for the answer	Men	Women
Satisfied	5	6
Rather satisfied	8	7
Rather not satisfied	6	10
Not satisfied	11	17

Do men and women have the same perception about working conditions?

Example 6.11

Data on seven years of crop yields of experimental and control groups on agricultural land is given below, cwt/ha.

Year	1	2	3	4	5	6	7
Experimental land	22.9	20.2	19.5	30.5	35.6	31.9	27.7
Control land	19.4	16.2	16.9	29.3	31.4	28.5	25.6

Does a pre-sowing treatment of the land have an effect on crop yield?

Example 6.12

The pH values of 10 samples of a solution were determined to be 7.48 ± 0.21 ($t = 2.28$). Can we consider this solution to be alkaline (if the reactions are alkaline at pH > 7.0)?

There are many examples like this. From a statistical point of view, the question can be framed as follows: for two given populations (the ones to be compared), are the samples from the same general population or from two different general populations?

In fact, there is always a difference between two populations (distributions) in terms of certain features. The question is whether the difference is random or not, i.e. whether it is reliable, essential, and ***significant***. Thereby, a hypothesis about ***the absence*** of a real difference can be checked; this is called the ***zero hypothesis*** and denoted H_0. Different methods (estimates) are used to test it according to **statistical criteria**.

The level of significance determines the extent of a researcher's risk in mistakenly rejecting H_0. It is important to remember:

1. The word "significant", which we use to define the difference between distribution rows characterizes "non-randomness" rather than size.
2. A lack of sufficient grounds to reject H_0 does not prove the absence of a difference.
3. Depending on the context, one and the same criterion may need to be calculated with different formulas.
4. A researcher has to choose the level of significance, assessing the likelihood of possible outcomes. Values of 1 %, 5 %, or others, which are traditionally mentioned in the literature, are only ***suggested***, but not required.

Hence, it is obvious that a statistical criterion has to be correctly used in research, taking into account the various possible outcomes of the decisions made. The nature of populations to be compared, in terms of feature type, form of distribution, and sample size etc., are of critical importance when choosing a criterion for an individual case.

Some criteria can be used in the case of a distribution that is close to normal. Differences between distributions can be estimated by comparing their averages. Student's *t*-criterion can be used for this purpose (Example 6.8). Example 6.9 shows the absurdity if one uses average levels (e.g., wages, radiation, number of fur coats per woman), ignoring a distribution pattern consciously or by mistake. In fact, the same "averages" can show exactly opposite principles of remuneration.

When distributions of a qualitative feature are compared (Example 6.10), Pearson's *chi*-square criterion can be used. Comparable samples should have at least 20-30 units each and a sampling frequency of at least 4-5 units. The use of certain criteria does not require the calculation of distribution parameters (average, variance and others) and these are described as ***nonparametric***. They fall into three groups:

1. Criteria with difference in central characteristics.
2. Criteria enabling the identification of difference, but that do not explain it.
3. Criteria for matching paired populations.

V.Yu. Urbach gives two examples of the calculation of the criteria belonging to each group. The explanations are detailed below.

The criteria are:

1. White test and criterion *X* (**Van der Waerden test**).
2. Serial criterion (Kendall Rank Correlation) and Kolmogorov-Smirnov criterion.
3. Criterion of signs and Wilcoxon criterion.

Student's *t*-criterion is used to test a hypothesis about the averages of two normally distributed general populations. A frequent mistake is to ignore assumptions about the variance of distributions. Formulas to calculate the *t*-criterion differ so as to take into account the parameters of the data; four distinct cases can be observed.

Let us consider independent and dependent (interconnected) samples. Example 6.8 presents the case of an independent sample. The criterion is calculated with the formula $t = \frac{\tilde{x}_1 - \tilde{x}_2}{\sqrt{\mu_1^2 + \mu_2^2}}$, where $\tilde{x}_1$ and $\tilde{x}_2$ are the average values in two samples and μ_1 and μ_2 are the standard errors of samples, $\mu = \frac{\sigma}{\sqrt{n}}$ (6.1).

The number of degrees of freedom is given by: $f = n_1 + n_2 - 2 = 32 + 23 - 2 = 53$.

From the data in Example 6.11, let us estimate the effect of a pre-sowing treatment of agricultural land on yield capacity. We can calculate the sample averages where the harvest ratio is 26.9 centner/hectare for the experimental land

$$\tilde{x}_1 = \frac{22.9+20.2+19.5+30.5+35.6+31.9+27.7}{7} = \frac{188.3}{7} = 26.9 \text{ centner/hectare.}$$

For the control land, the average harvest ratio is lower, with 23.9 centner/hectare

$$\tilde{x}_1 = \frac{19.4+16.2+16.9+29.3+31.4+28.5+25.6}{7} = \frac{167.3}{7} = 23.9 \text{ centner/hectare.}$$

Now, we can calculate the variances of the harvest ratio on each plot of land using the formula. For the experimental land, the variance is equal to $\sigma_{6\,1} = \frac{22.9^2 + 20.2^2 + 19.5^2 + 30.5^2 + 35.6^2 + 31.9^2 + 27.7^2}{7} - 26.9^2 = 32.85$;

for the control plot, $\sigma^2 = \frac{19.4^2 + 16.2^2 + 16.9^2 + 29.3^2 + 31.4^2 + 28.5^2 + 25.6^2}{7} - 23.9^2 = 34$.

Using the data in formula 6.1, $t = \frac{\tilde{x}_1 - \tilde{x}_2}{\sqrt{\mu_1^2 + \mu_2^2}} = \frac{26.9-23.9}{\sqrt{\frac{32.85}{7} + \frac{34}{7}}} = 0.97$.

We can find a theoretical value for Student's criterion from the table of critical values (Table 5) at α = 0. 05 and for degrees of freedom, *k* = 7 + 7-2 = 12. This gives us a value that is equal to $t_{0.05}(12) = 1.78$, which is bigger than the real value (0.97). Thus, the null hypothesis on the absence of effect of the pre-sowing treatment on yield capacity is not valid.

The value in the denominator (6.1.) is known as ***an average error of a difference*** and is calculated as, $\bar{\mu} = \sqrt{\mu_1^2 + \mu_2^2}$ where samples are independent. However, quite frequently, and especially in biological research, this happens in a different way. For example, this would be different if we studied the harvest ratio of two adjacent agricultural plots or the cholesterol level in a certain group of people before and after treatment. We may infer that, in the first case, the connection is similar weather conditions; in the second case, it is found in the similar composition of the human body.

In such cases, populations are described as independent and paired match. Then, $\bar{\mu} = \sqrt{\frac{\sum(\Delta_i - \bar{\Delta})^2}{n\,(n-1)}}$ (6.2.), where x_i and y_i are paired match variants: $\Delta_i = x_i - y_i$; $\bar{\Delta} = \bar{x} - \bar{y}$.

Example 13

What rule is used to determine the marginal error of small sampling?

Example 14

Using examples 6.8–6.11, define the null and alternative hypotheses and test them to a 0.05 level of significance.

Practice Exercises

Exercise 6.1

A 2 % mechanical sampling of 10000 depositor accounts in a city's commercial banks found an average deposit of 2350 euro with a standard deviation of 520 euro. Determine the probability that the marginal sampling error of an average deposit does not exceed 100 euro.

Exercise 6.2

A sample observation of a telephone network was done to study the average duration of telephone calls. Having analyzed 5000 observations, it was found that the average duration of a telephone conversation was 25 minutes with a standard deviation of 16 minutes. Determine the limits for the average duration of a conversation with a probability of 0.954 in the general population. Make a conclusion about the relative error.

Exercise 6.3

A cultivated area of corn covered 5000 ha. Using random non-repetitive sampling, 200 ha were observed. 160 ha of hybrid corn were found. Determine the limits for the general proportion of hybrid corn fields with a probability of 0.954, estimate the relative sampling error, and draw conclusions about the error size.

Exercise 6.4

A sample observation of 2500 passengers was done to study the average distance of a passenger's trip by commuter train. It was found that the average passenger trip distance was 35 km with a standard deviation of 10 km. Determine the limits for the general average passenger trip distance with a probability of 0.954 and make conclusions about the size of the relative sampling error.

Exercise 6.5

A 10 % sample observation was conducted in a city to study the free time of the population. 900 people were in the sample. The expenditure of free time on different activities by resident per week was tabulated in minutes.

Type of leisure activity	Average, minutes	Standard deviation, minutes
Reading newspapers	136.8	120
Reading fiction	276.0	180
Entertainment	102.0	90

Determine the marginal sampling error for average leisure time by activity with a probability of 0.954. Make a conclusion about the average size of sampling errors. Build appropriate confidence limits.

Exercise 6.6

The average service life of 120 machine tools at a machine tool plant was determined by mechanical sampling to be 12 years. The standard deviation was 3.5 years. Determine the marginal sampling error and the confidence interval for the average service life of a machine tool in a general population of 1250 with a probability of 0.954.

Exercise 6.7

From a 5 % sample observation of 15000 employees in state-run institutions, it was found that the average employee takes 65 minutes to get to work at s = 10 min. Determine the relative sampling error with a probability of 0.954 and define the confidence interval. Make some conclusions.

Exercise 6.8

A 5 % sample observation of 5000 passengers was carried out to study the average distance of passenger trip by long-distance train over a year. The results showed that the average passenger trip distance was 750 km with a standard deviation of 125 km. Determine the limits for the general average passenger trip distance with a probability of 0.954 and make some conclusions about the size of the relative sampling error.

Exercise 6.9

A 10 % sample observation was conducted on the interest rates (%) of deposits in some banks.

16.2	19.9	13.3	14.2	16.3	18.2
12.6	18.3	11.6	16.7	11.9	10.8
16.9	12.9	19.7	16.2	19.9	17.9
11.5	27.9	12.9	18.6	16.1	12.9
12.9	9.8	12.8	15.6	10.3	17.7

Using the given data, determine the possible level of interest rates on deposits in a general population with a probability of 0.997. Establish the reliability of the results.

Exercise 6.10

A 10 % sample observation of students in an evening class was carried out at an institute to study the match of job position with chosen career path. The results showed that 100 people were working in their chosen profession. 1900 people were in the evening class. Determine the general proportion and the number of students who do not work in their chosen career with a probability of 0.954.

Exercise 6.11

From the data of a 2 % sample observation of the size of some parts, with a total of 30000 pieces, the average size of the parts was found to be 15 cm with a variation coefficient of 26 %. Determine the relative sampling error for the average part size with a probability of 0.997. Make some conclusions.

Exercise 6.12

A 3 % sample observation of lamps for quality checking purposes at an electric lamp manufacturing plant showed that 32 lamps out of 1600 were rejected. Find the limits of the general proportion of rejected lamps with a probability of 0.954 and the possible number in the whole population.

Exercise 6.13

Data from a 4 % sample observation showed that the amount of free time of one city resident per day was 210 minutes, when s = 40 min. 1000 people were observed. Find the marginal sampling error for the average amount of

free time with a probability of 0.954 and make a conclusion about the relative value.

Exercise 6.14

A sample observation of the quality of some electric lamps, with 250 thousand items in total, was conducted at an electric lamp manufacturing plant. Determine the quantitative size of sampling at which the marginal error of the proportion of rejected lamps will not exceed 0.4 percentage points at a probability of 0.997.

Exercise 6.15

A company decided to conduct a sample observation of rejected bottles from a total of 100 thousand bottles per day coming from trade businesses. What should the number of samples be to ensure that the marginal sampling error is not larger than 3 percentage points at a probability of 0.954?

Exercise 6.16

From a 5 % sample observation at a post office we need to determine the proportion of letters sent by private individuals. No preliminary data are available about the specific weight of these letters in terms of general correspondence. The results from the sample should be estimated with a level of accuracy of 4 percentage points and a probability of 0.954. What should the number of this sample be?

Exercise 6.17

We need to determine potential wood supply from woodland with an area of 5000 ha, using mechanical sampling. The test plot is 0.05 ha. Determine the size of the sample needed to achieve a level of accuracy at 2.5 m^3 and with a guaranteed probability of 0.954. From previous observations, the standard deviation of wood output per 0.2 ha was found to be 15 m^3.

Exercise 6.18

The quality of 5 % of output was checked using non-repeated sampling. 10 out of 500 items were rejected. Calculate the standard error of sampling of the proportion of rejected items and estimate:

1) how the sampling error will change, if the sampling percentage decreases by 2.5 times;

2) how the sampling error will change, if the number of rejected items increases by 2 times;
3) how the sampling error will change, if the number of rejected items decreases by 2 units and the size of the sample population decreases by 100 units;
4) how the sampling error will change, if the size of the sample population increases by 1.5 times.

Exercise 6.19

A moment of observation was used to analyze a taxi fleet of 100 vehicles to check the efficiency of vehicle use. How many observations should be conducted to get a sampling error for the proportion of cars on downtime not exceeding 2.5 percentage points with a probability of 0.997? How many observations of each car should be performed every day if the observation lasts 7 days?

Exercise 6.20

From the data of a 2 % sample observation of consumer expenses of 500 households in a city, the level of expenditure on certain groups of items was found: foodstuffs, 55 %; non-food products, 25 %; and service consumption, 20 %. What will the limit of the marginal sampling error be with a probability of 0.954? How many households have to be observed to confirm that the relative error remains within 5 %?

Exercise 6.21

There are 2000 trees in an orchard. A sample observation was conducted to determine yield capacity. The orchard was divided into 100 plots and 5 plots were sampled with a mechanical method. The yield capacity (cwt/tree) was as below.

Plot №	1	2	3	4	5
Yield capacity	12	15	10	11	14

Determine:

1) the range of average yield capacity of the orchard with a probability of 0.954;

2) the number of plots that need to be observed to ensure that the relative sampling error of the average yield capacity of the orchard remains within 5 %.

Exercise 6.22

Serial sampling was used to study the quality of some produce in glass jars. 5 jars were sampled from 100 boxes with 10 jars in each box. The results of observation showed that there were some cases of poorly sealed jars.

Box №	1	2	3	4	5
Number of jars	1	0	3	0	1

Determine:

1) the range for the proportion of poorly sealed jars with a probability of 0.954;
2) the number of boxes that have to be checked to ensure that the relative sampling error of the proportion of poorly sealed jars remains within 5 %.

Exercise 6.23

A sample observation was conducted to study the effect of new harvesting technology on yield losses. Traditional harvesting practice was applied to 10 hectares and the yield loss was 10 %; the new practice was used on 8 hectares and the yield loss was 5 %. Define the null and alternative hypotheses. Test the null hypothesis with Student's criterion for a probability of 0.975.

Exercise 6.24

120 women and 80 men were interviewed to study the vulnerability of men and women to a risk. 50 women and 55 men were found to be vulnerable to the risk. Define the null and alternative hypotheses. Test the null hypothesis with Student's criterion for a probability of 0.95 and make a conclusion.

Exercise 6.25

Below are some data about the average grade point of 10 students on a school certificate and in the first examination session.

№ /п	Average grade point		№	Average grade point	
	school certificate	exam session		school certificate	exam session
1	4.8	4.7	6	3.,3	4.1
2	4.4	4.2	7	4.0	3.7
3	4.2	4.4	8	3.9	3.0
4	5.0	5.0	9	4.7	4.3
5	4.5	4.9	10	3.7	3.2

Define the null and alternative hypotheses. Use Student's criterion and make conclusions about the effect of school achievement on the results of the first examination session at a probability of 0.9.

Exercise 6.26

The term of use of short-term credit (number of days) by companies in the region (with mechanical sampling of 2 %).

74	80	76	45	55	25	24	31	58	73
53	60	48	75	39	73	71	55	75	42
40	33	66	79	90	58	92	26	58	64
60	23	60	66	72	70	78	50	82	40
27	37	41	72	62	20	70	43	38	54

Determine the range of general average duration of credit use with a probability of 0.954 and make conclusions about the relative error size.

Exercise 6.27

After mechanical sampling of 20 % of 10000 depositor accounts in the saving banks of a city, it was found that the average account balance was 1200 euro with a standard deviation of 100 euro. Determine the relative sampling error and the probability when the marginal error of the average deposit balance does not exceed 4 euro.

Exercise 6.28

From a sample observation at a post office, determine the proportion of letters sent by private individuals. No preliminary data about the specific weight of these letters in a general quantity of correspondence are available. Estimate the results of sampling with an accuracy of 2 percentage points and with a probability of 0.954 when the general quantity of correspondence is 20000 letters per day. Determine the number of this sample.

Exercise 6.29

To study the average duration of telephone conversations, a sample observation of a telephone network was conducted. 500 observations showed the average duration of a conversation to be 15 minutes with an average quadratic deviation of 5 minutes. Determine the limits of the general average duration of a conversation with a probability of 0.954. Draw some conclusions on the relative sampling error.

Exercise 6.30

Determine the potential supply of wood from woodland with an area of 1000 hectares using mechanical sampling. The size of the test plot is 0.25 hectare. Determine the size of the sample to be estimated with accuracy to 2 m^3 on the test plot for a probability of 0.954. The standard deviation of wood output per 0.25 hectare is 20 m^3.

Exercise 6.31

A cultivated area of corn was 10000 ha. Using random non-repeated sampling, 1000 hectare were observed to have 800 hectares of hybrid corn. Determine the limits of the general proportion of hybrid corn fields for a probability of 0.954, estimate the relative sampling error, and make conclusions on the size of the error.

Exercise 6.32

A sample observation of 1000 passengers was conducted to study the average distance of passenger trips by commuter train. The average distance of one passenger trip was 24.2 km with a standard deviation of 8 km. Determine the limits of the general average passenger trip distance for a probability of 0.954 and make some conclusions about the size of the relative sampling error.

Exercise 6.33

To determine the proportion of letters addressed beyond Ukraine's borders, a15 % random sample observation was planned at a post office. How many letters have to be sampled? Estimate the results with an accuracy limit of 2.5 percentage points and a probability of 0.954.

Exercise 6.34

A company decides to conduct a sample observation of the proportion of rejected bottles. Determine the size of the sample for a probability of 0.954 with a standard sampling error not exceeding 3 percentage points.

Exercise 6.35

The proportion of rejected products was 3 % at the first enterprise and 5 % at the second one. Find the enterprise where the sampling error is larger, if the size of the sample is the same for both enterprises. Justify your choice.

Exercise 6.36

A sample observation was conducted at a lamp a manufacturing plant to check the quality of 1 million lamps. What should the size of the sample be for a probability of 0.954 with the standard error of the proportion of rejected lamps not exceeding 1.5 percentage points?

Exercise 6.37

A standard sampling error for the proportion of rejected products was 1.2 percentage points and the specific weight of the rejected products in the sample was 3 %. With what probability can we state that the proportion of rejected products in a general population would not exceed 7 %?

Exercise 6.38

The work experience (in years) of bank employees (from a 5 % mechanical sampling) is as follows.

9	4	13	8	5	10	9	3	10	12
5	6	7	6	8	5	4	11	3	7
8	9	6	2	4	8	6	12	8	11
2	3	5	4	6	7	8	4	9	8
7	5	8	7	5	3	5	7	7	10
6	12	10	12	9	6	7	8	6	12

Determine the sampling error for a probability of 0.954 and define the confidence interval for average work experience. Draw some conclusions about the error level.

7. Statistical Methods for Measuring Interconnection

7.1. Types of interconnections between events

All events and processes that exist in nature and society are interconnected and interdependent. As such, the study of interconnections and relationships of cause and effect is one of the most important areas of statistics. A cause-effect relationship is the primary form of these regular connections, but the cause itself does not define the effect to a full extent; the latter depends on the conditions within which a cause operates. Conditions and causes are factors. A feature that characterizes an effect is called *effective*, whereas one that characterizes a cause is called *factorial*.

Statistical regularity is one of the manifestations of the regular connection between the preceding and the next state of the system. It is formed by a set of elements and is under the influence of constantly changing conditions. From a systems approach, the formal presence of regularity in a system means that its current state will define its future state in a non-unique fashion with some degree of probability, which is an objective measure of the possibility of change in a system from its past state.

A statistical regularity is considered to be a quantitative regularity of change in the space and time of mass events and processes of social life and consists of a set of elements. Statistical regularity, in accordance with the law of large numbers, is typical of the mass process, rather than of the individual elements of the process. It is only expressed as the average.

As the spatiotemporal intervals of the development of a phenomenon increase, its regularity becomes more and more established and, therefore, it becomes more and more subject to probabilistic determination. Thus, having information on the developmental regularities of a socioeconomic system, it is possible to envisage its subsequent development with a certain degree of probability and to determine the size of an initial indicator characterizing a system's effectiveness both in general and in terms of its individual elements. However, it should be taken into account that significant changes in external conditions can lead to significant changes in the strength of such regularity.

The connections between events existing in nature and society are classified as functional and stochastic. In a ***functional*** connection, each possible value of a factor, *x*, matches a determined value of a result, *y*. For example, we see such relationships in physical and chemical processes, etc. This can be presented in a schematic diagram (Figure 7.1).

$$x_1 \rightarrow y_1$$

$$x_2 \rightarrow y_2$$

$$x_3 \rightarrow y_3$$

$$x_4 \rightarrow y_4$$

--- --- ---

$$x_n \rightarrow y_n$$

Figure 7.1. Diagrammatic representation of a functional connection

Most frequently, in any social process, this is seen in the connection between the elements of computational formulas, for example, a weighted relationship between yield capacity and cultivated area.

There is no one-to-one correspondence between the features that characterize cause and effect. As such, several values of a result (effect) correspond to one value of a factor (cause). The elements of such a set form conditional distributions. Statistical regularities exist in the forms of distributions, interconnections, and tendencies, etc. They are not available to direct observation and cannot be studied empirically. In the case of a stochastic connection, when several values of a result correspond to one value of a factor, a difference is observed within the distribution. The more noticeable such a difference, the closer the connection between the cause and the effect. A correlation is where a variant of a stochastic connection is observed in the event of a change in average conditional distributions.

In a *stochastic* connection, each value of a factor, *x*, matches a certain set of a results, *y*, which vary and form a distribution row, described as *conditional*. A stochastic connection is observed in the event of a change in conditional distributions. This can be shown in a schematic diagram (Figure 7.2). An example of such a connection is the relationship between qualifications and remuneration.

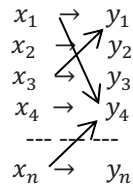


Figure 7.2. Diagrammatic representation of stochastic connection

Methods of statistical simulation play a dominant role in identifying and measuring statistical regularities. Through combining components of qualitative and quantitative analysis in a model, statistics not only identifies and measures regularities, but also comprehensively studies the mechanisms of their occurrence. It is possible to study functional and stochastic connections with statistical models; these models are constructed through spatial sampling and time rows. They measure both the direct and indirect effects of the factors, as well as studying their individual and cumulative effect.

7.2. Stages of studying interconnections between events

The study of the interconnections of complex events and processes consists of several stages. Firstly, statistical indicators are calculated from the statistical observation of an economic phenomenon and the nature of an event.

The second stage is to establish the quantitative connection between the chosen features. A quantitative estimate of a connection can be determined by different methods. If only the closeness of the interconnection of qualitative indicators is to be estimated, this stage becomes the final one. If the interconnection of quantitative indicators is to be estimated, then confirmation of a hypothesis about the existence of a connection leads to the third stage, which is to determine any analytical dependence between the features.

The type of analytical dependence or a concrete formula that determines a mutual correspondence between factors is chosen to undertake the informative analysis of an event. If nothing is known about the nature of the interconnection in advance, then, in the process of research, different hypotheses are checked and various formulas are tested to divine the one that is most probable and corresponds to the available data.

The third stage of research may be conducted with the use of methods of correlation-regression analysis. Here, the change of an average level of one feature or several effective features is determined from changes in the real values of another indicator or values of several features-factors, through paired or multiple regression, respectively.

The fourth stage of studying interconnection envisages estimation of the reliability of the results generated. The estimation of reliability is made with the hypothesis that the data generated from observation and the results of processing are a sample of the general population. The results of reliability estimation of the calculated values of the parameters of the feature of interconnection specify the existence of a connection and its form. We choose the most significant features and build a system of interconnections and groups of indicators.

7.3. Method of analytical grouping: Variance analysis

A general presentation of the method of analytical grouping is given in Chapter 2. The main idea of this method is that of grouping all the elements of a population by a factorial feature. An average value of the effective feature is calculated for each group.

For example, we singled out groups of workers by category and calculated an average wage for each group. We found that higher wages corresponded to groups with more qualified workers implying a direct connection between these two factors ("qualification" and "wage"). We can affirm this with a certain probability only once we prove non-randomness – the ***significance*** of a difference in the averages, and thereby, a ***significant*** connection. This can be done with Student's criterion. Thus, we can determine the presence of a connection and its direction.

Besides qualifications, other factors also influence the average wage of workers, such as sickness, nature of production, and gender etc. The method of variance analysis allows us to determine the effect of each factor, as well as the proximity of the connection. Example 7.1 illustrates the method.

Example 7.1

Some data about the hourly output of parts produced by workers in two groups who had been re-trained (N_1) and who had not been re-trained (N_1) are presented. Each group consisted of 5 workers.

Table 7.1. Analytical grouping of the dependence of hourly output of workers who had and did not have re-training

Group	Number of people f_j	Hourly output of parts (pieces) y_j	Average output $\overline{y}_j$
Group 1	5	40; 48; 43; 45; 44	220/5 = 44
Group 2	5	62; 66; 60; 68; 64	320/5 = 64
Total	10	540	540/10 = 54

Variance analysis makes it possible to define the role of systematic and random variation in terms of general variation and to define the role of the factor in the variation of the result (effect). Therefore, the rule of the addition of variance is used where a general variance is equal to the sum of two variances – an inter-group one and an average one from inter-group variances $\sigma^2 = \delta^2 + \overline{\sigma^2}$. Comparison of an intergroup variance with the general one characterizes the proximity of the connection/density. This relationship is described as the ***correlation ratio***: $\eta^2 = \frac{\delta^2}{\sigma^2}$. Let us calculate these parameters for a given example for group and general averages. Columns 4–7 in Table 7.2 represent the result of calculations.

Table 7.2. Calculation table for variance

№ п/п	Hourly output of parts		Individual deviation from a general average		Square of individual deviation	
	group 1	group 2	group 1	group 2	group 1	group 2
1	2	3	4	5	6	7
1	40	62	40-54=-14	62-54=8	196	64
2	48	66	48-54=-6	66-54=12	36	144
3	43	60	43-44=-11	60-54=6	121	36
4	45	68	45-54=-9	68-54=14	81	196
5	44	64	44-54=-10	64-54=10	100	100
Total	220	320	-50	50	534	540

A general variance characterizes a general variation due to the effect of all the factors and is calculated with the formula

$$\sigma^2 = \frac{\sum(y_i - \bar{y})^2}{n} = \frac{534 + 540}{10} = 107.4, \text{ where } \bar{y} = \frac{220+320}{10} = 54 \text{ pieces.}$$

An intergroup variance characterizes a factor variation, i.e. differences in output due to the fact that some workers had re-training, and is calculated with the formula

$$\delta^2 = \frac{\sum_{j=1}^{f_j}(\bar{y}_j - \bar{y})^2 f_j}{\sum f_j} = \frac{5 \times (44-54)^2 + 5 \times (64-54)^2}{5+5} = 100,$$

where $\bar{y}_1 = \frac{220}{5} = 44$ pieces and $\bar{y}_2 = \frac{320}{5} = 64$ pieces. $\bar{y}_j$ is the group average; $\bar{y}$, is the general average; f_j is the number of units in a group; and j is the number of the group. The average of intergroup variances characterizes a residual variation, i.e. a variation of the result (effect) caused by the rest of the factors, and is calculated with the formula

$$\overline{\sigma^2} = \frac{\Sigma \sigma_j^2 \times f_j}{\Sigma f_j}, \sigma_j^2 = \frac{\Sigma (y_{ij} - \bar{y}_j)^2}{f_j},$$

where σ_j^2 is the group average and y_{ij} refers to the individual values of an effective feature in j-group.

Let us calculate the group variances from Table 7.3. The average of intergroup variances will be equal to

$$\sigma_1^2 = \frac{34}{5} = 6.8, \sigma_2^2 = \frac{40}{5} = 8,$$

$$\overline{\sigma^2} = \frac{6.8 \times 5 + 8 \times 5}{5+5} = 7.4.$$

Table 7.3. Calculation table for group variances

№ п/п	Hourly output of parts		Individual deviations from the group average		Square of individual deviation	
	group 1	group 2	group 1	group 2	group 1	group 2
1	40	62	40-44=-4	62-64=-2	16	4
2	48	66	48-44=4	66-64=2	16	4
3	43	60	43-44=-1	60-64=-4	1	16
4	45	68	45-44=1	68-64=4	1	16
5	44	64	44-44=0	64-64=0	0	0
Total	220	320	0	0	34	40

Now, we check the variance sum rule: 100 +7.4 =107.4. As such, the correlation ratio is $\eta^2 = \frac{100}{107.4} = 0.931$ i.e. 93.1 %. Thus, 93.1 % of output variation is from the grouping factor, i.e. taking a re-training course, and only 6.9 % of variation is due to the other reasons. For example, this could be the worker's age or work experience etc. A correlation changes from 0 to 1 when an intergroup variance is equal to zero. This is possible only when all group averages are equal and there is no correlation connection between the averages.

A connection of ***correlation*** here refers specifically to a kind of statistical connection where, with a change in factor, x, the average value of the result (effect), y, changes. Thus, an intergroup variance is equal to a general variance and the average from an intergroup variance lies between 1 and 0. This means that one value of the result corresponds to every value of a factor, i.e. there is a functional connection between the factors.

Now, let us divide workers into two groups by the feature of the number of letters in their last name (odd or even). Calculated group averages also differ and this can be a matter of randomness, but not a characteristic of the correlation. Thus, the significance (non-randomness) of a connection requires checking.

The significance of deviations in the group average is examined with the help of statistical criteria. In this case, we can use the Fisher F-criterion or compare a real value with a corresponding value from the critical table. We can see (appendix, Table 2) that the distribution depends on the number of degrees of freedom of factorial, k_1, and random, k_2, variances, where $k_1 = m\text{-}1$; $k_2 = n\text{-}m$ (m is the number of groups and n is the general size of the population).

In our example, $k_1 = 2\text{-}1 = 1$ and $k_2 = 10\text{-}2 = 8$.

From the table of critical values for the level of probability a = 0.05, we define $\eta^2_{0.05} = 0.399$ (appendix, Table 2). This means that a correlation randomly occurs in only 5 out of 100 cases and will not exceed 0.399. Now, we compare a real value with a critical one. If a real value is larger than the critical value, then the connection between the factor and the result is considered to be significant and any differences between the groups are not random.

$$0.931>0.399 \Longrightarrow \eta^2 > \eta^2_{0.05}.$$

The connection between worker re-training and an increase in labor productivity is thus significant.

The Fisher F-criterion is also used for checking the level of significance of a connection. With a large value for degrees of freedom, the values in the table do not change much.

To determine the proximity of a connection, the index of correlation is calculated as a square root from a correlation ratio $\eta = \sqrt{\eta^2} = \sqrt{0.931} = 0.965$. Using a scale of connection proximity (Table 7.4), we can infer that

there is a close connection between the re-training of workers and the increase in labor productivity.

Table 7.4. Scale of closeness of connection

Value of correlation index	Level of connection
0–0.3	Weak
0.3–0.5	Medium
0.5–0.7	Noticeable
0.7–0.9	Close
Over 0.9	Very close

In variance analysis, a factor can be both quantitative and qualitative. Despite its advantages compared to other methods, variance analysis does not allow study of the form of a connection. When we have a sufficient number of groups and a quantitative factor with a proven connection significance, we can find certain points on the X and Y coordinates, connect them on a curve, and formulate a model of the form of connection.

7.4. Correlation-regression analysis

The main characteristic of a connection of correlation is the regression line. ***A regression line between x and y*** is the function that connects the average values of feature *y* with the average values of feature *x*. Depending on the form of the regression line, we can distinguish linear and non-linear connections. A regression line can be presented in the form of a table or a figure, or analytically. In a correlation-regression analysis (CRA), the estimate of a regression line is not done at different points, as in an analytical grouping, but rather at every point of the intervals in changes of factor *x*. A regression line is continuous and is shown in the form of a function, $Y = f(x)$. This is called a regression equation. Y is the theoretical value of a result (effect).

We can illustrate the sense of LSM (least square method) using a simple example (with the assumption that all numerical values are nominal). If we know that a segment of a metal pipe of a certain diameter is 1 m long and weighs 10 kg, we can estimate the exact weight. To be more exact, we can

estimate the mass of any segment of the metal pipe with the same diameter. Several segments would constitute a statistical population. We can measure the length of each segment and calculate its mass with the formula $y = m \times x$, where m is the mass of a segment of 1 meter long and x is the length of the pipe.

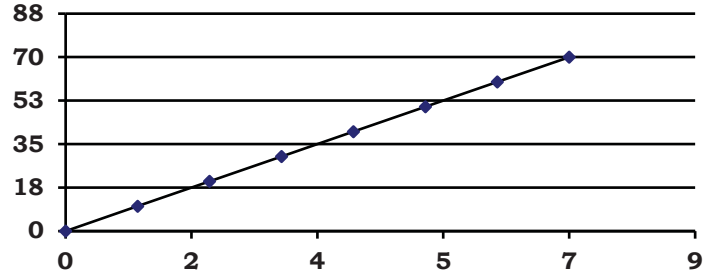


Figure 7.3. Functional dependence of pipe mass on pipe length

Points on the graph that correspond to each segment of the pipe in the figure will be on one line – the correlation is both functional and linear (Figure 7.3).

Let us take another population: a large group of men between aged 20-45 years. They have a common body type, being neither fat nor thin, nor tall nor short. Here, we find a corresponding point for each man in the group on the axes "height-weight" (Figure 7.4).

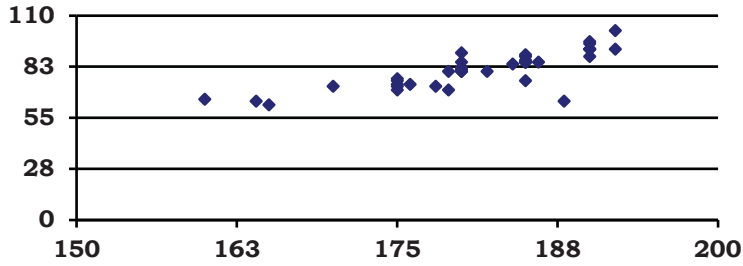


Figure 7.4. Correlation field of factors, "height-weight"

If the number of points on the line increases (Figure 7.3), the straight line becomes more vivid. With the increase in the male population (Figure 7.4), what is known as a *correlation* field appears in the figure, elongated and looking like an ellipse. It is obvious that a certain value of the feature

"height" (factor) at the value of 180 cm matches ***a set*** of values of a result (effect) "weight". These points are marked in Figure 7.4. It can be seen that all the men with the same height have different weights, ranging from 65 kg to 105 kg, or 90 $\pm$ 15 kg. We can practically talk of ***an average value*** for the weight of the sample population. Here, we have a ***conditional*** distribution of the result (effect) in "weight". This is characterized by a set of parameters, as in any other distribution row of a quantitative feature. We have already determined some of them visually and can calculate the others. Let us assume that $\overline{x}$ = 80 kg, s = 5 kg.

It is interesting to note that, if the population of men is rather large, their distribution by weight is closer to a normal distribution, which is very common in nature for mass events. We can find many examples from biology of this norm, but it is not applicable where we are considering a pathology. For example, in a population of humans, people are normally distributed by height, weight, blood pressure, and lung volume etc. In comparison, a normal distribution occurs quite rarely in socioeconomic events. It is important to remember that the form of distribution influences the choice of methods for statistical analysis, in particular, when it concerns hypothesis checking and studying connections.

There is a statistical, correlation, and direct connection between the features "height-weight". When the value of a result (effect) "height" increases, an average ***probable*** value of the feature "weight" also increases. Given a concrete value for a factor, we can determine a ***probable*** value of the result. If a correlation field is rather elongated, it can be simulated in the form of a certain function. In our example, it is linear in form (a regression equation), $= \alpha + bx$, where y is theoretical value of a result (effect). Let us assume a correlation field in considering the interconnection of the features "height-floor". It is most likely something similar to what is shown in Figure.7.5. It is not difficult to come to a conclusion: if there is no connection between features, a correlation field has no definite form (Figure 7.5). With the increasing connection of the relationship, some points move closer to that imaginary line: the regression line (Figure 7.4).

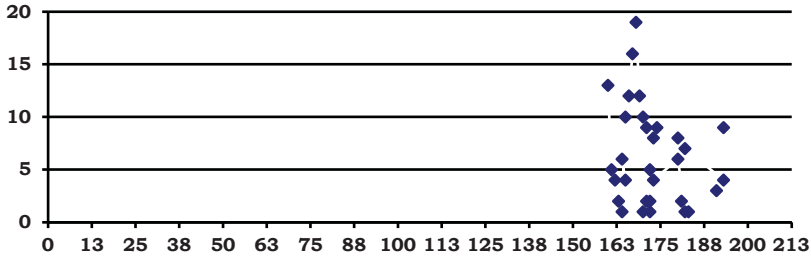


Figure 7.5. Correlation field of example, "height-floor"

A correlation-regression analysis involves the following stages:

- choosing a regression form;
- determining the parameters of the regression equation;
- estimating the connection proximity;
- and checking the significance of the connection.

Figures, analytical groupings, and theoretical substantiveness are all important in choosing functions. It is possible to sort the functions when calculating regression equations of different types and choosing the best for the purpose. The linear function is the most widely used in statistical analysis $y_x = \alpha + bx$. Parameter b is *the coefficient of regression*. It indicates how much the result will change if the factor increases by one unit. The parameter a is a free member of the equation; it may not have a proper interpretation as this is the value of y at $x = 0$. If x cannot take a zero value, then a is not interpreted; as a free member of the regression equation, a has only a secondary value for the calculation of the theoretical values of the result (effect).

In this example, we suggest to our reader, from our own personal experience, the use of some real parameter values to compare the dependence of a person's mass on height. Sometimes, studying the sense of an event requires the use of non-linear regression equations. In such a case, we more often use a degree function, $y_x = a \times x^b$; a semi-logarithmic function, $y = \alpha + b \times logx$; or a hyperbola, $y_x = \alpha + b\frac{1}{x}$. The parameters of the regression equation are determined by the least squares method, the main condition of which is minimization to the sum of squares of the deviation of empirical values from theoretical ones. This makes it possible to achieve the best estimates of parameters a and b: $\sum(y_i - Y_i)^2 \Rightarrow min$.

To compute this, we solve a system of normal equations $\begin{bmatrix} n\alpha+b\sum x=\sum y \\ \alpha\sum x+b\sum x^2=\sum xy \end{bmatrix}$.

To solve the system, a method of determinants is used:

$$\alpha = \frac{\sum y \sum x^2 - \sum xy \sum x}{n\sum x^2 - (\sum x)^2},$$

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}.$$

Using the regression coefficient b, we can determine the coefficient $\beta = b \times \frac{\bar{x}}{\bar{y}}$ of elasticity, which reflects the percentage change in value of a result (effect) when a factor changes by 1 %. To check the significance of a regression coefficient, Student's t-criterion is used, determined with the formula $t_{\alpha_1} = \frac{\alpha_1}{\mu_{\alpha_1}}$, $\mu_{\alpha_1} = \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_x^2 (n-2)}}$, where σ_ε^2 is the residual variance and σ_x^2 is the variance of a factor.

Determination of the proximity of connection in LSM, follows the rule of the sum of variances; however, if the estimates of the regression line in the first method are the average group values of the result (effect), in LSM these are the theoretical values of the latter. The variance of theoretical values is described as *factorial* and is calculated with the formula $\sigma_Y^2 = \frac{\sum(Y_i - \bar{y})^2}{n}$. This characterizes the variation of a result (effect) connected with the variation of a factor. A residual variance and a random variance are calculated instead of the average from the group variances $\sigma_\varepsilon^2 = \frac{\sum(y_i - Y_i)^2}{n}$. When the variation of a result (effect) is not connected to the variation of a factor, then a general variance is calculated with the formula $\sigma_y^2 = \sigma_Y^2 + \sigma_\varepsilon^2$; $\sigma_y^2 = \frac{\sum(y_i - \bar{y})^2}{n}$, where y_i is the real value of a result (effect); Y_i is the theoretical value of a result (effect); and n is the size of a population. The measure of proximity of connection in LSM is the coefficient of determination analogous to the correlation ratio $R^2 = \frac{\sigma_Y^2}{\sigma_y^2}$. This takes a value between 0 (no connection) and 1 (a functional connection between features).

The strength of connection is also characterized by a correlation index, $R = \sqrt{\frac{\sigma_Y^2}{\sigma_y^2}}$.

If the connection between factors is linear, a linear coefficient of correlation is used. It takes a value between -1 to +1 and indicates not only connection proximity, but also its direction. Its absolute value coincides with the index of correlation and is calculated with the formula

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum y^2 - (\sum y)^2] \times [n\sum x^2 - (\sum x)^2]}} \text{ or } r = \frac{\overline{x \times y} - \bar{x} \times \bar{y}}{\sigma_x \times \sigma_y}.$$

The significance of a connection in LSM is checked with the Fisher *F*-criterion, $F_R = \frac{R^2}{1-R^2} \times \frac{n-m}{m-1}$, where *m* is the parameter of the regression equation. If the real value of the Fisher *F*-criterion goes beyond a theoretical one, with a significance level of *a* and degrees of freedom $k_1 = m-1$ and $k_2 = n-m$ (m is a parameter of the regression equation and n is the general size of the population), then the adequacy of the suggested model is not random and the connection between the factor and the result (effect) is considered to be significant, with a probability of 1–α.

The relationship between the production cost of an output unit and the volume of production can be approximately expressed by an equation of hyperbolic regression $= \alpha + b\frac{1}{x}$. It differs from linear regression – instead of value *x* we have the value *1/x*.

The system of normal equations is $\begin{bmatrix} n\alpha + b\sum\frac{1}{x} = \sum y \\ \alpha\sum\frac{1}{x} + b\sum\frac{1}{x^2} = \sum\frac{y}{x} \end{bmatrix}$.

To solve the system, a method of determinants is used

$$a = \frac{\sum y \sum\frac{1}{x^2} - \sum\frac{y}{x}\sum\frac{1}{x}}{n\sum\frac{1}{x^2} - \left(\sum\frac{1}{x}\right)^2}; \; b = \frac{n\sum\frac{y}{x} - \sum\frac{1}{x}\sum y}{n\sum\frac{1}{x^2} - \left(\sum\frac{1}{x}\right)^2}.$$

To calculate the parameters of a regression equation that has the form of a degree function, it is imperative to give that function a linear form with the logarithms $\log Y = \log a + b \times \log x$. This equation differs from the equation for a common linear regression using logarithms of *y* and *x*.

Example 7.2

Determine the availability and nature of a statistical connection between the features “age of equipment” and “repair expenses” using LSM . Take the data and intermediate calculations from Table 7.5.

$$\bar{x} = \frac{70}{10} = 7 \text{ years; } \bar{y} = \frac{27}{10} = 2.7 \text{ thousand euro.}$$

Using the data in the table, we can calculate the parameters of the regression equation:

a = (27 × 536 - 217.1 × 70) / (10 × 536 – 70 × 70) = -1.576.

b = (10 × 217.1 – 70 × 27) / (10 × 536 -70 × 70) = 0.611 thousand euro.

Hence, we can see that there is a direct connection between the age of equipment and the expense of repair. The equation of linear regression: $Y = -1.576 + 0.611 \times x$.

Table 7.5. Equipment age and expense of repair for a group of companies

№	Equipment age, years (x)	Repair expenses, thousand euro (y)	x^2	xy	Y	$(y_i - Y_i)^2$	$(y_i - Y_i)^2$
A	1	2	3	4	5	6	7
1	4	1.5	16	6.0	0.868	0.399	1.44
2	5	2.0	25	10.0	1.479	0.271	0.49
3	5	1.4	25	7.0	1.479	0.006	1.69
4	6	2.3	36	13.8	2.090	0.044	0.16
5	8	2.7	64	21.6	3.312	0.374	0.00
6	10	4.0	100	40.0	3.312	0.285	1.69
7	8	2.3	64	18.4	4.534	1.024	0.16
8	7	2.5	49	17.5	2.700	0.040	0.04
9	11	6.6	121	72.6	5.145	2.117	15.21
10	6	1.7	36	10.2	2.090	0.152	1.00
Total	70	27	536	217.1	27.010	4.712	21.92

The coefficient of elasticity is $\beta = 0.611 \times \frac{7}{2.7} = 1.584$. With an increase in the age of equipment by 1 %, the expense of repair increases by 1.6 %. Let us calculate theoretical values for Y (column 5, Table 7.5), using the value of x in a regression equation. For the first value, Y_1 = - 1.576 + 0.611 x 4 = 0.868 thousand euro.

The residual variance is equal to $\sigma_\varepsilon^2 = \frac{4.712}{10} = 0.4712$.

The general variance is $\sigma_Y^2 = \frac{21.92}{10} = 2.192$.

Then, a factorial variance is calculated using the rule of the sum of variances $\sigma_Y^2 = 2.192 - 0.4712 = 1.7208$.

The determination coefficient is equal to $R^2 = \frac{1.7208}{2.192} = 0.785$ (alternately, 78.5 % of the general variation of repair expenses depends on variation in equipment age). We can calculate the correlation coefficient with the formula $R = \sqrt{\frac{\sigma_Y^2}{\sigma_y^2}} = \sqrt{0.785} = 0.886$.

We can calculate the linear correlation coefficient $r = \frac{\frac{217.1}{10} - 7 \times 2.7}{\sqrt{4.6 \times 2.192}} = 0.885$.

As we can see, the results are the same from both formulas. Using the scale of connection proximity (Table 7.4), the conclusion can be drawn that there is a rather close and direct connection between equipment age and expense of repair. To check the significance of the correlation coefficient, a special table of critical values is used. The value of *n* has two units fewer than the number of observations. In our example *n* = 10-2 = 8. A coefficient is significant if it exceeds the corresponding value in the table.

Let us check the significance of the correlation coefficient with the *F*-criterion:

$$F_R = \frac{0.785}{1 - 0.785} \times \frac{10-2}{2-1} = 29.2, \text{ where a} = 0.05\ F(1{:}8) = 5.32.$$

The result is smaller than the computed value, 29.2. As such, the computed correlation coefficient is significant, confirming the connection proximity between equipment age and expense of repair. Also, a table of critical values for the *t*-criterion can be used. The degrees of freedom depend on the number of parameters of the regression, equation *m*.

Now, let us check the significance of the regression coefficient. We calculate a real value for Student's criterion. For this purpose, we calculate the variance of equipment age with the formula

$\sigma_x^2 = \overline{x^2} - \bar{x}^2 = \frac{536}{10} - \left(\frac{70}{10}\right)^2 = 4.6$; $\mu_{\alpha_1} = \sqrt{\frac{0.4212}{4.6 \times (10-2)}} = 0.11$ and $t_{\alpha_1} = \frac{0.611}{0.11} = 5.55$.

For the table value of the *t*-criterion at a significance level of 0.05 for degrees of freedom 10-2 = 8, is equal to 1.64; this is much lower than the computed value. Hence, the regression coefficient is considered significant.

7.5. Multifactorial correlation

To describe the dependence of a result (effect) for several factors, a multifactor regression model is used: $\boldsymbol{Y = f(x_1, x_2, \ldots\ldots\ldots, x_n)}$. Because of the inherent difficulties of explaining the forms of connection, multifactor linear equations and equations leading to a linear form via transformations are used: $\boldsymbol{Y = \alpha + b_1 x + b_2 x^2 + b_3 x^3 + \cdots + b_n x^n}$.

The equation parameter b_i is *a partial coefficient of regression*, which shows how, on the average of n result (effect), *Y* changes when a factor, x_i, changes by one, provided the other factors remain unchanged.

It is possible to solve a regression equation using the least squares method:

$$\begin{cases} \text{na} + b_1 \sum x_1 + b_2 \sum x_2 + \ldots + b_n \sum x_n = \sum y \\ a \sum x_1 + b_1 \sum x_1^{\,2} + b_2 \sum x_1 x_2 + \ldots + b_n \sum x_1 x_n = \sum x_1 y \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a \sum x_1 + b_1 \sum x_1 x_n + b_2 \sum x_2 x_n + \ldots + b_n \sum x^2{}_n = \sum x_n y \end{cases}$$

To study the developmental mechanism of statistical regularities, we use the concept "factor" and examine the determination of the result by factors. These statistical methods enable the combination of different factors that have common features or traits and the peculiarities of their impact on the researched process into particular groups. The aim is to decrease the number of factors, as well as to eliminate the effect of multicollinearity of the factors, which distort the results of interconnection analysis.

The choice of factors is a very important step in building a regression model, which is closely connected to the choice of model. This is a constant and difficult problem. In addition to the in-depth understanding of the sense of an event being studied, the researcher is required to follow some formal presuppositions:

1) the availability of random samples from a general population;
2) a sufficient number of observations;
3) the independence of the observations;
4) a considerable excess of the number of population units over the number of factors (by 6–8 times);
5) uniformity in the population;
6) a quantitative level of variable estimates.

7.6. Nonparametric methods for studying interconnections between events

Against the background of different methods of studying statistical connections, it is important to understand the specificity and conditions of applications for such methods. In CRA, the factorial and the result (effect) belong to a metric scale; here, the methods of analytical grouping and variance analysis can be realized if a factor is qualitative and the result (effect) is quantitative. If both the factor and result (effect) are qualitative, then it belongs to a nominal or ordinal scale, which do not require the calculation of distribution parameters, and nonparametric methods are used. What is the sense of this principle?

We discussed earlier that a criterion may be used to compare the distribution rows of a qualitative feature. We present a further illustration here with data in Table 7.6.

Table 7.6. Relationship of attitude to working conditions in an enterprise to gender

Gender	Attitude to working conditions					Total
	Satisfied	Rather satisfied	Rather unsatisfied	Unsatisfied	Do not know	
Male	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_1
Female	f_{21}	f_{22}	f_{23}	f_{24}	f_{25}	f_2
Total	$F_{0.1}$	$F_{0.2}$	$F_{0.3}$	$F_{0.4}$	$F_{0.5}$	n

To calculate a horizontal structure (risk ratio), we need to divide the frequency of each sub-group by the general frequency of the corresponding group, $\overline{w} = \frac{f_{ij}}{f_{0.j}}$. This is presented in Table 7.7.

Table 7.7. Horizontal structure of the dependence of attitude to working conditions on gender

Gender	Attitude to working conditions					Total
	Quite satisfied	Rather satisfied	Rather unsatisfied	Quite unsatisfied	Do not know	
Male	$\overline{w}_{11} = \frac{f_{11}}{f_{0.1}}$	$\overline{w}_{12} = \frac{f_{12}}{f_{0.2}}$	$\overline{w}_{13} = \frac{f_{13}}{f_{0.3}}$	$\overline{w}_{14} = \frac{f_{14}}{f_{0.4}}$	$\overline{w}_{15} = \frac{f_{15}}{f_{0.5}}$	1.00
Female	$\overline{w}_{21} = \frac{f_{21}}{f_{0.1}}$	$\overline{w}_{22} = \frac{f_{22}}{f_{0.2}}$	$\overline{w}_{23} = \frac{f_{23}}{f_{0.3}}$	$\overline{w}_{24} = \frac{f_{24}}{f_{0.4}}$	$\overline{w}_{25} = \frac{f_{25}}{f_{0.5}}$	1.00
Total	$\overline{w}_{0.1} = \frac{f_{0.1}}{n}$	$\overline{w}_{0.2} = \frac{f_{0.2}}{n}$	$\overline{w}_{0.3} = \frac{f_{0.3}}{n}$	$\overline{w}_{0.4} = \frac{f_{0.4}}{n}$	$\overline{w}_{0.5} = \frac{f_{0.5}}{n}$	1.00

To calculate a vertical structure, the frequency of each sub-group in a group is divided by the frequency of that group with the formula $\overline{w}_{ij} = \frac{f_{ij}}{f_i}$. Then, we can create Table 7.8.

Table 7.8. Vertical structure of the dependence of attitude to working conditions on gender

Gender	Attitude to working conditions					Total
	Quite satisfied	Rather satisfied	Rather unsatisfied	Quite unsatisfied	Do not know	
Male	$\overline{w}_{11} = \frac{f_{11}}{f_1}$	$\overline{w}_{12} = \frac{f_{12}}{f_2}$	$\overline{w}_{13} = \frac{f_{13}}{f_3}$	$\overline{w}_{14} = \frac{f_{14}}{f_4}$	$\overline{w}_{15} = \frac{f_{15}}{f_5}$	$\overline{w}_1 = \frac{f_1}{n}$
Female	$\overline{w}_{21} = \frac{f_{21}}{f_1}$	$\overline{w}_{22} = \frac{f_{22}}{f_2}$	$\overline{w}_{23} = \frac{f_{23}}{f_3}$	$\overline{w}_{24} = \frac{f_{24}}{f_4}$	$\overline{w}_{25} = \frac{f_{25}}{f_5}$	$\overline{w}_2 = \frac{f_2}{n}$
Total	1.00	1.00	1.00	1.00	1.00	1.00

Vertical and horizontal structures help us to study distributions by factors and results (effect). Two lines of the table, i.e. distributions, are compared by checking a hypothesis on ***uniformity***: do men and women show similar attitudes to working conditions? This question can be put in a different way: is there a connection between gender and attitude to working conditions?

Thus, we are dealing with a hypothesis on ***independence***. In fact, if the attitude of men and women to working conditions differs significantly, then we can infer a significant statistical connection between the features, "gender-attitude to working conditions". This is presented in Table 2.1. There are 4 lines for distribution by hair color and 3 lines for distribution by eye color. If the distribution by the hair color of people with blue eyes differs significantly from the distribution of the same feature for people with grey eyes, and from those with hazel eyes, then there is a non-random statistical connection between these features.

Hence, a criterion can be used to prove the ***presence of significance of a connection***. Using CRA, we can also determine the form (direction) and the proximity. In the case of qualitative factors, there is probably no sense in discussing the form of the model and also the direction. This can be determined visually from the table of interdependence (TI) and we can see there is a direct relationship between eye color and hair color (or vice versa).

Vertical and horizontal structures can be studied with the help of cross tabulation. To determine the correlation, mutual conjugation/contingency is used.

For non-square tables, the Chuprov coefficient $C = \sqrt{\frac{\chi^2}{n \times \sqrt{(m_1-1)\,(m_2-1)}}}$, where m_1 and m_2 are the number of lines and the column of the table and n is the number of population elements.

The Kramer coefficient is calculated with the formula $= \sqrt{\frac{\chi^2}{n(m-1)}}$, where $m = min(m_1, m_2)$.

The formula for Pearson's *chi*-square depends on the supporting information available. If the cross tabulation is presented with frequencies, then the criterion can be calculated with the formula

$$\chi^2 = n \times \left(\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\frac{f_{ij}^2}{f_i f_j} - 1\right),$$

where f_{ij} is the frequency of sub-group *j* in group *i*; f_i is the frequency of group *i*; f_j is the frequency of sub-group *j*; *i* is the number of a group with the first factor; *j* is the number of a group with the second factor; m_1 is the number of groups with the first factor; m_2 is the number of groups with the second factor; and n is population size.

If the cross tabulation is presented by factor (a horizontal structure), then the criterion can be calculated with the formula

$$\chi^2 = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\frac{\left(\bar{w}_{ij} - \bar{w}_{0.j}\right)^2}{\bar{w}_{0.j}},$$

where w_{ij} is the fraction of sub-group *j* in group *i* and $w_{0.j}$ is the fraction of group *j* in the whole population. In the case of independence, $w_{ij} = w_{0.j}$ and

$\chi^2 = 0$. The Chuprov and Kramer coefficients can take values from 0, when there is no correlation, up to 1, when there is a functional connection.

In practical statistical observations, there is quite often a necessity to analyze alternative distributions, where a population is divided by each factor into two groups with opposite characteristics. For instance, we can analyze the study progress of students by gender, dividing them into two groups in terms of who passed the examination and who did not.

Example 7.3

Table 7.9. Relationship of study progress of students by gender

Gender	Number of students		Total
	Passed	Failed	
Women	a=25	b=2	a+b=27
Men	c=20	d=3	c+d=23
Total	a+c=45	b+d=5	50

The correlation can be estimated by a coefficient of association (contingency)

$$A = \frac{ad-bc}{\sqrt{(a+c)(c+d)(a+c)(b+d)}} = \frac{25 \times 3 - 2 \times 20}{\sqrt{27 \times 23 \times 45 \times 5}} = 0.09.$$

Then, Pearson's criterion is calculated with the formula

$$\chi^2 = A^2 \times n = 0.09^2 \times 50 = 0.405.$$

Using the tables of critical values for a significance level of 0.05 and one degree of freedom, the *chi*-square value is 3.84, which is larger than the real value. As such, the null hypothesis on the absence of differences in the distributions of women and men by level of progress in study does not deviate and the correlation significance is classified as not-proven.

This conclusion is indeed true because the real factors in study progress are not gender, but rather level of attendance of lectures and practical lessons and the number of hours devoted to self-study etc. Using 2×2 tables, we can calculate an **odds ratio** with the formula $= \frac{a \times d}{b \times c}$.

The odds ratio for our example is equal to $W = \frac{25 \times 3}{20 \times 2} = 1.9$, implying that women were more successful in examinations than men by a multiple of 1.9. The odds ratio can be also calculated for tables using larger dimensions, by changing them into 2×2 tables. Two techniques can be used:

1) eliminate intermediate values for factors and results (effect);
2) combine the lines and (or) columns according to their logical meaning.

Example 7.4

Using these data, determine the connection proximity between attitudes to smoking and lung condition (men).

Table 7.10

Attitude to smoking / Test value	Do not smoke	Smoke	Stopped smoking	Total
Normal	4	32	8	44
Abnormal	64	83	46	193
Total	68	115	54	237

Let us calculate Pearson's criterion with the formula of frequencies

$$\chi^2 = 237 \times \left[\frac{4^2}{44 \times 68} + \frac{32^2}{44 \times 115} + \frac{8^2}{44 \times 54} + \frac{64^2}{193 \times 68} + \frac{83^2}{193 \times 115} + \frac{46^2}{193 \times 54} - 1\right] = 14.23.$$

Now, we can check the criterion for significance. The theoretical value of the criterion at a significance level of 0.05 and degrees of freedom, k = (2-1) × (3-1) = 2, is 5.99, which is lower than the real value. So, there is a connection.

To calculate *chi*-square with another formula, it is first necessary to calculate a horizontal structure. For this purpose, a share of those who

smoke, those who do not smoke, and those who have stopped smoking are determined from among those people whose lungs are in a normal condition (Table 7.11). For those who do not smoke, 4/44 = 0.091, and so on.

Let us check the significance of the coefficient *chi*-square from the table of critical values. First, we calculate the number of degrees of freedom, $k = 2$. The coefficient is significant if it exceeds the corresponding table value.

Table 7.11. Calculation of horizontal structure (%)

Test value / Attitude to smoking	Do not smoke	Smoke	Smoking stopped	Total
Normal	0.09	0.73	0.18	1.00
Abnormal	0.33	0.43	0.24	1.00
Total	0.29	0.48	0.23	1.00

Now, we calculate χ^2 with the formula

$$\chi^2 = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\frac{\left(\bar{w}_{ij}-\bar{w}_{0.j}\right)^2}{\bar{w}_{0.j}} = 44 \times \left[\frac{(0.09-0.29)^2}{0.29} + \frac{(0.73-0.48)^2}{0.48} + \frac{(0.18-0.23)^2}{0.23}\right] + 193\left[+\frac{(0.33-0.29)^2}{0.29} + \frac{(0.43-0.48)^2}{0.48} + \frac{(0.24-0.23)^2}{0.23}\right] = 12.276 + 2.154 = 14.43.$$

At $\alpha = 0.05$, χ^2 (2) = 5.99. This is also lower than the real value of 14.43. As such, the differences between the groups are not random and the connection between attitude to smoking and condition of lungs is significant with a likelihood of 95 out of 100 cases. Next, the correlation needs to be measured.

We put the calculated value of χ^2 into the formula for the Kramer coefficient $V = \sqrt{\frac{\chi^2}{n(m-1)}} = \sqrt{\frac{14.43}{237}} = 0.247$,

and into the formula for the Chuprov coefficient

$$C = \sqrt{\frac{\chi^2}{n \times \sqrt{(m_1 - 1)\,(m_2 - 1)}}} = \sqrt{\frac{14.43}{237 \times \sqrt{(2-1)(3-1)}}} = 0.207.$$

We interpret that the correlation between the studied features is weak.

Let us calculate the **odds ratio** for this example by changing Table 7.10 into a 2×2 table, i.e. by removing the last column "Stopped smoking". Then, the **odds ratio** is $W = \frac{4 \times 83}{32 \times 64} = 0.162$.

For a more convenient interpretation, let us take the inverse value, i.e. $\frac{1}{0.162}$ = 6.2.

The probability of falling sick is 6 times higher among those who smoke than those who do not smoke.

When the interconnection of features of different natures is studied, the use of cross tabulation is common in a wide range of fields – economics, sociology, biology, and medicine. Compared to CRA, it is easier to justify the choice, meet the required conditions for application, and interpret the results. We have already discussed the factors for choosing the LSM method, although the corresponding assumptions are not always valid (interconnection of factors, normality of distribution, and correspondence of scales etc.).

The desire to overcome similar obstacles sometimes encourages researchers to use acrobatic mathematical tricks and *far-fetched* scientific methods. To solve practical tasks with statistical software packages, a specialist in economics and management has to know ***the statistical instrument and related applications***, so as to be able ***to interpret*** the program outputs. As such, we recommend using examples to relate theory to possible applications. This will help:

a) to connect the method used with the real-life applications;
b) apply the available techniques (using statistical programs) to compare the resulting values, their interpretation, and the use of terminology.

Parametric and nonparametric methods are not interchangeable, but, in some cases, for convenience, instead of one component you can use another; for example, replacing metric scales with ordinal ones. It is worth noting, though, that a shallow analysis can be justified by an insightful argument

and appropriate reliability tests. Connection proximity, however, and its significance can only ever be determined with nonparametric methods; LSM facilitates the study of form.

There are some well-known measures of interconnection not based on statistics. When a cross tabulation is built on the basis of a set of factors, one or more of them can be measured on an ordinal scale, for example "eye color-hair color". The most common methods of range correlation are the measuring procedures of Kendall, Stewart, and Spearman. If the features in such a cross tabulation are only attributive (e.g. "gender", "specialty"), the Goodman-Kruskal method is recommended. Finally, there is a group of methods particularly used with 2×2 cross tabulation in the form of a four-cell table of interdependence. A common example of this would be "smoke-don't smoke; fall sick-don't fall sick". Obviously, using these as examples, one needs to ask questions of such distributions: "what does 'fall sick' mean?"; "how often?"; "what are the diseases at issue?"; "what does 'don't smoke' mean?"; "does this refer to never smoking or only smoking occasionally?" etc.

In general, simple answers are not always conclusive. In fact, there are also alternative features, for example, "gender" (at least in its currently accepted form). However, if a population size allows, we can try to "extend" a measurement scale. For instance, if we formulate a question on a survey for the sociological observation of company workers like "are you satisfied with your working conditions?", the range of answers may be: dissatisfied; more dissatisfied than satisfied; difficult to answer; more satisfied than dissatisfied; satisfied. If you offer two options so that the interviewee can only respond with "satisfied" or "unsatisfied", you cannot gauge the true picture of people's perceptions. However, sometimes it is useful to act to the contrary.

Theoretically, we can practically avoid the use of a coefficient of mutual conjugation in those cases where the values of some cells of cross tabulation are smaller than 5. The means to overcome such a situation can be found in the approaches of Yates, Cochran, and Mantel. However, in those cases it is advisable to combine the lines or columns in the table of interdependence. Clearly, this can be done, for instance, for features such as "satisfaction with working conditions" or "hair color", but would not be appropriate for a feature like "specialty". An insightful analysis does not just mean checking the hypothesis of independence, but also comparing the inherent criteria to gain a more comprehensive understanding of inference from the results.

7.7. Range correlation

The measurement of a relationship with correlation-regression analysis and variance analyses brings associated complexity and requires extensive calculation. A screening analysis that seeks to get an indicative estimate of connection proximity includes:

1) the Fechner correlation coefficient of signs;
2) Spearman's rank and the Kendall rank correlation coefficients.

A Fechner correlation coefficient of signs is derived from a comparison of the signs of deviation from the average and the calculation of the number of coincidences and non-coincidences of the signs. T is determined with the formula $= \frac{u-v}{u+v}$, where u is the number of pairs with similar deviation signs x and y from $\bar{x}$; and $\bar{y}$ and v are the number of pairs with different deviation signs x and y from $\bar{x}$ and $\bar{y}$. A correlation coefficient of signs ranges from -1 to +1. The closer to 1, by module, the closer the connection. The sign +/– refers to the direction of a connection. If $u = v$, then $i = 0$ and the connection is absent.

Example 7.5

We can study the connection between the cost of fixed assets and the volume of production from calculation of the Fechner correlation coefficient of signs (Table 7.12).

$\bar{x}$ = 108/10=10.8 euro

$\bar{y}$ = 47.2/10=4.72 euro. So, u = 9, v = 1. Then, $i = \frac{u-v}{u+v} = \frac{9-1}{9+1} = +0.8$.

Table 7.12. Cost of fixed assets and production volume (million euro)

№	Cost of fixed assets (*x*)	Production volume (*y*)	Deviation sign	
			$x - \bar{x}$	$y - \bar{y}$
A	1	2	3	4
1	6	2.4	-	-
2	8	4.0	-	-
3	9	3.6	-	-
4	10	4.0	-	-
5	10	4.5	-	-
6	11	4.6	+	-
7	12	5.6	+	+
8	13	6.5	+	+
9	14	7.0	+	+
10	15	5.0	+	+
Total	108	47.2	x	x

We can see that the connection between the cost of fixed assets and the volume of production is direct and rather close/tight. Let us consider one more method to estimate a correlation from the calculation of a correlation coefficient range. The distinctiveness of this method is that the results are not calculated from the preliminary data, but rather from ranges that are given for all values of the features under study ranked in ascending order. If the values coincide, the ranges are determined by division of the sum of ranges by the number of values. Spearman's rank correlation coefficient is determined with the formula $\rho = 1 - \frac{6 \sum d^2}{n\,(n^2-1)}$, where d^2 is the square of the range difference for each unit of $d = x - y$. Spearman's rank correlation coefficient also covers the range -1 and +1. The closer to 1 by module, the

closer the connection is. The sign +/– refers to the direction of a connection. If the ranges coincide by two factors, the connection is both complete and direct. If ρ = 0, there is no connection between the factors.

Now we can calculate (Table 7.13) the coefficient of a correlation range from the data in the previous example.

Table 7.13. Calculation of Spearman's rank correlation coefficient

Ranges	*R* by *x*	*R* by *y*	Difference of ranges, *d*	Square of range difference, d^2
A	1	2	3	4
1	1	1	0	0
2	2	3.5	-1.5	2.25
3	3	2	+1	1
4	4.5	3.5	+1	1
5	4.5	5	-0.5	0.25
6	6	6	0	0
7	7	8	-1	1
8	8	9	-1	1
9	9	10	-1	1
10	10	7	+3	9
Total	x	x	x	16.5

Ranges in the cost of fixed assets for the fourth and fifth enterprises were determined as the arithmetic mean = (4+5)/2 = 4.5. Similarly, the volume of production for the second and fourth enterprises were estimated. The data were put into the formula

$$\rho = 1 - \frac{6 \sum d^2}{n\,(n^2-1)} = 1 - \frac{6 \times 16.5}{10 \times (100-1)} = 1 - 0.1 = 0.9.$$

This confirms that the connection between the cost of fixed assets and the volume of output is both direct and close. We can check the coefficient of the correlation rank for significance. In Table 2.7, supplement 2, for a significance level of 0.05 and 10 units of a population we find the theoretical value of the coefficient, which is equal to $\rho_{0.05}(10) = 0.54$. This is smaller than the real value. As such, the connection between the cost of fixed assets and the volume of production is significant.

A rank coefficient of correlation is more accurate compared to a correlation coefficient of signs, as it calculates both the deviation of signs and the position of the value of a factor in a given row. In addition to the above described coefficients, in practice, the Kendall rank correlation coefficient is used to determine a rating and connection proximity $\tau = \frac{2\sqrt{S_i}}{n\,(n-1)}$, where S_i is the sum of points.

The basis of this method lies in the calculation of the points for each unit of a population. We compare a range for the first population unit by effect, *y*, in a row, arranged by factor x with the rest of the units of a population, placed close together in a list. If it is smaller, we put +1 and if it is bigger, -1. Let us consider the example in Table 7.14.

Table 7.14. Calculation of the number of points

№	R_x	R_y	S_{ji}								
			S_{1i}	S_{2i}	S_{3i}	S_{4i}	S_{5i}	S_{6i}	S_{7i}	S_{8i}	S_{9i}
1	1	1									
2	2	3	+1								
3	3	2	+1	-1							
4	4	4	+1	+1	+1						
5	5	5	+1	+1	+1	+1					
6	6	6	+1	+1	+1	+1	+1				
7	7	8	+1	+1	+1	+1	+1	+1			
8	8	9	+1	+1	+1	+1	+1	+1	+1		
9	9	10	+1	+1	+1	+1	+1	+1	+1	+1	
10	10	7	+1	+1	+1	+1	+1	+1	-1	-1	-1
Total	x	x	+9	+6	+7	+6	+5	+3	+1	0	-1

$$\tau = \frac{2\sqrt{S_i}}{n\,(n-1)} = \frac{2 \times (9+6+7+5+3+1+0-1)}{10 \times (10-1)} = 0.67.$$

The computed coefficient confirms the presence of quite a close and direct correlation between the cost of fixed assets and the volume of production. The critical value of the Kendall rank coefficient for the level of significance $\alpha = 0.05$ at $n = 10$ is equal to 0.467. The real value is larger than the critical one establishing the previously made conclusion on the significance of the correlation between these factors.

Practice Exercises

Exercise 7.1

Data below from a sociological interview (poll) of 10 students is arranged by two features: “active participation in class” and “grade”.

Student №	1	2	3	4	5	6	7	8	9	10
Active participation	7	3	4	1	8	2	6	10	5	9
Grade	5	3	4	2	9	1	7	10	6	8

Estimate the correlation between the factors. Check for significance at $\alpha = 0.05$. Make some conclusions.

Exercise 7.2

From the data, determine the presence of a **correlation** between the factors “monthly wage–worker’s age”.

Monthly wage, euro	Number of workers by age group, years			Total
	20–35	35–50	over 50	
200–400	1	2	15	18
400–600	1	40	15	56
600–800	48	247	53	348
800–1000	22	54	2	78
Total	72	343	85	500

Check for significance at $\alpha = 0.05$. Make some conclusions.

Exercise 7.3

Identify the presence of a **correlation** between two factors.

B A	B1	B2	B3	Total
A1	10	5	65	80
A2	90	15	15	120
Total	100	20	80	200

Check for significance at $\alpha = 0.05$. Make some conclusions.

Exercise 7.4

In the data below, determine the relationship between two factors.

Marital status	Availability of an apartment		Total
	Have	Do not have	
Married	371	55	426
Single	44	30	74
Total	415	85	500

Check for significance at $\alpha = 0.05$. Make some conclusions.

Exercise 7.5

From the data, discover the relationship between specialty and worker's daily wage (euro) using analytical grouping.

№	Specialty	Daily wage	№	Specialty	Daily wage
1	Turner	210.42	11	Turner	90.66
2	Fitter	99.54	12	Miller	96.84
3	Miller	110.05	13	Miller	110.24
4	Miller	98.12	14	Fitter	213.53
5	Fitter	98.45	15	Miller	109.83
6	Miller	117.54	16	Turner	111.33
7	Turner	212.25	17	Miller	128.74
8	Fitter	211.22	18	Fitter	210.34
9	Miller	99.32	19	Miller	97.96
10	Turner	214.15	20	Fitter	129.77

Check for significance at $\alpha = 0.05$. Make some conclusions.

Exercise 7.6

Below, we have some data about the average grade point of 10 students from an entrance examination and the first examination session.

№	Average grade point		№	Average grade point	
	Entrance exams	Examination session		Entrance exams	Examination session
1	4.8	4.7	6	3.3	4.1
2	4.4	4.2	7	4.0	3.7
3	4.2	4.4	8	3.9	3.0
4	5.0	5.0	9	4.7	4.3
5	4.5	4.9	10	3.7	3.2

Determine the correlation between average grade points, using the Spearman's rank and Kendall rank correlation coefficients. Check for significance at $\alpha = 0.05$. Make some conclusions.

Exercise 7.7

From the data, determine whether there is a relationship between smoking and having a lung condition.

Test value	Attitude to smoking			Total
	Do not smoke	Smoke	Smoked, gave up	
Within the norm/standard	4	32	8	44
Higher than the norm	64	83	46	193
Total	68	115	54	237

Check for significance at $\alpha = 0.05$. Make some conclusions.

Exercise 7.8

Out of a group of 112 patients on an intensive care ward, 77 were in shock and 37 of them died. 5 types of shock are known. From the data below, determine whether the chance of survival depends on the presence of shock and its type. Make some conclusions.

Type of shock	Chance of survival	
	Survived	Did not survive
Hypotonic	7	8
Cardiologic	11	11
Neurologic	10	6
Septic	9	7
Endocrinological	3	5
No shock	32	3
Total	72	40

Exercise 7.9

Calculate a correlation to characterize the relationship proximity between the level of mechanization (percentage) for harvesting and production cost of 1 ton of sugar beets. Use the following data.

Mechanization level, %	Number of enterprises	Production cost of 1 ton of sugar beets, euro
40–60	10	36.0
60–80	25	30.0
60–100	15	26.0
Total	50	x

The general variance of the sugar beets is 20. Check the significance of connection at $\alpha = 0.05$. Make some conclusions.

Exercise 7.10

Calculate a correlation to characterize the relationship between the output rate percentage and worker qualifications. Use the following data.

Qualification	Number of workers	Output rate, %	Intergroup variance
Non-qualified	15	100	6
Qualified	50	105	4
Highly qualified	35	110	6
Total	100	x	x

Check for significance at $\alpha = 0.05$. Make some conclusions.

Exercise 7.11

Below are some data characterizing the relationship between disposition to risk and income level.

Income level	Disposition to risk		Total
	Avoids risks	Takes risks	
High	14	9	23
Medium	37	27	64
Low	100	28	128
Very low	151	30	181
Total	302	94	396

Estimate the correlation and check its significance with *chi*-square at a significance level of $\alpha = 0.05$. Calculate the **odds ratio** and make some conclusions.

Exercise 7.12

Below are some data characterizing the relationship between investment goals and level of income.

Income level	Goal of investment			Total
	To achieve profit in the form of dividends	To protect money from inflation	Calculation of rise in stock	
High	20	15	18	53
Medium	90	43	50	183
Low	165	80	67	312
Total	275	138	135	548

Estimate the correlation and check its significance with *chi*-square at a significance level of $\alpha = 0.05$. Calculate the **odds ratio** and make some conclusions.

Exercise 7.13

From the data, estimate the correlation between smoking and the condition of the lungs. Check its significance with *chi*-square at a significance level of $\alpha = 0.05$.

Testing value	Attitude to smoking			Total
	Do not smoke	Smoke	Gave up smoking	
Normal	4	32	8	44
Abnormal	64	83	46	193
Total	68	115	54	237

Calculate the **odds ratio** and make some conclusions.

Exercise 7.14

Below are some data characterizing the dependence between preference for taking risks and gender.

Preference for taking risks	Gender		Total
	Male	Female	
Avoids taking risks	110	190	300
Likes to take risks	60	40	100
Total	170	230	400

Estimate the connection proximity and check its significance with *chi*-square at a significance level of $\alpha = 0.05$. Calculate the **odds ratio** and make some conclusions.

Exercise 7.15

Below are some data characterizing the dependence of age and desire to take part in certain activities related to securities (financial assets).

Desire to participate	Age group			Total
	16–29	30–59	over 60	
Yes, I have a desire	130	200	10	340
No, I have no desire	170	400	160	730
Total	300	600	170	1070

Estimate the connection proximity and check its significance with *chi*-square at a significance level of $\alpha = 0.05$. Calculate the **odds ratio** and make some conclusions.

Exercise 7.16

Below are some data characterizing the dependence of investment type by age.

Investment type	Age group			Total
	16–29	30–59	over 60	
Securities	38	132	20	190
Currency	35	48	12	95
Real estate	27	57	15	99
Total	100	237	47	384

Estimate the closeness of the connection and check its significance with *chi*-square at a significance level of $\alpha = 0.05$. Calculate the **odds ratio** and make some conclusions.

Exercise 7.17

Below are some data on the distribution of 100 employee families by the level of education of the husband and wife.

Husband's education	Wife's education			
	Secondary	Secondary special	University	Total
Secondary	15	10	5	30
Secondary special	8	26	10	44
University	5	8	13	26
Total	28	44	28	100

Estimate the correlation between the level of education of the husband and wife with a coefficient of interdependence. Check the significance of the connection at $\alpha = 0.05$. Calculate the **odds ratio** and make some conclusions.

Exercise 7.18

Below we have some data on the distribution of income dynamics in a sampled population by level of income.

Income level	Income dynamics			Total
	Notably improved	Did not change much	Notably worsened	
High	27	30	3	60
Medium	45	150	15	210
Low	10	185	175	370
Total	82	365	193	640

Estimate the correlation between the level of income and the income dynamics with a coefficient of interdependence. Check the significance of the connection at $\alpha = 0.05$. Make some conclusions.

Exercise 7.19

From the data below on wage (euro) and the level of worker qualifications calculate the correlation and make a conclusion about the presence of a connection between these features. Check the significance level of the connection at $\alpha = 0.05$. Make some conclusions.

Wage	Qualification	Wage	Qualification
450	3	450	3
460	3	500	4
460	3	490	4
470	3	550	5
480	4	600	5
450	3	450	3
460	3	500	4
470	3	600	5

Exercise 7.20

Below are some data about the time spent by students on homework per week and data on the study progress of 10 students (results of an examination session).

№	Time spent, min.	Average grade point for the session	№	Time spent, min.	Average grade point for the session
1	120	4.7	6	100	4.1
2	110	4.2	7	90	3.7
3	120	4.4	8	80	3.0
4	130	5.0	9	110	4.3
5	140	4.9	10	75	3.2

Determine the connection proximity from the given data using the Fechner correlation coefficient of signs. Make some conclusions.

Exercise 7.21

From the data in Exercise 3.3 (chapter 3), build a cross tabulation. Estimate the correlation between two factors, "family size-housing area supply" with a coefficient of interdependence. Check the significance level of the connection at $\alpha = 0.05$. Make conclusions.

Exercise 7.22

Using the tabulated data in Exercise 3.3 (chapter 3), construct an analytical grouping (first determine the factorial cause and results (effects)) and estimate the correlation between the given features using a rule of variance. Additionally, calculate the correlation η^2. Check for significance at $\alpha = 0.05$. Make some conclusions.

Exercise 7.23

Using the tabulated data in Exercise 3.3 (chapter 3), determine parameters of an equation of linear regression for two factors "family size-housing area supply" and give an economic interpretation. Using a linear coefficient of correlation and a coefficient of determination, estimate the correlation and check its significance by the *F*-criterion at a significance level of $\alpha = 0.05$. Make some conclusions.

Exercise 7.24

Calculate a correlation to characterize the relationship between level of income per capita and level of foodstuff expenses.

Income level	Number of families, thousands	Foodstuff expenses, euro	Intergroup variance
Low	45	110	64
Medium	25	200	225
High	30	250	144
Total	100	x	x

Check for significance at $\alpha = 0.05$. Make some conclusions.

Exercise 7.25

Calculate a correlation to characterize the relationship between level of income per capita and level of foodstuff expenses.

Income level	Number of families	Share of foodstuff expenses, %	Intergroup variance
Low	15	85	64
Medium	50	75	100
High	35	50	144
Total	100	x	x

Check for significance at $\alpha = 0.05$. Make some conclusions.

Exercise 7.26

Calculate a correlation to characterize the relationship between wage and worker qualifications.

Qualification	Number of workers	Wage, euro	Intergroup variance
Unqualified	15	150	15
Qualified	50	200	12
Highly qualified	35	250	10
Total	100	x	x

Check for significance at $\alpha = 0.05$. Make some conclusions.

Exercise 7.27

Calculate a correlation to characterize the relationship between turnover size and turnover expenses (in percentage of turnover).

Turnover, million euro	Number of stores	Turnover expenses, %
Up to 10	5	10.8
10–30	10	8.6
30–50	28	6.5
Over 50	7	4.0
Total	50	x

The general variance of turnover expenses is 5.5. Check for significance at $\alpha = 0.05$. Make some conclusions.

Exercise 7.28

Calculate a correlation to characterize the relationship between percentage of output rate and wage size.

Output rate, %	Number of workers	Wage, euro
Up to 100	20	1250
100–150	70	1700
Over 150	10	2100
Total	100	x

The general variance of the wages is 660. Check for significance at $\alpha = 0.05$. Make some conclusions.

Exercise 7.29

Calculate a correlation to characterize the relationship between level of education and remuneration.

Education level	Number of respondents	Remuneration, euro	Intergroup variance
University	30	250	140
Secondary technical	20	200	150
Secondary	50	150	60
Total	100	x	x

Check for significance at $\alpha = 0.05$. Make some conclusions.

Exercise 7.30

Below are some data to characterize the dependence of income level on type/nature of activity.

Income level	Nature of activity			Total
	Executives/ managers	Workers and employees	Business people	
High	14	24	12	50
Medium	27	103	20	150
Low	32	200	12	244
Total	73	327	44	444

Estimate the correlation and check its significance with *chi*-square at a significance level of $\alpha = 0.05$. Calculate the **odds ratio** and make some conclusions.

Exercise 7.31

In some data from an analytical grouping characterizing the relationship between turnover and shopping area, the intergroup variance of turnover was 150 and the residual one was 46. Calculate the percentage of the variation of turnover explained by the factor of shopping area. Estimate the correlation and check its significance at $\alpha = 0.05$ if the number of trading companies is 100 units and the number of groups by shopping area size is 8 units.

Exercise 7.32

In some data from an analytical grouping characterizing the relationship between turnover and expenses, the intergroup variance of turnover was 125 and the residual one was 36. Calculate the percentage of the variation of turnover explained by the factor of shopping area size. Estimate the connection proximity and check for significance at $\alpha = 0.05$ if the number of trading companies is 100 units and the number of groups by shopping area size is 7 units.

Exercise 7.33

In the results of a regression model with a direct dependence between size of turnover and turnover expenses, the factorial variance of turnover was 125 and the residual one was 36. Calculate the model's adequacy and estimate the correlation. Check for significance at $\alpha = 0.05$ with the F-criterion if the number of trading companies is 100 units.

Exercise 7.34

In the results of a regression model with a direct dependence between turnover and number of trading companies in the regions, the factorial variance of turnover was 186 and the residual one was 120. Calculate the model's adequacy, estimate the correlation, and check for significance at $\alpha = 0.05$ with the F-criterion if the number of trading companies is 100 units. Make some conclusions.

Exercise 7.35

In the results of a regression model on turnover and expenses, the factorial variance of turnover was 186 and the residual one was 169. Calculate the model's adequacy and estimate the correlation. Check for significance at $\alpha = 0.05$ with the F-criterion if the number of trade companies is 100 units. Make conclusions.

Exercise 7.36

In the results of a regression model with an exponential function on the dependence between turnover and expenses, the factorial variance of turnover was 230 and the residual one was 50. Calculate the model's adequacy and estimate the correlation. Check its significance at $\alpha = 0.05$ with the F-criterion if the number of trading companies is 120 units. Make some conclusions.

Exercise 7.37

Below are some data about the average grade point of 10 students for an entrance examination and the first examination session.

№	Average grade point		№	Average grade point	
	Entrance examination	Session		Entrance examination	Session
1	4.8	4.7	6	3.3	4.1
2	4.4	4.2	7	4.0	3.7
3	4.2	4.4	8	3.9	3.0
4	5.0	5.0	9	4.7	4.3
5	4.5	4.9	10	3.7	3.2

Determine the correlation between average grade points using the Fechner correlation coefficient of signs. Make some conclusions.

Exercise 7.38

Below are some data on the time spent on homework per week and data about the study progress of 10 students (examination results).

№	Time spent, min.	Average grade point at examination	№	Time spent, min.	Average grade point at examination
1	120	4.7	6	100	4.1
2	110	4.2	7	90	3.7
3	120	4.4	8	80	3.0
4	130	5.0	9	110	4.3
5	140	4.9	10	75	3.2

Determine the correlation between the given factors using the Kendall rank correlation coefficient. Make some conclusions.

Exercise 7.39

Using the data below on wages (euro) and work experience (in years), calculate and make some conclusions about the presence of a correlation between these factors. Check for significance at $\alpha = 0.05$. Make some conclusions.

Wage	Work experience	Wage	Work experience
450	5	450	5
460	5	500	10
460	5	490	10
470	5	550	10
480	5	600	15
450	5	450	5
460	5	500	10
470	5	600	10

8. TIME SERIES

8.1. Principles and practices of time series statistical analysis

When statistical research into the dynamics of indicators of socioeconomic development is undertaken, the following tasks are carried out:

- estimation of the intensity of changes in indicators of socioeconomic development over time;
- determination of the average values for the studied indicators characterizing socioeconomic development;
- identification of the regularities in changes of indicators of socioeconomic development over time;
- determination of the impact of different periods of processes of development on the emergence of corresponding periods of socioeconomic development of a country;
- predictions on socioeconomic development;
- measurement of factors that define indicative dynamics of socioeconomic development.

Time series refers to an ordered sequence of random values over time. Requirements for the creation of time series are:

1. Ensure the uniformity of the levels in the time series – accommodating the same population strata, the same industrial sector etc.
2. Observe the conditions of comparison of time series as regards the uniformity of measurement, the equality of time intervals, the periodicity of registration, and the uniformity of the objects etc.
3. Ensure regularity between time intervals and the intensity of the studied processes – smaller intervals are taken for more variable events and larger intervals for less variable events.
4. Meet the requirements of the ordering of levels over time. Time series analysis with missing levels is unacceptable. If such gaps are unavoidable, they are filled with calculated nominal values.

5. Complete a development phase (periodization). Any study of events over time is divided into uniform stages of development – uniform historic periods.

This procedure enables:

- comparison of the levels and time intervals of the time series;
- adherence to the principle of a uniform sequence and continuity of the time series over time;
- ordering of the time series;
- the consistency of the time series;
- periodization (division into periods) of an event's development.

Let us discuss these principles in more detail. To make a comparison using such criteria involves the comparison of such things as territory, objects, units of measurement, time of registration, prices, and methodology of calculation. A comparison by territory means that the data are recalculated for countries, republics, and regions that have changed their geographical boundaries. A comparison by objects means comparison of populations that do not change in terms of qualitative composition. A territory and an object comparison are done by closing the time series, i.e. either absolute levels are replaced with relative ones or recalculated into conditional absolute values.

Another principle is the sequence and continuity of time series over time. It is advisable and, in some cases, practical for the time series to sequentially cover the whole of the relevant period from beginning to end. A lack of data can considerably distort the general picture of the time series. In such situations, non-available data demand interpolation. To solve the issue of the size of temporary intervals between registration dates, it is recommended that the contents of the process be taken into account. The larger the variation in the row levels, the more often the measurements must be done. This means that intervals can be increased for stable processes.

Row levels should necessarily reflect the quality of the events. In particular, it is necessary to perform a typological grouping within each interval to which the time series is related. After uniform time periods are singled out, it is possible to define time levels. In other words, this principle guarantees a comparison by population structure. A standard structure should be used.

8.2. The concept and types of time series

A time series is a time sequence of statistical indicators. Any time series includes two mandatory elements: a time variable and a set of indicator levels. There are many definitions of a time series:

A time series is an ordered sequence of numbers that can characterize the change in an event over time. It is a sequence of two elements – row level (y_t) and the time to which it belongs (t, moment or interval).

A time series is a successive row of values of a socioeconomic event that changes over time.

A time series is a placement of values of a certain statistical indicator in chronological order.

Depending on the indicator, **there are rows of**:

- absolute values (e.g., dynamics of investment in fixed assets, scope of construction and installation work, Table 8.1);
- medium/average values (dynamics of stock exchange prices);
- relative values (rates of investment growth in fixed assets, Table 8.2); coefficient dynamics of fixed asset replacement, dynamics of branch and regional concentration of investments).

Table 8.1. Domestic investments in fixed assets and scope of construction and installation work over 7 years (in current prices, million euro)

Year	Investments in fixed assets	Scope of construction and installation work
1	23629	10162
2	32573	13751
3	37178	14799
4	51011	19894
5	75714	32126
6	93096	40031
7	125254	53937

Table 8.2. Domestic indices of physical volume of investments in fixed assets and construction and installation work over 7 years (percentage of the preceding year)

Year	Investments in fixed assets	Construction and installation work
1	114.4	109.1
2	120.8	116.7
3	108.9	101.7
4	131.3	126.2
5	128.0	117.2
6	101.9	93.4
7	119.0	109.9

The time feature rows are classified as:

- **moment time series** (cost of fixed assets, Table 8.3);
- **interval time series** (putting fixed assets into operation, Table 8.4).

Table 8.3. Dynamics of fixed assets in the national economy over 7 years

Year	In current prices at the end of the year, billion euro	in percentage of the preceding year
1	829	101.0
2	915	102.4
3	965	101.1
4	1026	103.3
5	1141	104.2
6	1276	103.7
7	1406	104.0

Table 8.4. Putting fixed assets into operation over 5 years (current prices at the end of the year, million euro)

Year	Putting fixed assets into operation
1	21774
2	29362
3	34547
4	44165
5	61468

In a ***moment*** time series, the levels fix the state of an event ***at certain moments of time*** (***definite dates***). In an ***interval*** time series, they fix an aggregate result ***at a certain interval of time***. For example, quarterly volumes of production fit into an interval row; the size of a country's population at the beginning of the year fits a moment row. If a time series characterizes a change in one indicator, it is called ***one-dimensional***; if two or more indicators are present, it is ***multidimensional***. This latter form comes in two types:

- **parallel** time series;
- time series of **interconnected indicators** (Table 8.5).

Table 8.5. Main indicators of domestic investment and construction activity over 7 years (current prices, million euro)

Year	Gross accumulation of fixed capital	New fixed assets put into operation	Investments in fixed capital	housing projects[1] put into operation
1	33427	21774	23629	5.6
2	40211	23726	32573	5.9
3	43289	35025	37178	6.1
4	55075	44165	51011	6.4
5	77820	61468	75714	7.6
6	96965	…	93096	7.8
7	129037	…	125254	8.6

[1]) million square meters by area.

Parallel time series characterize the dynamics of either a single indicator, concerning various objects, or different indicators, concerning one object. A connection between the indicators of a multidimensional time series can be either ***functional*** or ***correlative***.

The following methods are applied when multidimensional dynamic rows are analyzed:

- calculation of absolute, relative, and average indicators of the time series;
- fitting of the time series;
- analysis of fluctuations in the processes;
- analysis of the interconnections in the time series;
- periodization of the time series.

8.3. Calculation of the average levels of a time series

One of the summarizing characteristics of a time series is an ***average*** level. In an interval time series with equal intervals, the simple arithmetic mean is calculated $\bar{y} = \frac{\sum y_t}{n}$, where y_t is the value of a dynamic row and n is the number of row levels.

As an example, using the following data on quarterly volumes of production, we can determine the average production volume per quarter.

Quarter	Volume of production, thousand euro
I	200
II	210
III	250
IV	240
Total	900

This gives $\bar{y} = \frac{900}{4} = 225$ thousand euro

Hence, output equal to 225 thousand euro was produced on average per quarter. For an interval row with unequal intervals, the weighted arithmetic mean $\bar{y} = \frac{\sum y_i t_i}{\sum t_i}$ is calculated, where y_i is the levels of the time series that characterize a certain period of time and t_i is the duration of that time period. It should be noted that such rows rarely appear and, as such, we will not discuss the calculation of an average for these rows in detail.

In a moment row with equal intervals of time, the average level of a row is calculated as a chronological average $\bar{y} = \frac{\frac{y_1}{2} + y_2 + y_3 + \cdots + y_{n-1} + \frac{y_n}{2}}{n-1}$.

As an example, using the data below about the number of workers at the beginning of the quarter, we can determine the average quarterly number of workers.

Quarter	Worker numbers at the beginning of the quarter
I	200
II	210
III	250
IV	240
I of a following year	220
Total	x

$$\bar{y} = \frac{\frac{200}{2} + 210 + 250 + 240 + \frac{220}{2}}{5-1} = 228 \text{ number of workers.}$$

So, the average yearly number of workers is 228 people. If the time intervals between moments are different, we use the formula of the weighted arithmetic mean $\bar{y} = \frac{\sum y_i' t_i}{\sum t_i}$, with $y_i' = \frac{y_t + y_{t+1}}{2}$, where y_t is the average of the levels of different periods of time and t_i is the duration of those periods.

For example, using the data below about in-stock balance in a warehouse in March, we can determine the average in-stock balance per month.

Date	In-stock balance, thousand pieces	Average in-stock balance, thousand pieces (y_i')	Duration period, days (t_i)	$y_i' \times t_i$
01.03	200	x	x	x
10.03	210	(200+210)/2=205	10-1=9	205×9
15.03	250	(210+250)/2=230	15-10=5	230×5
25.03	240	(250+240)/2=245	25-15=10	245×10
01.04	220	(240+220)/2=230	32-25=7	230×7
Total	x	x	31	7055

We find that $\bar{y} = \frac{7055}{31}$ = 227.581 thousand pieces giving us an average in-stock balance per month of 227581 pieces.

8.4. Time series indicators

Analysis of the developmental features of socioeconomic events and processes is made with absolute, relative, and average indicators of dynamics. These include: absolute increase (absolute decrease); growth rate; rate of increase (decrease); absolute value of 1 % increase (decrease); average absolute increase; average growth rate; and average rate of increase (decrease).

The calculation is based on a comparison of the levels of a dynamic row. There are two systems of comparison, known as basic and chain. If an initial level of y_1 row, or the level that is prior to the y_0 time series, is taken as the basis, then the comparison base is called *constant* and the indicators of dynamics are called *basic*.

If the prior level is a base, it is called a "variable base" and the indicators of dynamics are described as "chain". Two systems of comparison are graphically presented in Figure 8.1.

Let us examine the methodology for calculating time series indicators in different systems for comparative purposes.

1. ***The absolute increase (decrease)*** [Δ_i] characterizes the absolute change of a comparable level from the base of the comparison.

BASIC	*CHAIN*
$\Delta_i^b = y_t - y_1$	$\Delta_i^c = y_t - y_{t-1}$

Interconnection: the sum of all absolute increases in the chain is equal to the final basic absolute increase $\sum \Delta_i^c = \Delta_K^b = y_n - y_1$.

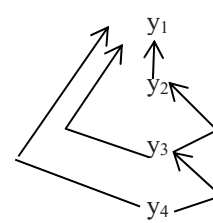


Figure 8.1. Comparative reflection when basic and chain comparison systems are used

The average absolute increase is given by $\bar{\Delta}=\frac{\sum \Delta_i^c}{n-1}=\frac{y_n-y_1}{n-1}$, where n is the number of row levels; y_n is the last level of the time series; y_1 is the first level of the time series; y_t is the current level of the time series; and y_{t-1} is the preceding level of the time series.

2. ***Growth (decreasing) rate*** [t] is characterized by how many times a comparable level of the time series is bigger than the comparison base. It is calculated as a coefficient or a percentage.

BASIC $t_i^b=\frac{y_t}{y_i}$ *CHAIN* $t_i^c=\frac{y_t}{y_{t-1}}$

Interconnection: the product of all chain growth rates is equal to the final basic growth rate $\prod t_i^c=t_K^b=\frac{y_n}{y_1}$.

Average growth rate $\bar{t}=\sqrt[n-1]{\prod t_i^b}=\sqrt[n-1]{\frac{y_n}{y_1}}$, where n is the number of row levels; y_n is the last level of the time series; y_1 is the first level of the dynamics row; y_t is the current level of the time series; and y_{t-1} is the preceding level of the time series.

3. ***The rate of increase (decrease)*** [T] is characterized by the percentage change in a value of a comparable level compared to the base of the comparison. It is calculated as the relation of an absolute increase to a prior level for a chain system, the first level for a basic system, or in reference to a growth rate.

BASIC	*CHAIN*
$T_i^b = \frac{\Delta_i^b}{y_i}$	$T_i^c = \frac{\Delta_i^c}{y_{t-1}}$

$$\text{or} = (t-1) \times 100\ \%.$$

An arithmetic operation cannot not take place with an increase in rates, as they cannot be added, subtracted, or multiplied. If increases belong to different events between which there is a connection, then a coefficient of elasticity can be calculated as a correlation between the increasing rate of an effective indicator and the increasing rate of a factorial indicator $K_{EL} = \frac{T_y}{T_x}$.

If a coefficient of elasticity by module is larger than 1, the result is elastic in a factor, otherwise it is non-elastic. For example, if a price rise of 10 % led to a decline in demand of 15 %, then the coefficient of demand elasticity in price is equal to $K_{EL} = \frac{-15\%}{10\%}$ = -1.5.

The inference is that a price rise of 1 % led to a decline in demand of 1.5 % and thus demand is elastic in terms of price. To find an average among growth rates at unequal time intervals, a weighted geometric average is used

$$\bar{t} = \sum t_i \sqrt{\prod \left(\frac{y_t}{y_{t-1}}\right)^{t_i}},$$

where t_i is the duration of the time intervals. To give an example, the average yearly growth rate of output volume for three years was 1.07; for the following two years it was -1.10. To give the average yearly growth rate of output volume for five years $\bar{t} = \sqrt[5]{1.07^3 \text{ x } 1.1^2}$ = 1.082.

4. ***The absolute value of a* 1 % *increase*** indicates how much of the absolute increase (decrease) is in the 1 %. It is calculated as a correlation between the absolute increase and rate of increase.

BASIC	*CHAIN*
X	$A\% = \frac{\Delta_i^c}{T_i^c} = \frac{y_{t-1}}{100}$

Values of A % are the same for basal rates of increase and are equal to the first level divided by 100. When the development intensity of the events, characterized by two dynamic rows, is compared, ***an advance coefficient*** is

calculated. This gives the relation of basic growth rates for two dynamic rows at equal time intervals $K_{ADV} = \frac{t_i'}{t_i''}$.

If the speed of the process in the study period is not the same, aa speedup or slowdown in growth is determined. If the time intervals are the same, the basic characteristics of the speed can be compared; if they are different, the average characteristics are compared.

Absolute speedup (slowdown) can be determined with the formula $\delta_i = \Delta_i - \Delta_{i-1}$.

The speedup of absolute growth is indicated by a positive value, d > 0; slowdown is indicated by a negative value, d < 0.

The coefficient of speedup (slowdown) of a relative dynamic speed is given by $K_i = \frac{t_i}{t_{i-1}}$.

We can illustrate this with examples for the calculation of these indicators.

Example 8.1

Below, we have some data about putting into use common areas of residential buildings in the country over 6 years (Table 8.6).

Table 8.6

Year	1	2	3	4	5	6
Area, million m^2	5.6	5.9	6.1	6.4	7.6	7.8

1. We determine the type of time series. In this case, it is an interval time series because its levels can be added and the sum of the levels represents the common residential area that was put into use in the period 2000-2005. This amounted to 39.4 million m^2.
2. We determine the average yearly area that was put into use. We have an interval time series with equal intervals. The average level is calculated with the formula of the simple arithmetic mean $\bar{y} = \frac{\sum y_i}{n} = \frac{39.4}{6}$ = 6.567 million m^2.
3. We calculate the indicators of dynamics.

Let us use this date to build a table (Table 8.7).

Table 8.7. Calculation of indicators of dynamics

Year	Area, million m^2	Absolute increase, million m^2		Growth rate		Rate of increase, %		Absolute value of 1% increase, thousand m^2
		B	C	B	C	B	C	
2000	5.6	0.0	x	1.000	x	0.0	x	x
2001	5.9	5.9-5.6=0.3	0.3	5.9/5.6=1.054	1.054	(1.054-1)×100%=5.4	5.4	(5.6/100)×1000=56
2002	6.1	0.5	0.2	1.089	1.034	8.9	3.4	59
2003	6.4	0.8	0.3	1.143	1.049	14.3	4.9	61
2004	7.6	2.0	1.2	1.357	1.188	35.7	18.8	64
2005	7.8	2.2	0.2	1.393	1.026	39.3	2.6	76
Total	39.4	x	2.2[1]	x	1.393[2]	x	39.3[3]	56[4]

m^2 = square meter; B = basic; C = chain.
[1]) 0.3+0.2+0.3+1.2+0.2 = 2.2;
[2])1.054×1.034×1.049×1.188×1.026 = 1.393;
[3]) (1.393-1) ×100 % = 39.3%;
[1]) (2.2/39.3) ×1000 = 56 thousand m^2.

Over a period of six years, the common area of residential buildings increased by 39.3 % equaling 2.2 million m^2. The highest level of dynamics were recorded for 2004 – the indicator value increased by 18.8 % and amounted to 1.2 million m^2.

We analyze the yearly dynamics. The average absolute increase is calculated from the final basal absolute increase $\bar{\Delta} = \frac{2.2}{6-1}$ 0.44 million m^2.

We determine the average yearly growth rate using the final basal growth rate $\bar{t} = \sqrt[6-1]{1.393}$ = 1.0685 or +6.85%.

This implies that, during the 6 year period, the total common residential area put into use in the country increased yearly by 6.85 % on average, amounting to 440 thousand m^2. A comparative analysis of chain growth rates substantiates the conclusion on the acceleration of yearly growth rates

of the total common residential area put into use up to the 5th year – the growth rate slowed down from the 6th year.

8.5. Statistical analysis of developmental tendencies

The empirical levels of a time series are changed by the impact of various factors. We can single out some of the most characteristic features, such as a developmental trend or seasonal fluctuations, and measure them quantitatively with the help of statistical methods.

Tendency (trend) refers to the developmental direction of any event or process. In some cases, a trend can be determined by the values of the row levels. Let us examine the graphical expression of the main trends.

A steady increase or decrease is shown in Figure 8.2.

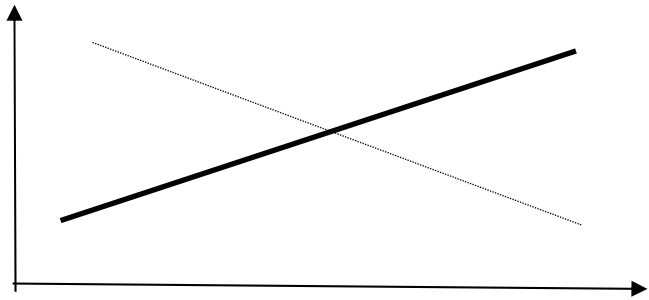


Figure 8.2. Diagram of steady increase (___), decrease (-----)

A slowing increase or decrease is shown in Figure 8.3.

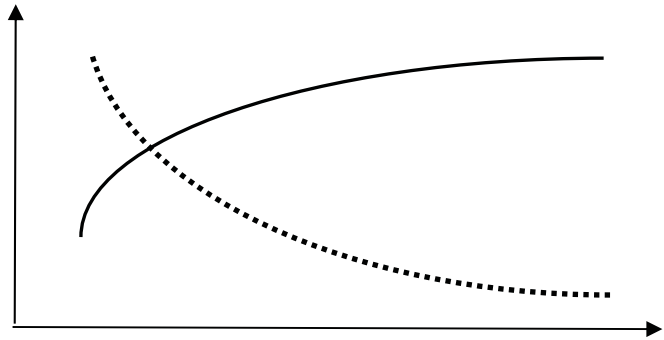


Figure 8.3. Diagram of slowing increase (_____), decrease (-----)

An accelerating increase or decrease is shown in Figure 8.4.

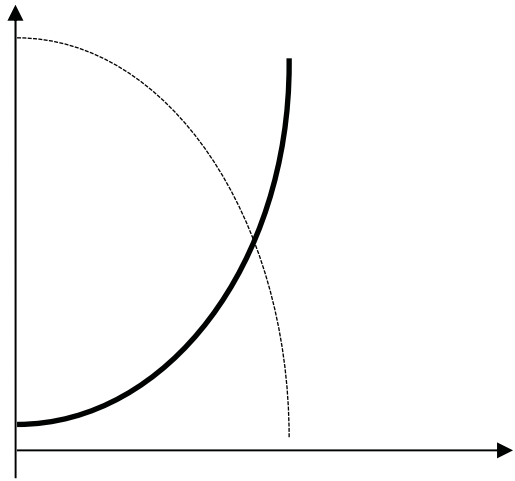


Figure 8.4. Diagram of accelerating increase (_____), decrease (-----)

When a clearly visible developmental trend, under the effect of random factors, is not identifiable, special statistical methods are used to determine (describe) this trend. The methods of the accelerated and moving mean belong to the simplest methods of row smoothing. The calculation of the ***accelerated mean*** is done using increased time intervals. Here, primary (empirical) levels are replaced with average levels. If periodic fluctuations are recorded in a dynamic row, then an increased interval should be equal to the fluctuation period. Such an interval "smoothens" random fluctuations, but does not show any change in the levels within an increased interval. A drawback of this method of smoothing is that the new row is shorter than the empirical one by m times (m is the size of an increased interval).

The essence of the method of the ***moving (sliding) mean*** is that the mean (the average) is calculated using increased intervals with a gradual shift down one level. The use of this method of smoothing enables leveling of dynamic row fluctuations; however, the drawback is that the smoothed row is shorter than the empirical one by m-1 levels. Besides, this only shows a trend, but does not allow it to be measured quantitatively.

Example 8.2

An example of a smoothed time series.

Table 8.8. Output of a company by months in thousand euro

Month	Output		
	Factual level	Smoothed by method of	
		moving mean	stepped mean
January	118	-	122
February	124	122	
March	124	125	
April	128	123	129
May	127	129	
June	132	132	
July	136	133	134
August	131	134	
September	135	136	
October	141	138	142
November	139	142	
December	146	-	

We smooth a row using the method of the moving mean. The mean for the first row levels is calculated as $y_1 = \frac{118+124+124}{3} = 122$ thousand euro.

Let us refer to February – leaving out the first level (January) and adding the fourth level (April), we calculate the mean $y_2 = \frac{124+124+128}{3} = 125$ thousand euro.

Then, we move on to the level for March with this indicator and, subsequently, proceed to the last row and calculate all the means one after another. Solving up to the end of a row, we calculate all the means.

Using the method of the accelerated mean, we calculate the output volume for three months (a quarter) and then determine the average monthly volume

$y_1 = \frac{118+124+124}{3} = 122$ thousand euro.

$y_2 = \frac{128+127+132}{3} = 129$ thousand euro, and so on (Table 8.8).

Taking into account the characteristics of the internal structure of the time rows (in terms of the dependence of the row levels), we can move the calculation of a moving mean to the end of the smoothing interval; we use balances (weights) due to the shifting of a row level from a time moment, rather than using symmetric balances. Such a mean is called *exponential* and is a linear combination of two values: a row level and an average level. The exponential mean is balanced (weighted) on a smoothing parameter, α, which is derived from the length of the smoothing interval, m: $\alpha = 2 / (m + 1)$.

The exponential mean is calculated with the formula $E(y_t) = \alpha y_t + (1 - \alpha)E(y_{t-1})$, where $E(y_t)$ is the exponential mean at moment t and α is the parameter representing the weight of the current row level, $0 < \alpha \leq 1$. From practical experience, a smoothing parameter should be limited to between the values 0.1 and 0.3. A value of 0.1 generally gives good results. In choosing a smoothing parameter, we can practically take into consideration that, with an increase in the values, the speed of the response to changes in the process increases, but the filtration qualities of an exponential mean decrease.

Let us calculate the exponential mean for different values of α from our example (Table 8.9).

Table 8.9. Exponential mean of output, thousand euro

Month	**Factual volume thousand euro**	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$
January	118	118.0	118.0	118.0
February	124	118.6	119.2	119.8
March	124	119.1	120.2	121.1
April	128	120.0	121.7	123.1
May	127	120.7	122.8	124.3
June	132	121.9	124.6	126.6
July	136	123.3	126.9	129.4
August	131	124.0	127.7	129.9
September	135	125.1	129.2	131.4
October	141	126.7	131.5	134.3
November	139	127.9	133.0	135.7
December	146	129.8	135.6	138.8
Total	1581	1475.2	1510.5	1532.5

The exponential mean has a very important attribute: it is easily adaptable to new conditions. This attribute of the exponential mean can be useful for short-term prediction. The methods of analytical smoothing make it possible to identify a trend and measure it quantitatively.

"Trend curves" are mathematical functions used to describe the main developmental trends. The type of function depends on the specificity of the process and the nature of its dynamics: steadiness, acceleration or slowing-down of an increase or decrease in the row levels (figures 8.2–8.4).

The developmental trends of an event over time are studied through the calculation of indicators of dynamics, with the use of simple methods of the analysis of dynamics, and the determination of the main developmental trend of the event under study, including trend, regressive, autoregressive, and combined models.

Trend models are used mostly to smoothen a time series with the aim of determining the main developmental tendency. Commonly, the smoothing process does not involve causal factors, which influence the levels of the time series. Often, there is no information about the specific reasons for the emergence of an event in time or about the nature of their effect, and the development is recognized as a time-dependent trend only. When smoothing a time row, we achieve a more aggregated summary row, which represents the combined effect of all factors over time. The deviation of a definite row level from levels corresponding to the general trend explain the effect of random factors.

To describe a developmental trend and make a prediction, we use growth curves of a certain time function, where fixed and random components are combined, and form a stationary random process. This gives us the trend model $Y = f(t) + \varepsilon$, where $f(t)$ is the level of the developmental trend and ε is the random deviation from the trend.

Indeed, such a presentation of a dynamic row appears to be an aberration that allows us to separate out regularity (trend) and randomness (residual variation). A trend $f(t)$ characterizes the main tendency and the direction and the form of movement, while the residual concerns the level of adequacy (approximation) of a theoretical model for a real development process.

The purpose of smoothing a time series is to determine the analytical or graphic trend $f(t)$. When a functional type of trend equation is chosen, importance is given to functions that have economic meaning and are easily interpreted. Usually, these parameters measure either an absolute speed (absolute increase) or a relative speed (growth rate).

In practice, we choose a type and determine the function parameters $f(t)$ from an available time row. Then we analyze the deviation from the trend. The function $f(t)$ is chosen to give a clear and informative explanation of the studied process. These functions are:

a) linear function: $Y_t = \alpha_0 + \alpha_1 t$, where parameter α_1 characterizes a stable absolute speed;
b) a parabola of the 2nd order: $Y_t = \alpha_0 + \alpha_1 t + a_2 t^2$, where a_2 indicates a stable increase in absolute speed (speed up or slow down);
c) an exponential function: $Y_t = \alpha_0 + \alpha_1^t$;
d) a power function: $Y_t = \alpha_0 t^{\alpha_1}$.

In all functions, t is an ordinal number indicating a period and α_0 is a row level at $t = 0$. Thus, the content/meaning of the other parameters depends on the type of function. In particular, a linear function is used for an absolute yearly increase; a demonstrative function is used for an average yearly growth rate; and a primary absolute increase is given from a parabola of the 2nd order. Two types of functions are most frequently used when smoothing is done: parabolic (second order) $Y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2$ and exponent of the first order $Y_t = e^{\alpha_0 + \alpha_1 t}$, or of the second order $Y_t = e^{\alpha_0 + \alpha_1 t + \alpha_2 t^2}$.

A parabolic function is chosen when an increase in the absolute chain shows some developmental tendency, but the absolute chain increase of the absolute chain increase (difference of the second order) does not show any developmental tendency. Exponential dependences are chosen if more or less constant relative growth is seen in an initial time series (stability of chain growth rates, rates of increase, growth coefficients) or (lack of) stability in change in the indicators of relative growth.

For function $Y_t = e^{\alpha_0 + \alpha_1 t}$, a relation of levels $f(t+1)$ to $f(t)$ is equal to e^{α_1}, i.e. a value of some indicator of a chain growth coefficient that does not depend on time. For function $Y_t = e^{\alpha_0 + \alpha_1 t + \alpha_2 t^2}$, $e^{2\alpha_2}$ is the value that is independent of time for a relation of the growth coefficient $\big(f(t+2) / f(t+1)\big)$ to the preceding $\big(f(t+1) / f(t)\big)$. Additionally, $f(t)$ can be transformed as $Y_t = e^{\alpha_0 + \alpha_1 t} = e^{\alpha_0} \times e^{\alpha_1 t} = A_0 \times A_1^t$, where A_0 summarizes an initial level of the time series and A_1^t presents some constant growth coefficient.

Quantitative values of the parameters of the function describing a general trend can determined by various methods. The simplest method, which gives quite satisfactory results, is the mean. It implies that, depending on the number of unknown parameters of function $f(t)$, an initial row is divided into a corresponding number of intervals.

A mean value is determined for each interval separately (the logarithm of previous initial levels are taken for the exponential trend). Quantitative values of the unknown parameters are determined from an equation system. In most calculations, the least squares method (LSM) is used for the smoothing of the time series. This ensures the smallest sum of the squared deviations of factual levels from the smoothed ones. In this case, we consider a common paired regression of the studied indicator over time. The results of LSM are more accurate and allow us to estimate the statistical significance of the parameters. Let us do an analytical smoothing with a straight line for our example $y = \alpha_0 + \alpha_1 t$, where α_0 and α_1 are the parameters for finding a straight line. They are determined by the least squares method

$$\begin{cases} n\alpha_0 + \alpha_1 \sum t = \sum y \\ \alpha_0 \sum t + \alpha_1 \sum t^2 = \sum ty \end{cases}.$$

This expression is simplified if the middle row is taken as the reference time. Then, $\sum t = 0$. The expression becomes $\begin{cases} n\alpha_0 = \sum y \\ \alpha_1 \sum t^2 = \sum ty \end{cases}$.

The solution of this expression is quite simple, even without the use of the indicator method (Table 8.10).

Table 8.10. Calculation of equation parameters of a linear trend

Month	Factual Volume, thousand euro	Indicator calculation			
		t	t^2	y_t	Y_t
January	118	-11	121	-1298	120
February	124	-9	81	-1116	122
March	124	-7	49	-868	124
April	128	-5	25	-640	126
May	127	-3	9	-381	128
June	132	-1	1	-132	131
July	136	1	1	136	133
August	131	3	9	393	135
September	135	5	25	675	137
October	141	7	49	987	139
November	139	9	81	1251	142
December	146	11	121	1606	144
Total	1581	0	572	623	1581

$12 \times \alpha_0 = 1581$, $\alpha_0 = 1581/12 = 131.75$ thousand euro.
$572 \times \alpha_1 = 623$, $\alpha_1 = 623/572 = 1.089$.

The results of calculation give the trend equation

$$Y = 131.75 + 1.089 \times t.$$

Hence, the monthly output volume increased by 1089 euro.

For a parabola of the second order, we need to solve a system with three equations

$$\begin{cases} n\alpha_0 + \alpha_1 \sum t + \alpha_2 \sum t^2 = \sum y \\ \alpha_0 \sum t + \alpha_1 \sum t^2 + \alpha_2 \sum t^3 = \sum ty \\ \alpha_0 \sum t^2 + \alpha_1 \sum t^3 + \alpha_2 \sum t^4 = \sum t^2 y \end{cases},$$

since $\sum t = \sum t^3 = 0$, the expression becomes $\begin{cases} n\alpha_0 + \alpha_2 \sum t^2 = \sum y \\ \alpha_1 \sum t^2 = \sum ty \\ \alpha_0 \sum t^2 + \alpha_2 \sum t^4 = \sum t^2 y \end{cases}$

$$\begin{cases} n\alpha_0 + \alpha_2 \sum t^2 = \sum y \\ \alpha_1 = \frac{\sum ty}{\sum t^2} \\ \alpha_0 \sum t^2 + \alpha_2 \sum t^4 = \sum t^2 y \end{cases} \Rightarrow \begin{cases} n\alpha_0 + \alpha_2 \sum t^2 = \sum y \\ \alpha_0 \sum t^2 + \alpha_2 \sum t^4 = \sum t^2 y \end{cases} \Rightarrow$$

$$\begin{cases} \alpha_0 = \frac{\sum y \sum t^4 - \sum t^2 \sum t^2 y}{n \sum t^4 - (\sum t)^2} \\ \alpha_2 = \frac{n \sum t^2 y - \sum y \sum t^2}{n \sum t^4 - (\sum t)^2} \end{cases}.$$

Example 8.3

Let us consider the smoothing of a time series with a parabola of the second order and the example of the volume of domestic money supply over a period of 10 years (Table 8.11).

$$\begin{cases} \alpha_0 = \frac{\sum y \sum t^4 - \sum t^2 \sum t^2 y}{n \sum t^4 - (\sum t)^2} = \frac{1242209 \times 19338 - 330 \times 54026825}{10 \times 19338 - (330)^2} = 73307 \text{ million euro} \\ \alpha_1 = \frac{\sum ty}{\sum t^2} = \frac{6126677}{330} = 18566 \text{ million euro} \\ \alpha_2 = \frac{n \sum t^2 y - \sum y \sum t^2}{n \sum t^4 - (\sum t)^2} = \frac{10 \times 54026825 - 1242209 \times 330}{10 \times 19338 - (330)^2} = 1984 \text{ million euro} \end{cases}$$

.

Table 8.11. Calculation of the parameters of the trend equation with a parabola of the second order

Year	Money supply (M2) by the end of the year, million euro	t	$t \times y$	t^2	t^4	$t^2 \times y$
1	15432	-9	-138888	81	6561	1249992
2	21714	-7	-151998	49	2401	1063986
3	31387	-5	-156935	25	625	784675
4	45186	-3	-135558	9	81	406674
5	64321	-1	-64321	1	1	64321
6	94855	1	94855	1	1	94855
7	125483	3	376449	9	81	1129347
8	193145	5	965725	25	625	4828625
9	259413	7	1815891	49	2401	12711237
10	391273	9	3521457	81	6561	31693113
Total	1242209	0	6126677	330	19338	54026825

The trend equation is $y = 73307 + 18566t + 1984t^2$. From this, we can see that money supply increases annually by 18566 million euro with an average positive growth rate of 1984 million euro. We can determine the theoretical levels and estimate the levels of adequacy of the chosen trend model (Table 8.12). The adequacy level can be estimated with the help of a determination coefficient, as in a correlation-regression analysis, using the formula

$$R^2 = 1 - \frac{\sum((Y_t - y_t)^2 / n)}{\frac{\sum y_t^2}{n} - \left(\frac{\sum y_t}{n}\right)^2} = 1 - \frac{5911818551 / 10}{\frac{29031E+11}{10} - \left(\frac{1242209}{10}\right)^2} = 0.956.$$

The determination coefficient's theoretical value taken from the table of critical values (appendix Table 1) is $R^2_{0.05}(2;7) = 0.575$. This is much lower than the calculated value. As such, the second order parabola

adequately describes growth in the money supply. Another technique to estimate the adequacy of a trend equation is to calculate a standard error of approximation with the formula

$$v = \sqrt{\frac{1}{n-m-1}\sum\left(\frac{Y_t - y_t}{y_t}\right)^2}\ 100\%.$$

If the value remains within 15 %, the model can be used predictively.

Table 8.12. Calculation of the theoretical levels and level of adequacy

Year	Money supply (M2) at the end of the year, million euro,		$Y_t - y_t$	$(Y_t - y_t)^2$	$(y_t)^2$	$\frac{Y_t - y_t}{y_t}$
	Factual levels, y_t	Theoretical levels, Y_t				
1	15432	66917	51485	3E+09	2.4E+08	11.131
2	21714	40561	18847	4E+08	4.7E+08	0.753
3	31387	30077	-1310	2E+06	9.9E+08	0.002
4	45186	35465	-9721	9E+07	2E+09	0.046
5	64321	56725	-7596	6E+07	4.E+09	0.014
6	94855	93857	-998	1E+06	9E+09	0.000
7	125483	146861	21378	5E+08	1.6E+10	0.029
8	193145	215737	22592	5E+08	3.7E+10	0.014
9	259413	300485	41072	2E+09	6.7E+10	0.025
10	391273	401105	9832	1E+08	1.5E+11	0.001
Total	591523	1387790	145581	6E+09	2.9E+11	12.014

To estimate the extent of dynamic stability, we use the residual dispersion of the time rows. This is calculated from the deviation of theoretical and calculated values with the formula $\sigma_\varepsilon^2 = \frac{\sum_{t=1}^{n}(y_t - Y_t)^2}{n-m}$, where *n* is the number

of levels of a time series and m is the number of parameters of the trend equation.

In choosing a trend function, empirical methods are used alongside theoretical methods for the analysis of regularities: a comparison of several functions is based on the average of the squared error

$$\sigma_\varepsilon = \sqrt{\frac{1}{n-m}\sum_{t=1}^{n}(y_t - Y_t)^2}.$$

The smaller the value of this error, the greater the adequacy of the trend model. To identify and check the adequacy of a trend, a cumulative criterion is also used, which is calculated from the relation of the accumulated sum of deviations at row levels from an average level and their internal deviations $= \frac{\sum_{l=1}^{n}\left(\sum_{t=1}^{l}(y_t - \bar{y})\right)^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2}$.

Accumulated deviations have the ability to distinguish changes in the regularities of deviation. The deviations in one sign form a series and, when added at certain intervals, they are not cancelled. The procedure of checking a criterion is standard – a calculated value of a criterion is compared with a theoretical one at a given level of significance for α. If a calculated value is larger than a critical one, the hypothesis for the absence of the trend is otherwise rejected and the trend is significant (appendix Table 9).

Let us check a hypothesis on the presence of a trend in the time series indicating GDP volume using the data in Table 8.13.

Table 8.13. Calculation of cumulative criterion for GDP dynamics over 11 years, million euro

Year	**GDP** y_t	$(y_t - \bar{y})$	$\sum_{t=1}^{l}(y_t - \bar{y})$	$\left(\sum_{t=1}^{l}(y_t - \bar{y})\right)^2$
1	81519	-154604	-154604	23902396816
2	93365	-142758	-297362	88424159044
3	102593	-133530	-430892	1.85668E+11
4	130442	-105681	-536573	2.87911E+11
5	170070	-66053	-602626	3.63158E+11
6	204190	-31933	-634559	4.2665E+11
7	225810	-10313	-644872	4.1586E+11
8	267344	31221	-613651	3.76568E+11
9	344822	108699	-504952	2.54977E+11
10	441500	205377	-299575	89745180625
11	535700	299577	2	4
Total	$\bar{y}$ =236123	x	x	2.8888E+12

From the calculated results (Table 8.13), we determine a value for the criterion *T*:

$$T = \frac{2.48888E+12}{2031694234 \quad 6x11} = 11.14.$$

The calculated value is higher than the critical value $T_{0.05}(11)$=5.02 = 5.02 (appendix, Table 9). This confirms the significance of the trend for the time row indicating GDP. If there is a tendency towards stability in the growth rates when the row levels change, then time series smoothing must be done using the exponential function $Y_t = a_0 a_1^t$, where a_1 is the growth rate.

The technique of smoothing using a demonstrative function is similar to smoothing with a straight line, except for the fact that, in this case, row logarithms are measured, but not their levels $\log y = \log a_0 + t \log a_1$. Thus, the expression becomes

$$\begin{cases} n \log a_0 + \log a_1 \sum t = \sum \log y \\ \log a_0 \sum t + \log a_1 \sum t^2 = \sum t \log y \end{cases}.$$

Since $\sum t = 0$, we get $\begin{cases} n \log a_0 = \sum \log y \\ \log a_1 \sum t^2 = \sum t \log y \end{cases}$

then, $\log a_0 = \frac{\sum \log y}{n}$; $\log a_1 = \frac{\sum t \log y}{t^2}$.

From the logarithmic values, we can determine the values of the equation parameters using an exponential function.

An analytical equation presents a mathematical model of an event and gives an expression of statistical regularity observed in a time series. Following the technical conventions for the analytical smoothing of a time series is very important. This is due to the fact that the levels are considered as a time function. In fact, the development of an event is caused by the effect of forces having an impact on such a development in terms of their direction and intensity, rather than the effect of time. The development of an event over time presents an external manifestation of the effect of these forces and their combined effect influences changes in the level at different time intervals.

Models that account for the general regularities of an event that sees changes at time intervals and changes from a complex of factors are called *multi-factorial dynamic models*. If the value of indicator *y* depends on the change in a complex of factors $x_1, x_2, \ldots\ldots\ldots, x_k$, then, from the data about a population over a period of time, one can build a correlation model to characterize the dependence of y on the mentioned factors in each time period. We can present such a dependence as a linear function

for t = 1, $y\,(1) = \alpha_{01} + \alpha_{11}x_1 + \alpha_{21}x_2 + \ldots\ldots\ldots + \alpha_{k1}x_k$

for t = 2, $y\,(2) = \alpha_{02} + \alpha_{12}x_1 + \alpha_{22}x_2 + \ldots\ldots\ldots + \alpha_{k2}x_k$

for t = 3, $y\,(3) = \alpha_{03} + \alpha_{13}x_1 + \alpha_{23}x_2 + \ldots\ldots\ldots + \alpha_{k3}x_k$

etc.

For all the periods, we have a system of n equations and n coefficients from a regression for each of the factors. Thus, we can generate a matrix of regression coefficients $k \times n$ as

$$\alpha_{01}\ \alpha_{11}\ \alpha_{21}\ \ldots\ \alpha_{k1}$$

$$\alpha_{02}\ \alpha_{12}\ \alpha_{22}\ \ldots\ \alpha_{k2}$$

$$\alpha_{03}\ \alpha_{13}\ \alpha_{23}\ \ldots\ \alpha_{k3}$$

$$\ldots\ \ldots\ \ldots\ \ldots\ \ldots$$

$$\alpha_{0n}\ \alpha_{1n}\ \alpha_{2n}\ \ldots\ \alpha_{kn}$$

From each of these time rows, we can present α_{ij} as a function of time t and using analytical smoothing we can predict regression coefficients for time period t.

8.6. Prediction of the levels of a time series

Sometimes there is a need to discover the absence of intermediate row levels. This procedure is called ***interpolation*** and is done to understand a general developmental tendency for the period under study.

When predicting economic indicators, another statistical technique is used called ***extrapolation***. Using this technique, the values of the levels beyond available factual data are calculated. In extrapolation, the assumption is that the reasons underlying the main developmental tendency for a population move forwards and the identified tendency does not change in the short term. To carry out this operation, we should put a proper value for t into the trend equation, according to the extension of the initial row, and calculate Y_t. Thus, the generated value is called *a point estimate of prediction*.

In our example, a point estimate is $Y_{2008} = 73307 + 18566 \times 11 + 1984 \times 11^2 = 517597$ million euro. To make an interval estimate of prediction, confidence intervals are calculated within which the predicted value will lie at a certain probability level. A trust number t is the trust interval, which is determined from the quantile distribution tables of Student's distribution with a certain probability level, α, and degrees of freedom (n–m), as well as the standard prediction error

$$\sigma_p = \sigma_\varepsilon \times \sqrt{\frac{n+1}{n} + \frac{3(n+2v-1)^2}{n\ (n^2-1)}},$$

where v is the interval of prediction and σ_ε^2 is the estimate of residual dispersion calculated with the formula $\sigma_\varepsilon = \sqrt{\frac{\sum(Y_t - y_t)^2}{n-m}}$, where m is the number of parameters of the trend equation. In our example, $v = 2$, since an interval consists of two units for one year

$\sigma_\varepsilon = \sqrt{\frac{5.91E+09}{10-3}} = 29061$ million euro;

$\sigma_p = 29061 \times \sqrt{\frac{10+1}{10} + \frac{3(10+2\times 2-1)^2}{10\ (10^2-1)}} = 36899$ million euro;

$t_{0,05}(10-3) = 1.89$.

$Y_{2008} = 517597 \pm 1.89 \times 36899 = 517597 \pm 68208$ million euro.

giving us $449389 \leq Y_{2006} \leq 585805$ million euro.

The amount of money supply in 2008 (at the end of the year) was likely to lie between 449389 million and 585805 million euro. At the end of 2008, the factual value of the amount of money supply was 570706 million euro. Thus, a parabola of the second order can be used to predict estimates of the quantity of money supply.

8.7. The statistical study of seasonality

Many social and economic processes are impacted by seasonal fluctuations – from year to year the level of impact either increases or decreases in some places. These internal fluctuations, which are of a somewhat regular nature, are described as ***seasonal***.

The statistical study of seasonality seeks to:

1) quantify seasonal fluctuations and identify their intensity and nature;
2) identify factors that cause seasonal fluctuations;
3) estimate the consequences that result from seasonal fluctuations.

When analyzing seasonality, the levels of a time row can reflect the development of an event by months (quarters) or for several years. We can

identify and measure them using different methods. The most common of these are:

- the method of absolute differences;
- the method of relative differences;
- the use of indices of seasonality;
- the method of average quadratic deviations.

Indices of seasonality are the most frequently used. Here, a summary seasonality index is calculated for each month (quarter) with the arithmetic mean from similar indices of every year and using the formula $i_S = \frac{y_t}{y_t'}$, where y_t' is either an average value of a row level (a constant mean), or a smoothed value in a trend (a variable mean). The variable mean is used for rows that have a distinct developmental tendency with the formula $i_S = \frac{y_t}{Y_t}$.

The constant mean is used for rows with an unclear main developmental tendency, namely, when there is no trend, with the formula $i_S = \frac{y_t}{\bar{y}}$.

We calculate indices of seasonality in the example below (Table 8.14).

Table 8.14. Calculation of seasonality indices

Month	Output volume, thousand euro		Seasonality indices by the first method	Seasonality indices by the second method
	factual	smoothed		
January	118	121	0.975	0.894
February	124	123	1.008	0.939
March	124	125	0.992	0.939
April	128	127	1.008	0.969
May	127	130	0.977	0.962
June	132	131	1.008	1.000
July	136	133	1.022	1.030
August	131	135	0.970	0.992
September	135	137	0.985	1.022
October	141	139	1.014	1.068
November	139	141	0.986	1.053
December	146	143	1.021	1.106
Total	1581	1581	x	x

1) $y_1 = 118 / 121 = 0.975$.

2) $\bar{y} = \frac{\sum y_i}{n} = \frac{1581}{12} = 132$ thousand euro.

$y_1 = 118 / 132 = 0.894$ thousand euro.

Using these indices, we can build seasonality lines to describe ***a seasonal wave*** (Figure 8.5).

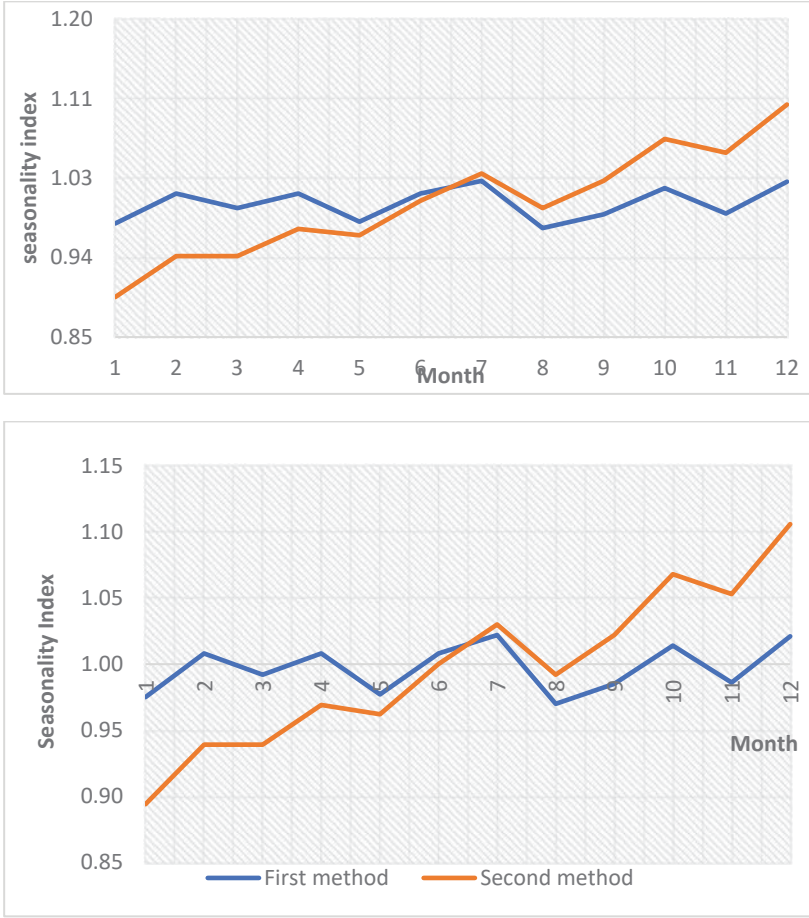


Figure 8.5. Seasonality line of output volume

Let us calculate a prediction for output volume using the trend model from Example 8.2 and considering a seasonal component (Table 8.15).

Table 8.15. Calculation of predicted values of output volume considering seasonality indices

Month	Output volume, thousand euro		Seasonality indices	Predicted output volume considering seasonal component
	factual	Smoothed by trend equation		
January	118	120	0.975	**120×0.975=117**
February	124	122	1.008	123.0
March	124	124	0.992	123.0
April	128	126	1.008	127.0
May	127	128	0.977	125.1
June	132	131	1.008	132.0
July	136	133	1.022	135.9
August	131	135	0.970	131.0
September	135	137	0.985	134.9
October	141	139	1.014	140.9
November	139	142	0.986	140.0
December	146	144	1.021	147.0
Total	1581	1581	x	**x**

8.8. Harmonic fluctuations

When there is a steady deviation from a tendency, showing an increase and/or decrease, recorded in an analyzed time sequence, one may assume the presence of some (one or several) fluctuation processes in the time series. This event becomes more distinct when we study processes that have a seasonal nature, such as when the increase or decrease of the levels is

regularly repeated with an interval of one year (e.g., the volume of investments in fixed capital). Obviously, in addition to processes that repeat with a one-year interval, dynamic rows can also include fluctuation processes with higher and lower periodicity. As such, one of the tasks in the analysis of a time series is to separate the fluctuation processes to estimate their significance and impact on the development of a further tendency.

The analysis is undertaken through treating a time series as a process of harmonic fluctuations. The following expression is appropriate for each point of the time row

$y_t = f(t) + \sum_{k=1}^{m} \left(a_k \, cos \left(kt \frac{2\pi}{n} \right) + b_k \, sin \left(kt \frac{2\pi}{n} \right) \right)$, where *t = 1,2,…,n.*

Here, y_t is the time series level at moment (interval) t; $f(t)$ smooths the row level at moment (interval) t; α_k and b_k are parameters of the fluctuation process (harmony) with number *k*, which estimate the range (amplitude) of deviation from a general tendency and the switching of fluctuations from the initial point. Since the developmental tendency $f(t)$ is determined by other methods, so as to analyze fluctuation processes, we can use row ε_t, which shows the deviation from a tendency. If the development of a tendency is not recorded in the initial row, analysis of fluctuation is done with primary data; in this case $f(t)$ is replaced with the general average level $\overline{Y}$.

The total number of fluctuation processes defined in a row consisting of *n* levels is determined as *n/2*, i.e. the maximal value *k* is equal to *n/2*. Usually, the most important harmonics are used. It should be noted that the superimposition of harmonics ensures a higher percentage of explained variation. To identify the events belonging to a periodic type, a Fourier series is used, $y = \alpha_0 + \sum(\alpha_k \cos kt + b_k \sin kt)$. In this equation, k indicates the Fourier harmonic and can be taken at different degrees of accuracy (the most frequent harmonics are 1 to 4). To discover the equation parameters, the method of least squares is used.

When the partial derivatives of this function are computed, they are approximated to zero and we get a series of normal equations. The solution of this system makes it possible to determine the equation parameters

$\alpha_0 = \frac{\sum y}{n}$;

$\alpha_k = \frac{2 \sum y \cos kt}{n}$;

$b_k = \frac{2\sum y \sin kt}{n}$.

It becomes clear that the parameters depend on the row levels and sequence of $\cos kt$ and $\sin kt$. When studying seasonal fluctuations over a year, n = 12 or n = 4 (according to the number of months or quarters, respectively).

A time series can be presented in the following sequence.

0	$\frac{\pi}{6}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$	$\frac{2\pi}{3}$	$\frac{5\pi}{6}$	π	$\frac{7\pi}{6}$	$\frac{8\pi}{6}$	$\frac{9\pi}{6}$	$\frac{10\pi}{6}$	$\frac{11\pi}{6}$
y_0	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}

A sequence of values, 0, $\frac{\pi}{6}$, $\frac{\pi}{3}$, etc., is calculated as $\frac{2\pi}{n}t$. To calculate the sines and cosines of various harmonics, we can use Table 8.16.

Table 8.16. Calculation of sine and cosine values for various harmonics

t	$\cos t$	$\cos 2t$	$\cos 3t$	$\cos 4t$	$\sin t$	$\sin 2t$	$\sin 3t$	$\sin 4t$
0	1	1	1	1	0	0	0	0
$\frac{\pi}{6}$	0.866	0.5	0	-0.5	0.5	0.866	1	0.866
$\frac{\pi}{3}$	0.5	-0.5	-1	-0.5	0.866	0.866	0	-0.866
$\frac{\pi}{2}$	0	-1	0	1	1	0	-1	0
$\frac{2\pi}{3}$	-0.5	-0.5	1	-0.5	0.866	-0.866	0	0.866
$\frac{5\pi}{6}$	-0.866	0.5	0	-0.5	0.5	-0.866	1	-0.866
π	-1	1	-1	1	0	0	0	0
$\frac{7\pi}{6}$	-0.866	0.5	0	-0.5	-0.5	0.866	-1	0.866
$\frac{8\pi}{6}$	-0.5	-0.5	1	-0.5	-0.866	0.866	0	-0.866
$\frac{9\pi}{6}$	0	-1	0	1	-1	0	1	0
$\frac{10\pi}{6}$	0.5	-0.5	-1	-0.5	-0.866	-0.866	0	0.866
$\frac{11\pi}{6}$	0.866	0,5	0	-0.5	-0.5	-0.866	-1	-0.866
2π	1	1	1	1	0	0	0	0

To find the parameters, we need to calculate the product of the levels for a given month using the sines and cosines of the corresponding harmonics. An equation for $k = 1$ is $y = \alpha_0 + a_1 \cos t + b_1 \sin t$.

The equation parameters are calculated with the formulas

$\alpha_0 = \frac{\sum y}{12}$;

$\alpha_1 = \frac{\sum y \cos t}{6}$;

$b_1 = \frac{\sum y \sin t}{6}$.

Let us build a seasonal wave from our example using the first and second harmonics (Table 8.17).

Table 8.17. Calculation of the parameters of Fourier equations for the first and second harmonics

Month	t	y_t	$y \cos t$	$y \sin t$	Y_t	$y \cos 2t$	$y \sin 2t$	Y_t
January	0	118	118.0	0.0	129.8	118.0	0.0	128.0
February	$\frac{\pi}{6}$	124	107.4	62.0	126.5	62.0	107.4	121.8
March	$\frac{\pi}{3}$	124	62.0	107.4	124.5	-62.0	107.4	121.7
April	$\frac{\pi}{2}$	128	0.0	128.0	124.5	-128.0	0.0	126.3
May	$\frac{2\pi}{3}$	127	-63.5	110.0	126.4	-63.5	-110.0	131.1
June	$\frac{5\pi}{6}$	132	-114.3	66.0	129.8	66.0	-114.3	132.6
July	π	136	-136.0	0.0	133.7	136.0	0.0	131.8
August	$\frac{7\pi}{6}$	131	-113.4	-65.5	137.0	65.5	113.4	132.4

September	$\frac{8\pi}{6}$	135	-67.5	-116.9	139.0	-67.5	116.9	136.1
October	$\frac{9\pi}{6}$	141	0.0	-141.0	139.0	-141.0	0.0	140.8
November	$\frac{10\pi}{6}$	139	69.5	-120.4	137.1	-69.5	-120.4	141.7
December	$\frac{11\pi}{6}$	146	126.4	-73.0	133.7	73.0	-126.4	136.6
Total	x	1581	-11.4	-43.4	1581.0	-11.0	-26.0	1581.0

For the first (fundamental) harmonic, the function becomes $y = 131.75 - 1.906\cos t - 7.236\sin t$, where

$$\alpha_0 = \frac{\sum y}{12} = \frac{1581}{12} = 131.75;$$

$$\alpha_1 = \frac{\sum y\cos t}{6} = \frac{-11.4}{6} = -1.906;$$

$$b_1 = \frac{\sum y\sin t}{6} = \frac{-43.4}{6} = -7.236.$$

For the second harmonic, we can use the following $y = 131.75 - 1.906\cos t - 7.236\sin t - 1.833\cos 2t - 4.33\sin 2t$, where

$$\alpha_2 = \frac{\sum y\cos 2t}{6} = \frac{-11}{6} = -1.833;$$

$$b_2 = \frac{\sum y\sin 2t}{6} = \frac{-26}{6} = -4.33.$$

From the harmonic coefficients, we can determine the amplitude of fluctuation for each harmonic $A_k = \sqrt{\alpha_k^2 + b_k^2}$.

The larger the amplitude of the fluctuation, the larger the role/effect of the corresponding harmonic on the general dispersion of the process. An apparatus of harmonic analysis allows us to estimate the role of each

fluctuation process in the general dispersion of the time row. To estimate the role of each harmonic, a coefficient of determination is used $R_k^2 = \frac{\delta_k^2}{\sigma^2}$, where $\delta_k^2 = \frac{A_k^2}{2}$ is the dispersion of *k*-harmonic and $\sigma^2 = \frac{\sum(y_t - \bar{y})^2}{n} = \bar{y}^2 - (\bar{y})^2$ is the general dispersion of the process.

For our example, the amplitude of the first harmonic is equal to

$A_1 = \sqrt{\alpha_1^2 + b_1^2} = \sqrt{(-1.906)^2 + (-7.236)^2} = 7.5.$

the general dispersion is calculated with the formula (Table 8.18)

$\sigma^2 = \bar{y}^2 - (\bar{y})^2 = \frac{209013}{12} - \left(\frac{1581}{12}\right)^2 = 59.9$, and the dispersion of the first (fundamental) harmonic is $\delta_1^2 = \frac{A_1^2}{2} = \frac{7.5^2}{2} = 28$.

Thus, the role/effect of the first harmonic on the fluctuation process is 46.7%, $R_1^2 = \frac{\delta_1^2}{\sigma^2} = \frac{28}{59.9} = 0.467$.

For the second harmonic, it is

$A_2 = \sqrt{\alpha_2^2 + b_2^2} = \sqrt{(-1.833)^2 + (-4.33)^2} = 4.7;$

$\delta_2^2 = \frac{A_2^2}{2} = \frac{4.7^2}{2} = 11.05;$

$R_2^2 = \frac{\delta_2^2}{\sigma^2} = \frac{11.05}{59.9} = 0.185 = 18.5\%.$

Table 8.18. Calculation of general dispersion

Month	Factual volume thousand euro, y_t	y_t^2
January	118	13924
February	124	15376
March	124	15376
April	128	16384
May	127	16129
June	132	17424
July	136	18496
August	131	17161
September	135	18225
October	141	19881
November	139	19321
December	146	21316
Total	1581	209013

Using a special calculation, we can estimate the role/contribution of each harmonic to the general sum of a row, e.g., we can show its effect/contribution to each fluctuation process for a total yearly output volume $\Delta_k = \sigma_k \sqrt[n]{2}$ for *k=1,2,...,n/2-1* then for $\Delta_{n/2} = \sigma_{n/2} n$.

Practice Exercises

Exercise 8.1

The table below gives the in-stock balance of a warehouse.

Date	thousand euro
01.01	50
14.01	56
20.01	45
01.02	52

Calculate the in-stock balance in the warehouse for January.

Exercise 8.2

Below, we have the in-stock balance of a warehouse in the first quarter.

Date	thousand euro
01.01	15
01.02	12
01.03	6
01.04	10

Calculate the average in-stock balance in the warehouse for the first quarter.

Exercise 8.3

An institute plans to increase first-year student admissions by 15 % over five years. If first-year admissions amounted to 24000 students in the current year, determine the percentage increase of yearly admissions and calculate the number of additional students admitted.

Exercise 8.4

The current population of a city is 158 thousand. Calculate the potential population in three years if the average annual rate of increase is 2 %.

Exercise 8.5

The annual increase in import volumes is represented by the data below (in percentage of previous year).

1st year	2nd year	3rd year	4th year	5th year	6th year
+11.9	+7.7	+34.6	+26.9	+25.7	+24.8

If the value is US $15.1 billion in the base year, determine the relative and absolute increases of import volumes for each year and the average per year.

Exercise 8.6

The yearly dynamics of gross wheat yield are represented by the data below (in percentage of the previous year). Determine the relative change in gross wheat yield for each year and the average per year.

1st year	2nd year	3rd year	4th year
108	96	109	102

Exercise 8.7

Using dynamic indicators of interconnection, determine the row levels (output volume) and chain indicators of dynamics, which are missing in the table below. Also determine the average yearly growth and rate of increase.

Year	Million euro	Chain dynamics indicators			
		Absolute increase, million euro	Growth rate, %	Rate of increase, %	Absolute value of 1 % increase
1	265	x	x	x	x
2		2.5			
3			102.5		
4					
5				+5	2.85

Exercise 8.8

Using dynamic indicators of interconnection, determine the row levels (total area of residential buildings put into operation by year) and indicators of base dynamics (the base is the 1st year), which are missing in the table. Also determine the average yearly growth and rate of increase.

Year	Million m^2	Indicators of base dynamics		
		Absolute increase, million m^2	Growth rate, %	Rate of increase, %
1	17.4	0.0	100	0.0
2				16.7
3			81	
4		-5.1		
5			58	

Exercise 8.9

Having calculated chain growth and rate of increase, analyze the dynamics of gross investments in fixed capital in comparable prices and also the average annual growth and rate of increase. Present the results of your calculations in the table. Make some conclusions.

Year	1	2	3	4	5	6	7
Investments in fixed capital in current prices, bln euro	23.6	32.6	37.2	51.0	75.7	93.1	125.3
Price index, % of previous year	117.7	114.1	103.5	105.9	115.0	120.7	113.0

Exercise 8.10

Some data is given below on the number of registered stock exchanges by year. Determine the average row level, calculate all chain indicators of a time series, and determine the average annual growth and rate of increase for the whole period.

Year	1	2	3	4	5
Total, un.	66	87	74	91	102

Exercise 8.11

Some data is given below about the number of registered marriages by year. Determine the average row level, calculate all chain indicators of a time series, and determine the average annual growth and rate of increase of marriages for the whole period.

1	2	3	4	5	6	7
274.5	309.6	317.2	371.0	278.2	332.1	355.0

Exercise 8.12

Some data on the dynamics of the population of a country is given below. Construct a trend model using a straight line and a parabola of the second order. Estimate the adequacy of each model and check its predictive potential. Build point and interval predictions for the next two years.

Year	2000	2001	2002	2003	2004	2005	2006
Population number at the beginning of the year, million	50.1	49.7	49.3	48.5	48.0	47.6	47.3

Exercise 8.13

Using the data below about the share of a country's population with average monthly expenses per capita that are lower than the cost of living (%) by aggregate and by monetary income over 7 years, construct a trend model using a straight line and a parabola of the second order. Estimate the adequacy of each model and check its predictive potential. Build point and interval predictions for the next two years.

Year	Population share with average monthly expenses per capita that are lower than the cost of living, %	
	money	aggregate
1	87.9	80.2
2	89.0	82.7
3	88.4	83.3
4	83.4	76.2
5	73.8	65.6
6	64.0	55.3
7	59.4	50.9

Exercise 8.14

Below are some data on the dynamics of private vehicle supply among a country's population. Build a trend model using a straight line and a parabola of the second order. Estimate the adequacy of each model and check its predictive potential. Build point and interval predictions for the next two years.

Year	1	5	10	11	12	13	14	15	16
(per 1000 people, pcs.)	63	87	104	105	108	105	108	113	115

Exercise 8.15

Using the data below on interest rates for credits and deposits over two years (%), estimate the seasonal and cyclical fluctuations and determine the contribution/effect of each harmonic in the fluctuation process using the apparatus of harmonic analysis. Build lines for seasonality.

Month	Interest rates on credits				Interest rates on deposits			
	in national currency		in foreign currency		in national currency		in foreign currency	
	1	2	1	2	1	2	1	2
January	16.1	14.8	11.3	11.4	8.1	8.0	7.0	6.3
February	16.1	14.5	11.4	11.6	7.8	8.1	6.6	5.8
March	16.0	14.5	11.4	11.6	7.3	8.1	6.3	5.9
April	15.9	14.6	11.3	11.7	7.8	7.8	5.7	5.9
May	15.6	14.2	11.0	11.3	7.8	7.7	5.4	5.7
June	15.7	14.6	11.2	11.2	7.2	7.7	5.3	5.7
July	15.3	14.5	11.0	11.4	7.0	8.2	5.5	5.8
August	15.2	14.0	11.2	11.2	7.3	8.0	5.3	5.6
September	15.1	14.3	11.3	11.1	7.4	8.3	5.3	5.6
October	14.9	14.2	11.3	11.0	7.9	8.3	6.0	5.8
November	14.9	14.4	11.2	11.1	7.4	8.6	5.8	5.8
December	15.4	14.8	11.5	11.2	7.9	8.6	6.1	6.3

Exercise 8.16

Using the data below on the dynamics of the number of people going abroad, make some conclusions about the nature of the developmental tendency. Estimate the adequacy of each model and check its predictive potential. Build point and interval predictions for the next two years.

Year	Thousand people
1	13422
2	14849
3	14729
4	14795
5	15488
6	16454
7	16875
1	17335

Exercise 8.17
Below is some data on in-stock balance in a warehouse for March.

Date	thousand euro
01.03	15
15.03	12
22.03	6
28.03	12
01.04	10

Calculate the average in-stock balance in the warehouse for March.

Exercise 8.18

Below is some data on in-stock balance in a warehouse for the second quarter.

Date	thousand euro
01.04	120
01.05	100
01.06	110
01.07	140

Calculate the average in-stock balance in the warehouse for the second quarter of the year.

Exercise 8.19

The current population of a city is 253 thousand. Calculate the potential population of the city in four years if the average annual rate of increase of 2.6 %.

Exercise 8.20

A plan is made to increase the operational housing area by 25 % over five years. If this area was 1.4 million m^2 in the current year, determine the necessary annual percentage increase and calculate the additional housing area put into operation.

Exercise 8.21

Having calculated the chain growth and rate of increase, analyze the dynamics of gross investments in fixed capital in comparable prices, average annual growth, and rate of increase. Present the results in a table. Make some conclusions.

Year	1	2	3	4	5
Investments in fixed capital in current prices, billion euro	12.3	14.5	16.9	17.2	17.2
Price index, times to the first year	1.0	1.12	1.14	1.15	1.15

Exercise 8.22

The annual increase in population is represented by the data below (as a percentage of the previous year).

1st year	2nd year	3rd year	4th year
+1.5	+1.2	-1.1	+0.5

Determine the relative change of the population for the whole period studied and the average per year.

Exercise 8.23

The annual dynamics of output volume are presented in the data (as a percentage of the previous year). Determine the relative change in output volume for the whole period studied and the average per year.

1st year	2nd year	3rd year	4th year
102	98	99	101

Exercise 8.24

Using indicators of the dynamics of interconnection, determine the row levels (the number of professors by years) and indicators of chain dynamics, which are absent in the table, and the average annual growth rate.

Year	People	Indicators of base dynamics		
		Absolute increase, people	Growth rate	Rate of increase, %
1	10339	0	1.000	0.0
2		920		
3				11.9
4			1.162	
5				20.8

Exercise 8.25

Using the data below on the number of retired people (pensioners) at the beginning of the year, determine the average row level by years and calculate all chain indicators of a time series. Determine the average annual growth and the rate of increase of the number of pensioners for the whole period (million people).

1	2	3	4	5	6	7
14447	14423	14376	14348	14065	14050	13937

Exercise 8.26

Using the data below on the number of unemployed people (according to the ILO's methodology), in millions at the end of the year, determine the average row level and calculate all base indicators of a time series (the 1st year is the base). Determine the average annual rate of change in the number of the unemployed for the whole period.

1	2	3	4	5	7	8
2655.8	2455.0	2140.7	2008.0	1906.7	1600.8	1515.0

Exercise 8.27

The quarterly increase in the output of a branch is presented in the table below (as a percentage of the previous quarter).

I quarter	II quarter	III quarter	IV quarter
+13	+12	+15	+14

If the gross output volume was 120 million euro in the fourth quarter of the previous year, determine the absolute and relative increase in output per year and on average per quarter. Make some conclusions.

Exercise 8.28

The annual increase (decrease) in operational housing area is presented in the table below (in percent of the previous year).

1-й рік	2-й рік	3-й рік	4-й рік
+9	-2	-1,5	+8

If the housing area put into operation was 20 million m^2 in the base period, determine the absolute and relative change in the operational housing area for the whole period studied and the yearly average. Make some conclusions.

Exercise 8.29

Using the dynamic characteristics of interconnection, determine the size of a joint venture capital project and its absolute and relative growth rate for the last five years. Make some conclusions.

Year	JV capital, million euro	Chain characteristics			
		Absolute increase, million euro	Growth rate, %	Rate of increase, %	Absolute value of 1% increase, million euro
1	336.0	x	x	x	x
2					
3				6	3.84
4		45			
5			115		

Exercise 8.30

Using the data in Exercise 8.29, calculate the average level for a joint venture capital, the average annual absolute increase (by two methods), the average annual growth rate (using two methods), and the average annual rate of increase.

Exercise 8.31

Using the dynamic characteristics of interaction, determine row levels to describe a country's export dynamics of goods and services for six years and the absolute and relative rate of their growth.

Year	Export volume, $ million	Chain characteristics			
		Absolute increase, $ million	Growth rate, %	Rate of increase, %	Absolute value of 1% increase, $ million
1	19.8	x	x	x	x
2				11.1	
3		5.3			
4			139.2		
5					
6				13.6	0.404

Exercise 8.32

Using the data in Exercise 8.31, calculate the average level of goods and service exports, the average annual absolute increase (using two methods), the average annual growth rate (using two methods), and the average annual rate of increase.

Exercise 8.33

Using the data below on the dynamics of the average exchange rate of the national currency established by the central bank, construct a trend model using a straight line and a parabola of the second order. Estimate the adequacy of each model and check its predictive potential. Build exact and interval predictions for the next two years.

Year	2000	2001	2002	2003	2004	2005	2006
1 English pound	8.2499	7.7394	7.9984	8.7128	9.7391	9.3376	9.2945
1 US dollar	5.4402	5.3721	5.3266	5.3327	5.3192	5.1247	5.0500
10 Russian rubles	1.94	1.84	1.70	1.74	1.85	1.81	1.86
1 euro	5.0289	4.8136	5.0301	6.0244	6.6094	6.3899	6.3369

Exercise 8.34

Using the data below on the average monthly nominal wage (euro) and indices (percentage comparison to a corresponding period of the previous year), determine seasonality indices for two years, compare the level of seasonal fluctuation, and construct lines of seasonality. Using the exponential mean, predict the levels of remuneration for the next year ($\boldsymbol{\alpha = 0.3}$).

Month	Nominal wage	Nominal wage index	Real wage index
January	865	135.0	122.8
February	905	135.7	122.4
March	987	136.7	125.5
April	984	134.1	124.4
May	1003	131.2	121.8
June	1064	129.2	120.6
July	1079	128.8	119.4
August	1073	129.1	119.7
September	1087	126.9	116.0
October	1088	123.3	111.0
November	1104	123.1	109.8
December	1277	125.2	111.7

Exercise 8.35

Using the data below on the share of a country's population with average monthly expenses per capita that are lower than the cost of living (%) by aggregate and monetary income over 7 years, construct a trend model using a straight line and a parabola of the second order. Estimate the adequacy of each model and check its predictive potential. Build point and interval predictions for the next two years.

Year	Share of population with average monthly expenses per capita that are lower than the cost of living, %	
	money	aggregate
1	87.9	80.2
2	89.0	82.7
3	88.4	83.3
4	83.4	76.2
5	73.8	65.6
6	64.0	55.3
7	59.4	50.9

Exercise 8.36

Using the production data below, estimate the cyclical fluctuations and determine the contribution/effect of each harmonic on the fluctuation process using the apparatus of a harmonic analysis. Build lines of seasonality.

Month	Production and distribution of electrical energy, gas, and water, million euro	Foodstuff, beverage, and tobacco production, million euro	Meat production, ton
January	8673	4968	9384
February	8793	5196	7994
March	17006	11170	9978
April	15533	11420	5061
May	22967	18055	3922
June	21377	18767	8210
July	28917	25587	5201
August	27413	26507	8723
September	34856	33285	13015
October	34557	35122	14289
November	44021	41843	12992
December	44801	44044	8056

9. INDICES

9.1. The concept of indices: individual and summary indices

An index is a relative value indicating the change in the level of an event over time, in space, or in comparison to a plan (norm or standard). For example, in January an associate professor's salary increased 1.48 times compared to December.

Depending on the purpose of the comparison, indices can be classified into three groups:

1. Dynamic indices indicate the change in an event over time.
2. Spatial indices capture the change in an event in space.
3. Norm indices (standard) measure the change in an event in comparison to a normative (standard or pattern) level.

The dynamic index for the unit price of a commodity is calculated as $i_p = \frac{P_1}{P_0}$.

The spatial index of a price can be given by $i_p = \frac{P_A}{P_B}$.

A norm index can be shown for the example of a plan fulfillment (p_{pl}) as follows $i_q = \frac{p_1}{p_{pl}}$.

Depending on the coverage of population units, we can use individual and summary indices.

Individual indices characterize the change in a single event (e.g., coal production in a mine or the price of some goods).

Summary indices characterize the change in the level of an indicator involving a population. A population may consist of uniform and non-uniform elements.

For example, in the first case we could have some data on the coal production of several mines, the yield capacity of different (i.e. uniform in a certain sense) crops, or the prices of potatoes charged by different sellers. An example of the second case could be the production volume of different kinds of products made by one or several companies or the prices of various commodities in a city. In the first case, we could calculate the average level for a population. A change in average levels is represented by ***a summary index of average values***; for the second case, ***a summary aggregated index*** is used. If the change in an event is studied for more than two periods, then each of them is denoted with a number, "0", "1", "2", "3" etc., respectively. If the level of a previous period is taken as a basis for comparison, then the dynamic indices are called ***chain*** indices; when the initial level is one and the same, we describe them as ***basic*** indices.

Example 9.1

Let us denote the production volume as q_0 for the year 2000, q_1 for 2001, and q_6 for 2006. Thus we get $i_{06/00} = \frac{q_1}{q_0} \times \frac{q_2}{q_1} \times \frac{q_3}{q_2} \times \frac{q_4}{q_3} \times \frac{q_5}{q_4} \times \frac{q_6}{q_5} = \frac{q_6}{q_0}$.

There is a general rule to estimate an interconnection for indices – indices are interconnected as well as their absolute values. For instance, if yield capacity (YC) increased by 1.2 times and the size of the cultivated area (CA) decreased by 10 %, the gross yield (GY) can be calculated as the product of yield capacity and cultivated area, i.e. GY = YC × CA.

Here $i_{GY} = i_{YC} \times i_{CA} = 1.2 \times 0.9 = 1.08$.

Therefore, gross yield increased by 8 %. Thus, we can determine the change in gross yield.

9.2. System of aggregated indices

Summary aggregated indices are used in the case of a non-uniform population. If a comparable commodity is sold/marketed (e.g. fruits and vegetables), then the index of physical gross output is $\boldsymbol{I_q = \frac{\sum q_1}{\sum q_0}}$.

If different goods (that are non-comparable) are sold, and their physical volume can be measured with different measurement units (e.g. in kg, pieces, and liters), as well as similar measurements units, then comparison of the physical volumes of these marketed goods is not relevant, e.g. the

sale by volume of foodstuffs and non-food products. As such, the general index of a physical volume cannot be $I_q = \frac{\sum q_1}{\sum q_0}$.

This means we have to adjust non-uniform commodities to comparable forms. In such cases, we use price as the value and thus a summary index of a physical volume would be $I_q = \frac{\sum pq_1}{\sum pq_0}$.

However, the question remains as to what level prices should be fixed? i.e., at current or at base prices? The sales volume (as turnover or proceeds from sales) in the current period compared to the base one can change as a result of two factors: a change in physical sales volume (for one or different kinds of goods) and a change in prices. Clearly, one of these two factors, or both together, can influence such a change. Moreover, the effect can be in the same direction or in opposite directions. To estimate the effect of all factors and of each of the factors separately, a system of indices is applied. If we fix prices at a base level, then the value $\sum p_0 q_1$ in the numerator position represents proceeds from sales in the current period at comparable prices – such an indicator has an economic meaning and can be interpreted. A system of weighting like this at a base level is called a ***Laspeyres system*** and the index of physical volume $I_q = \frac{\sum p_0 q_1}{\sum p_0 q_0}$ is called a ***summary aggregated index of physical turnover***.

When the price changes, a summary aggregated price index of turnover is calculated and the physical volume in this index system is fixed at its current level following the Paasche system. Thus the price index is $I_p = \frac{\sum p_1 q_1}{\sum p_0 q_1}$.

There is an interconnection between the index of physical volume and the price index $I_{pq} = I_q \times I_p = \frac{\sum p_0 q_1}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$ (9.1).

Indices of this type are described as aggregated since their numerator and denominator are aggregates, i.e. values that have an economic meaning.

So, $\sum p_0 q_1$ concerns the sales volume of the current period at base period prices, that is to say, at fixed prices; $\sum p_0 q_0$ is the sales volume of the base period; and $\sum p_1 q_1$ is the sales volume of the current period.

From the aggregated indices, we can calculate absolute changes in the proceeds from sales as a whole and from the effect of different factors, in terms of prices and physical volume as the difference between the numerator and the denominator of the corresponding formula.

An absolute change due to a price factor is calculated with the formula

$$\Delta_p = \sum p_1 q_1 - \sum p_0 q_1.$$

An absolute change due to a physical volume uses the formula

$$\Delta_q = \sum p_0 q_1 - \sum p_0 q_0.$$

A total absolute change due to both factors is calculated with the formula

$$\Delta_{pq} = \sum p_1 q_1 - \sum p_0 q_0 = (\sum p_0 q_1 - \sum p_0 q_0) + (\sum p_1 q_1 - \sum p_0 q_1).$$

Thus, there is also an interconnection between absolute changes. This makes it possible to estimate the share of the effect of each factor on the general change in proceeds from sales. A share of the effect of a price factor can be determined with the formula $d_{\Delta_p} = \frac{|\Delta_p|}{|\Delta_p| + |\Delta_q|} \times 100\%$.

The share of the second factor is 100 - d_{Δ_p}. Absolute changes should be taken by module because the effect of the factors can be multidirectional.

Example 9.2

From the data on sales of foodstuffs, create a summary index for price, physical volume, and turnover, as well as the absolute increase in turnover due to these factors. Present the calculations in the table. Analyze the results.

Commodity	Sales, ton		Price for 1 kg, euro	
	Base (q_0)	Accounting (q_1)	Base (p_0)	Accounting (p_1)
A	15	10	10	12
B	50	45	25	35
Total	65	55	x	x

Let us build an additional computational table.

Commodity	Sales proceeds, thousand euro		
	p_0q_0	p_1q_1	p_0q_1
A	10×15=150	12×10=120	10×10=100
B	25×50=1250	35×45=1575	25×45=1125
Total	1400	1695	1225

The summary turnover index is equal to $I_{pq} = \frac{\sum p_1q_1}{\sum p_0q_0} = \frac{1695}{1400} = 1.2107$ (21.07%).

The absolute general change in sales proceeds is

$$\Delta_{pq} = \sum p_1q_1 - \sum p_0q_0 = 1695 - 1400 = +295$$

thousand euro.

Thus, proceeds from sales increased by 21.07 % amounting to 295 thousand euro.

Let us analyze the effect of each factor:

a) Physical volume

$I_q = \frac{\sum p_0q_1}{\sum p_0q_0} = \frac{1225}{1400} = 0.875$ (-12.5%), the absolute change in sales proceeds due to the factor of physical volume is

$\Delta_q = \sum p_0q_1 - \sum p_0q_0 = 1225 - 1400 = -175$ thousand euro.

The physical volume of sales decreased by 12.5 % on average, resulting in losses of 175 thousand euro.

b) Price

$I_p = \frac{\sum p_1q_1}{\sum p_0q_1} = \frac{1695}{1225} = 1.3837$ (+38.4%), the absolute change in proceeds from sales due to the factor of price is

$\Delta_p = \sum p_1q_1 - \sum p_0q_1 = 1695 - 1225 = +470$ thousand euro.

Goods prices increased by 38.4 % on average, resulting in an increase in sales of 470 thousand euro.

Let us check the interconnection of indices and absolute changes:

$I_{pq} = I_q \times I_p = 0.875 \times 1.3837 = 1.2107,$

$\Delta_{pq} = \Delta_p + \Delta_q = 470 + (-175) = +295$ thousand euro.

Let us estimate the share of effect of each factor:

$d_{\Delta_p} = \frac{|\Delta_p|}{|\Delta_p|+|\Delta_q|} \times 100\% = \frac{470}{470+|-175|} \times 100\% = 73\%,$

$d_{\Delta_q} = 100 - d_{\Delta_p} = 100 - 73 = 27\%.$

Thus, we can see that more than two-thirds of the general change in proceeds from sales occurred due to a rise in price, resulting in a decrease in demand. In turn, losses were 175 thousand euro; however, these losses were totally compensated for by an increase in the proceeds from sales from the price rise. The value of proceeds from sales increased by 295 thousand euro.

9.3. Weighted average arithmetic and harmonic indices

Aggregated indices are computed from individual indices. For example, to calculate an index of physical volume when there are no individual data about prices and physical volume, but the availability of individual indices of physical volume, we use substitution, i.e. we substitute $\sum \boldsymbol{p_0 q_1}$ for $\sum \boldsymbol{i_q p_0 q_0}$, as $\boldsymbol{q_1 = i_q \times q_0}$.

Here, we use the following formula for the index of physical volume

$I_q = \frac{\sum i_q p_0 q_0}{\sum p_0 q_0}$ (9.2).

Such an index is called ***the weighted average arithmetic index of a physical volume***.

A weighted average index is an average of individual indices, weighted on aggregates having defined dimensions and a similar level of fixation. Aggregates can be of cost $\sum p_i q_i$ (turnover) indicators, $\sum c_i q_i$ (productive inputs), or labor $\sum t_i q_i$ (hours) indicators.

A similar replacement can be made for an aggregated price index, if $\sum p_0 q_1$ is substituted for $\frac{\sum p_1 q_1}{i_p}$, as $p_0 = \frac{p_1}{i_p}$. Then, the formula of the price index is $I_p = \frac{\sum p_1 q_1}{\sum \frac{p_1 q_1}{i_p}}$ (9.3).

Such an index is called ***the weighted average harmonic price index***.

Example 9.3

Below, we have some data on the volume of consumption of some foodstuffs and the price indices of some commodity groups.

Commodity groups	Consumption volume in current prices, euro		Price indices, i_p	$\frac{p_1 q_1}{i_p}$
	I quarter $p_0 q_0$	II quarter $p_1 q_1$		
Meat products	320000	315000	0.90	350
Dairy products	28000	26530	0.95	27926
Bakery products	32000	32817	0.98	33487
Total	380000	374527	x	411413

We can calculate the index of the level of consumption per capita, if the population increased by 4 %. The consumption level per capita is calculated as a relation of the summary volume of consumption on the population number $= \frac{q}{T}$. So, $I_\partial = \frac{I_q}{I_T}$.

Accordingly, $I_T = 1.04$. Now, we can calculate I_q.

$I_q = \frac{\sum (p_1 q_1)/ i_p}{p_0 q_0} = \frac{411413}{380000} = 1.081$ (108.1%), then $I_\partial = \frac{1.081}{1.040} = 1.038$ (103.8%).

Thus, the level of consumption per capita increased on average by 3.8 %, which led to an increase in general expenses of 31413 euro (411413–380000).

9.4. Indices of average values

Along with the need to estimate the dynamics of an indicator (e.g., a price level) for a company, statistics is used to study the dynamics of a population indicator (e.g., the price level of an individual commodity in a group of traded companies). As such, it is necessary to calculate the average level of an indicator and estimate its dynamics. If the dynamics of average indicator levels are studied, then a system of summary indices of average values are used.

We already know that two factors influence the value of an average level:

- a change in the value of an indicator itself;
- a change in the population structure, namely, the distribution of a population unit by an indicator.

Thus, a system of an average value indices includes three indices:

1. An index of the variable composition characterizing the dynamics of an average level from the effect of two factors – both the average change in an indicator and the change in a population's structure.

An index of a variable composition is written as $I_{\bar{x}}^{VC} = \frac{\bar{x}_1}{\bar{x}_0}$.

If we take into consideration that $\bar{x} = \frac{\sum xf}{\sum f}$, then the formula of an index of variable composition is $I_{\bar{x}}^{VC} = \frac{\sum x_1 f_1}{\sum f_1} \div \frac{\sum x_0 f_0}{\sum f_0}$.

A difference between two fractions reflects an absolute change in the average indicator level $\Delta\bar{x} = \frac{\sum x_1 f_1}{\sum f_1} - \frac{\sum x_0 f_0}{\sum f_0}$.

If we say that $\frac{f}{\sum f}$ indicates a structure, then denoting this ratio as *d*, we can write the index of variable composition as $I_{\bar{x}}^{VC} = \frac{\sum x_1 d_1}{\sum x_0 d_0}$, and the absolute change as $\Delta\bar{x} = \sum x_1 d_1 - \sum x_0 d_0$.

2. An index of fixed composition allows us to estimate the average dynamics of a level due to a change in the values of an indicator itself (x) in a fixed structure and at a current level following the Paasche

system. An index of fixed composition is written as $I_{\bar{x}}^{FC} = \frac{\sum x_1 f_1}{\sum f_1} \div \frac{\sum x_0 f_1}{\sum f_1} = \frac{\sum x_1 f_1}{\sum x_0 f_1}$.

Thus, an index of fixed composition is, in fact, an aggregated index. The difference between the numerator and denominator of this index indicates an absolute aggregated change from the effect of the change in an indicator, on average $\Delta_x = \sum x_1 f_1 - \sum x_0 f_1$.

Similarly, we can write an index of fixed composition in fractions $I_{\bar{x}}^{FC} = \frac{\sum x_1 d_1}{\sum x_0 d_1}$, and the absolute change as $\Delta_x = \sum x_1 d_1 - \sum x_0 d_1$.

3. An index of structural change indicates a change in the average level caused by the effect of changes in a population's structure $\frac{f}{\sum f}$ for a fixed value of an indicator (x) at a base level following the Laspeyres system. An index of structural changes can be written $I_{\bar{x}}^{SC} = \frac{\sum x_0 f_1}{\sum f_1} \div \frac{\sum x_0 f_0}{\sum f_0}$.

An index of structural change can also be written $I_{\bar{x}}^{SC} = \frac{\sum x_0 f_1}{\sum x_0 f_0} \div \frac{\sum f_1}{\sum f_0}$, and thus the structural change index is formed from two indices: an aggregated index of physical volume I_f and a general index of physical volume I^f.

A difference between two fractions represents the absolute change in an average indicator level caused by the effect of a structural component $\Delta_d = \frac{\sum x_0 f_1}{\sum f_1} - \frac{\sum x_0 f_0}{\sum f_0}$.

Similarly, we can write an index of structural change in fractions $I_{\bar{x}}^{sc} = \frac{\sum x_0 d_1}{\sum x_0 d_0}$, and the absolute change as $\Delta_d = \sum x_0 d_1 - \sum x_0 d_0$.

Thus, an index model characterizing the interconnection between the considered indices can be a two or three-factor model. A two-factor index model includes an index of fixed composition and an index of structural change. It is written as a product of these indices $I_{\bar{x}}^{VC} = I_{\bar{x}}^{FC} \times I_{\bar{x}}^{SC}$.

A three-factor index model includes two aggregated indices and one general index of physical volume. It is written $I_{\bar{x}}^{VC} = I_x \times I_f \div I^f$.

Let us consider an index system constructed using data on the average yield capacity of wheat in winter and in spring. Yield capacity (y) is found from the gross yield of a crop (GY) per unit (S) area. Then, the yield capacity of a different crop is calculated with the formula $YC = \frac{GY}{S}$, and the average yield capacity is $\bar{y} = \frac{\sum GY}{\sum S}$, $GY = y \times S$.

If gross yield is the product of yield capacity and cultivation area, then the formula of average yield capacity is $\bar{y} = \frac{\sum y \times S}{\sum S}$.

We have learned how to write an index system for the average values of a variable, a fixed composition, and a structural change. Now, let us consider an example of the use of an index system for average yield capacity.

Example 9.4

Below are some data on two years of wheat production in Ukraine.

Wheat cultivar/ type	Yield capacity, centner/hectare		Cultivation area, thousand hectares	
	1st year (y_0)	2nd year (y_1)	1st year (s_0)	2nd year (s_1)
Winter	38.0	31.0	7000	5205
Spring	27.5	25.0	20	55
Total	x	x	7020	5260

Calculate: (1) an index of variable composition for average yield capacity; (2) an index of fixed composition and structural change for average yield capacity; (3) the absolute change in gross yield due to the change in cultivation of each type of wheat; and (4) check the interconnection of the calculated indices. Present the calculations in a table and analyze the results.

An index of the variable composition of average yield capacity is

$$I_{\bar{y}}^{VC} = \frac{\sum y_1 \times s_1}{\sum s_1} \div \frac{\sum y_0 \times s_0}{\sum s_0}.$$

To calculate it, we build an additional computational table.

Wheat cultivation type	Gross yield, thousand centner		
	$y_0\ s_0$	$y_1\ s_1$	$y_0\ s_1$
Winter	38×7000=266000	31×5205=161355	38×5205=197790.0
Spring	27.5×20=550	25×55=1375	27.5×55=1512.5
Total	266550	162730	199302.5

We use the formula for an index of variable composition

$I_{\bar{y}}^{VC} = \frac{162730}{5260} \div \frac{266550}{7020} = \frac{30.94}{37.97} = 0.815$ (-18.5%).

As we can see, yield capacity decreased by 18.5 %, which was 7 centner/hectare (30.94-37.97).

Let us determine the factor of yield capacity of winter and spring wheat that influenced the decrease in average yield capacity.

We calculate the index of fixed composition as $I_{\bar{y}}^{FC} = \frac{\sum y_1\ s_1}{\sum y_0\ s_1} = \frac{162730}{199302.5} =$ 0.816 (-18.4%).

The absolute variable is $\Delta_y = \sum y_1\ s_1 - \sum y_0\ s_1$ = 162730 – 199302.5 = -36572.5 thousand centner.

Thus, the average decrease in yield capacity for both winter and spring crops of 18.4 % led to a gross yield loss of 3 million 657 thousand 250 ton (3657250 ton).

To estimate the effect of the structural component, we build an index of structural change for average yield capacity $I_{\bar{y}}^{SC} = \frac{\sum y_0 s_1}{\sum s_1} \div \frac{\sum y_0 s_1}{\sum s_1}$.

Then, using the data from the computational table

$$I_{\bar{y}}^{SC} = \frac{199302.5}{5260} \div 37.97 = 0.9979 \text{ (-0.21\%)}.$$

So, serious structural changes in cultivation area did not take place. Average yield capacity decreased by 0.21 % from the structural component and, thus,

the interconnection between the calculated indices is 0.816 × 0.9979 = 0.814.

Example 9.5

We have some data on the performance of two mines incorporated into a trust over two years. We can calculate the individual and summary indices of labor productivity, the total general increase in coal production, and the effect of some factors.

Mine №	Coal production, thousand ton		Worked thousand man-days		Labor productivity, ton/man-days		Share of labor hours	
	1st year q_0	2nd year q_1	1st year T_0	2nd year T_1	1st year w_0	2nd year w_1	1st year d_0	2nd year d_1
A	1	2	3	4	5=1/3	6=2/4	7	8
1	40	88.0	20	40	2.0	2.2	20/50=0.4	0.67
2	45	30.6	30	20	1.5	1.53	30/50=0.6	0.33
Total	85	118.6	50	60	85/50=1.7	118.6/60=1.977	1.0	1.00

We complete the table (columns 5-8) and calculate a summary line. Labor productivity is given by coal production per time unit. For mine № 1: $\boldsymbol{w_1}$ = 88:40 = 2.2 (ton/man-days) and $\boldsymbol{w_0}$ = 40:20 = 2.0 (ton/man-days).

So, the individual index of labor productivity for this mine is 1.1 (2.2/2.0). For mine № 2 it is 1.02 (1.53/1.5).

Thus, labor productivity for the first mine increased by 1.1 times, or 10 %. We describe this index as ***individual*** – it belongs to ***a population unit***, i.e. mine № 1. However, it is also clear that it reflects a change in the ***average***

labor productivity of all miners at mine № 1, that is, ***a population***. The same is true for the second mine. If labor productivity increased by 10 % at one mine and by 2 % at the other mine, this does not necessarily suggest a similar increase for the trust in general.

We calculate the summary index of an average labor productivity as

$$I_{\bar{w}}^{VC} = \frac{\bar{w}_1}{\bar{w}_0} = \frac{\sum q_1}{\sum T_1} \div \frac{\sum q_0}{\sum T_0} = \frac{118.6}{60} \div \frac{85}{50} = \frac{1.977}{1.7} = 1.163 \text{ (+16.3\%)}.$$

Thus, labor productivity in the trust increased by 16.3 %, which is more than the best figure at mine № 1. Such examples are numerous. For instance, despite the fact that one part of a collective (the staff) earned the same amount of money in May as in April, and the other cohort of workers received less money, the average wage increased. Thus, an index of variable composition can go beyond individual indices. In our example, we can see the change in average labor productivity due to changes in productivity for each mine and also due to changes in the structure of labor hours. To study the effect of the first component, we calculate an index of fixed composition

$$I_{\bar{w}}^{FC} = \frac{\sum w_1 T_1}{\sum w_0 T_1} = \frac{118.6}{2.0\ x\ 40 + 1.5\ x\ 20} = \frac{118.6}{110} = 1.078 \text{ (+7.8\%)}.$$

Contrary to an index of variable composition, an index of fixed composition never goes beyond individual indices. So, on average, labor productivity increased by 7.8 %.

In looking at the table of the computed data, it is easy to see that:

1) labor productivity at the first mine was higher in each year;

2) the share of working hours at the first mine (and its greater efficiency than the second mine) increased significantly (67 % against 33 %).

These changes in the structure of working hours had a positive effect on the change in average labor productivity

$$I_{\bar{w}}^{SC} = \frac{\sum w_0 T_1}{\sum T_1} \div \frac{\sum w_0 T_0}{\sum T_0} = \frac{I_{\bar{y}}^{VC}}{I_{\bar{y}}^{FC}} = \frac{1.163}{1.078} = 1.079 \text{ (+7.9\%)}.$$

Now, we determine the general change in coal production and the specific change caused by certain factors

$$\Delta_q = \sum q_1 - \sum q_0 = 118.6 = 33.6 \text{ thousand ton.}$$

Factors influencing the change in general coal production can be different. We can concentrate on two of them here: labor productivity and total labor hours.

$\Delta q_w = (\overline{w}_1 - \overline{w}_0) \sum T_1 = (1.977 - 1.7)$ x $60 = 16.6$ thousand ton;

$\Delta q_T = (\sum T_1 - \sum T_0)\overline{w}_1 = (60 - 50)$ x $1.7 = 17$ thousand ton.

We can confirm this with $\boldsymbol{\Delta q = \Delta q_w + \Delta q_T = 16.6 + 17 = 33.6}$ thousand ton.

9.5. Spatial indices

Spatial indices are a type of indices of average values. We can compare average indicator levels of different objects, such as regions or countries, using spatial indices. The formulation of spatial indices follows certain criteria:

- it requires us to explain the region (object) that is to be considered as a base for comparison;
- the identification of the order of fixation of the indicator values x_j and structural components d_j is of extreme importance.

The basis for comparison can be freely chosen (depending on the purposes of the comparison), but the same level has to be used – either an average level or a standard value. An average value of a feature is calculated from a vertical structure as the weighted arithmetic mean; an average value of a fraction is calculated from a horizontal distribution, also as the weighted arithmetic mean of two objects (territories).

Let us understand the rule and guidelines for the formulation of spatial indices of variable composition, fixed composition, and structural change.

An index of variable composition has a classical computation formula and deals with the number of times an average indicator level of object A is larger than the average value of object B

$$I\bar{x} = \frac{\sum x_A f_A}{\sum f_A} \div \frac{\sum x_B f_B}{\sum f_B} = \frac{\sum x_A d_A}{\sum x_B d_B}.$$

The use of an object with a smaller indicator level as a basis for comparison makes the interpretation of this index convenient. However, this is not a rule to be followed. It can also be interpreted in a different way. For example, if

an index of variable composition is 0.9 then, an average level in region A is 90 % of an average level in region B.

An index of fixed composition is built depending on whether object structures are comparable, because a structural fixation has to be determined. If the structures are comparable, then either object A or object B can be chosen for fixation. The distribution of fractions may take the following form.

Elements	Structure of object A, %	Structure of object B, %
1	25	20
2	75	80
Total	100	100

This is a comparable structure. The formula of an index of fixed composition is

$$I_{\bar{x}} \left| \begin{matrix} \frac{\sum x_A f_A}{\sum f_A} \div \frac{\sum x_B f_A}{\sum f_A} = \frac{\sum x_A d_A}{\sum x_B d_A} \\ \frac{\sum x_A f_B}{\sum f_B} \div \frac{\sum x_B f_B}{\sum f_B} = \frac{\sum x_A d_B}{\sum x_B d_B} \end{matrix} \right. .$$

If the structures are not comparable, then the distribution of fractions appears as

Elements	Structure of object A, %	Structure of object B, %
1	25	50
2	75	50
Total	100	100

Such a structure is described as *incomparable*. Then, the formula of an index of fixed composition is

$$I_{\bar{x}} = \left[\begin{array}{l} \dfrac{\sum x_A f_{AB}}{\sum f_{AB}} \div \dfrac{\sum x_B f_{AB}}{\sum f_{AB}} = \dfrac{\sum x_A d_{AB}}{\sum x_B d_{AB}} \\ \dfrac{\sum x_A d_{st}}{\sum x_B d_{st}} \end{array} \right. ,$$

where f_{AB} is the sum of the frequencies of two objects (territories) and d_{AB} is the structure of two regions and is calculated with the formula $d_{AB} = \frac{f_A + f_B}{\Sigma(f_A + f_B)}$.

An index of structural change is built through the fixation of an average level for two objects (territories) as regards different elements and characterizes a relationship of average levels in the fixed structure of the elements

$$I_{\bar{x}} = \frac{\sum \bar{x}_{AB} f_A}{\sum f_A} \div \frac{\sum \bar{x}_{AB} f_B}{\sum f_B} = \frac{\sum \bar{x}_{AB} d_A}{\sum \bar{x}_{AB} d_B},$$

$$\bar{x}_{AB} = \frac{x_A d_A + x_B d_B}{d_A + d_B}.$$

It is important to note that there is no interaction between the spatial indices of variable composition, fixed composition, and structural change. Let us better understand the index system with the following example.

Example 9.6

Data on the efficiency of vegetable production in two regions is presented below.

Production technology	Gross yield, thousand ton		Production cost of 1 ton, euro		
	Region A (f_A)	Region B (f_B)	Region A (x_A)	Region B (x_B)	On the average by two regions ($\bar{x}_{AB}$)
Intensive	90	210	130	110	(130×90+110×210)/ (210+90)=116
Traditional	210	140	150	160	(150×210+160×140)/ (210+140)=154
Total	300	350	x	x	x

Determine spatial indices for the average production cost of variable composition, fixed composition, and structural change. Make some conclusions and build an additional computational table.

d_A	d_B	d_{AB}	$x_A d_A$	$x_B d_B$	$x_A d_{AB}$	$x_B d_{AB}$	$\bar{x}_{AB} d_A$	$\bar{x}_{AB} d_B$
90/300 =0.3	0.6	(90+210)/ (300+350) =0.46	130× 0.3	110× 0.6	130× 0.46	110× 0.46	116× 0.3	116×0.6
210/300 =0.7	0.4	(210+140)/ (300+350) =0.54	150× 0.7	160× 0.4	150× 0.54	160× 0.54	154× 0.7	154×0.4
1.0	1.0	1.00	144	130	140.8	137	142.6	131.2

We can see from the table that the structures of gross yield by intensive and traditional technologies are comparable – a significant specific weight (60 %) of total production in region B is achieved using intensive technology, whereas only 30 % of vegetables are grown using intensive technology in region A. This leads to the assumption that the cost of production is lower

in region B and it is prudent to take it as the basis for comparison. An index of variable composition is calculated with the formula

$I_x = \frac{\sum x_A d_A}{\sum x_B d_B} = \frac{144}{130} = 1.108$, i.e. by more than 10.8 %.

Thus, the average cost of vegetable production in region A is higher by 10.8 % than in region B, amounting to 14 euro (144-130). Since the structures are incomparable, an index of fixed composition of average production cost is calculated with the formula

$I_{\bar{x}} = \frac{\sum x_A d_{AB}}{\sum x_B d_{AB}} = \frac{140.8}{137} = 1.028$, i.e. by more than 2.8 %.

This establishes that the average production cost in region A increased by 2.8 % when intensive technology was applied.

The index of structural change shows that the production cost in region A is 8.7 % higher than in region B due to the fact that vegetable production relies on the use of traditional technology

$I_{\bar{x}} = \frac{\sum \bar{x}_{AB} d_A}{\sum \bar{x}_{AB} d_B} = \frac{142.6}{131.2} = 1.087$, i.e. by more than 8.7 %.

Thus, the high production cost in region A is primarily attributable to the factor of production structure, without the use of intensive technology for vegetable production.

9.6. System of interdependent indices: factor index analysis

Index models are widely used for studying functional connections between results and the factors that define them. An index model is a system of interdependent indices that allows us to measure the value of degree and the absolute value of the effect of the factors included in the model on the change seen in the result. The degree of effect is characterized by an index (a growth coefficient), and the absolute value of the effect is characterized by an increase in the simulated indicator in absolute units of measurement. The base for the formulation of an index model of any qualitative indicator is a chain system of communication where a simulated indicator is viewed as a function of a definite set of factors and a multiplicative model is built on it $= \prod_{j=1}^{m} x_j$ (9.4).

It can be useful to analyze the level of interconnection between factors using simple static and dynamic stochastic models to choose the factors that should be included in the model. This makes it possible to take into account the most significant factors that define the level of an effective indicator.

Analysis of the effect of some factors on the dynamics of effective indicators is done using an index system. The results of analysis largely depend on the sequence by which the factors are included in the system and on the priority of their study. The basic requirement for an index analysis is to set up a clearly defined sequence of factors aligned with the importance of their economic essence, interconnections, and the calculation procedure. This is achieved by the extension of an initial two-factor model.

Such a model must have definite limits for a new factor being added and capture detailed reasons for effective change in a feature, but does not influence the change itself. It has to be specified that a steep increase in factorial space in most cases makes analysis more complicated, although it does not expand the analysis for the effective feature. When this type of model is used, it is practical to have three or four components that influence the result directly and significantly and define its absolute change.

An index system is built on the consequent change of "0" for "1" in the positions of numerator and denominator. As an example, we can examine how to build an index of the input cost of manufacturing different items if each of them is produced from one definite material.

Let us denote α as the number of items of one type; b as the material consumption per product unit; c as the unit price of the material; and y as the general material inputs. This simplified denotation of values with α, b, c in index factor analysis is quite common in the literature. More precisely, a general cost can be denoted $R = \sum \alpha \times b \times c$.

Let us imagine $I_{abc} = I_a \times I_b \times I_c = \frac{a_1 b_1 c_1}{a_0 b_0 c_0} = \frac{a_1 b_1 c_1}{a_0 b_1 c_1} \times \frac{a_0 b_1 c_1}{a_0 b_0 c_1} \times \frac{a_0 b_0 c_1}{a_0 b_0 c_0}$ (9.5).

It is not difficult to observe the rationale in the shift from (9.1) to (9.5). Absolute changes in the result due to particular factors are determined with the following formulas

$$\Delta y = \sum a_1 b_1 c_1 - \sum a_0 b_0 c_0 = \Delta y(a) + \Delta y(b) + \Delta y(c),$$

$$\Delta y(a) = \sum a_1 b_0 c_0 - \sum a_0 b_0 c_0 = \sum (a_1 - a_0) b_0 c_0,$$

$$\Delta y(b) = \sum a_1 b_0 c_0 - \sum a_1 b_0 c_0 = \sum (b_1 - b_0) a_1 c_0,$$

$$\Delta y(c) = \sum a_1 b_1 c_1 - \sum a_1 b_1 c_0 = \sum (c_1 - c_0) a_1 b_1.$$

Thus, the basis for the simulation of a multifactor index model of investment efficiency is the indicator of the return on invested capital. The index of investment profitability (I_R) is the product of the index of the return on invested capital (I_f) and the index of the investment level of fixed capital in the total amount of investment funds I_d (Eq. 9.7– Eq. 9.8).

The return from invested capital (f) can be determined by profitability of production (r) and the return from invested fixed capital (V): $f = r \times V$. So, the index of investment profitability is already the product of three factor-multipliers. The system of the indices of investment profitability includes three partial indices $I_R = I_r \times I_V \times I_d = \frac{\sum r_1 V_1 d_1}{\sum r_0 V_0 d_0}$ (9.6).

In building a partial index of each factor, we need to eliminate the effect of other factors included in the model to measure the degree and absolute scope of the effect of each factor on the dynamics of the simulated efficiency indicator. In our three-factor model of the dynamics of the index of investment profitability, this principle is as follows

$I_r = \frac{\sum r_1 V_1 d_1}{\sum r_0 V_1 d_1}$ (9.7);

$I_V = \frac{\sum r_0 V_1 d_1}{\sum r_0 V_0 d_1}$ (9.8);

$I_d = \frac{\sum r_0 V_0 d_1}{\sum r_0 V_0 d_0}$ (9.9).

The absolute change in a simulated indicator can be analyzed using factors with the index model. When there are too many factors, the calculation of absolute differences of the simulated indicator becomes bulky. As such, a compact method of calculation is suggested in the literature, which relies on the base level of a simulated indicator adjusted to the indices included in the model of factors. We suggest not using large index models when undertaking index analysis; it is advisable to choose only those factors that can influence a change in the level of an effective indicator.

Certain forms of indices have a property that is required to form an index system – the qualitative uniformity of the indicators that form it. However, combining heterogeneous indicators is often a necessity. Heterogeneous indicators have estimated values and represent a technological and

productive structure of investments in fixed capital within a single system. It is possible to solve the problem of switching from absolute (investment volume in fixed capital; cost of construction and installation work; investment cost in machines, equipment, tools, and implementation; investment in new construction; and expansion of technology through re-equipping and reconstruction) to relative values of comparison, structure, and coordination. To analyze the effect of factors on the change in the productive structure of investments in fixed capital, the following indicators are introduced:

1. The share of expenses for technical re-equipping and reconstruction in the total amount of investments in fixed capital (k_1). An increase in this coefficient improves the productive structure because priority is given to the intensification of investment activity, rather than to its extensive development (new construction and expansion).
2. The share of expenses for the purchase of machines, equipment, instruments, and implementation in the total amount of investments in fixed capital (k_2). An increase in this coefficient improves the productive structure and characterizes the intensive development of the investment process.
3. Coordination between expenses for the purchase of machines, equipment, instruments, implementation, and expenses for construction and installation work (k_3). An increase in this coefficient improves the technological structure of investment in fixed capital and characterizes the intensive development of the investment process.
4. The correlation of expenses for construction and installation work and on new construction and expansion (k_4). An increase in this coefficient worsens the productive structure of fixed capital and results in an extensive investment process.
5. The coordination of expenses for technical re-equipping and reconstruction and for new construction and expansion (k). An increase in this coefficient improves the productive structure of investments in fixed capital and characterizes the intensive development of the investment process.

Thus, we can simulate a four-factor index system $I_k = I_{k_1} \times I_{k_2} \times I_{k_3} \times I_{k_4}$ (9.9), where

$$I_k = \frac{\sum k_1^1 k_2^1 k_3^1 k_4^1}{\sum k_1^0 k_2^0 k_3^0 k_4^0} \qquad (9.10)$$

$$I_{k_1} = \frac{\sum k_1^1 k_2^1 k_3^1 k_4^1}{\sum k_1^0 k_2^1 k_3^1 k_4^1} \quad (9.11)$$

$$I_{k_2} = \frac{\sum k_1^0 k_2^1 k_3^1 k_4^1}{\sum k_1^0 k_2^0 k_3^1 k_4^1} \quad (9.12)$$

$$I_{k_3} = \frac{\sum k_1^0 k_2^0 k_3^1 k_4^1}{\sum k_1^0 k_2^0 k_3^0 k_4^1} \quad (9.13)$$

$$I_{k_4} = \frac{\sum k_1^0 k_2^0 k_3^0 k_4^1}{\sum k_1^0 k_2^0 k_3^0 k_4^0} \quad (9.14)$$

We can make a comparative analysis with this index system by time, region, branch, micro-level, and macro-level. The branch concept in making an index analysis of the development of industry and construction, in particular, in relation to capital intensive branches, such as machine-building and construction, enables determination of the effect of extensive and intensive factors on the increase in production volume.

We can use a more compact calculation method for the same purpose. It relies on a base level of profitability adjusted for the indices included in the factorial model. Here, the connections with a base level of profitability for individual factor indices are allocated computed values from the effect of *i*-factor by a multiplication series, while other factors included in the model remain unchanged.

With a base value of profitability, y^0, a computed value for the first factor, y^1, and for the second factor, y^2 etc., we can generate an order for determining the absolute effect

$$y^1 = I_{x_1} \times y^0$$

$$y^2 = I_{x_2} \times y^1$$

$$y^3 = I_{x_3} \times y^2$$

$$y^6 = I_{x_6} \times y^5$$

In a general form, the computation formula is

$$\Delta y\,(x_i) = y^0\left(\prod_{i=1}^{n} I_{x_i} - \prod_{i=1}^{n-1} I_{x_i}\right) \quad (9.15)$$

The use of sign "-" points to the reverse effect. Applying such a method, we measure the extent of the effect of each factor on general changes in the level of profitability. We can measure such an effect as a percentage

$$d_{x_i} = \frac{\Delta y(x_i)}{\Delta y}. \quad (9.16)$$

A sum of products by several objects and the absolute increase are calculated in the numerator and in the denominator of corresponding formulas to analyze the effect of factors for several objects (e.g., companies of the same branch). The effect of each factor is computed as the difference between the numerator and denominator of the corresponding index.

9.7. Characteristics of applying the index method

An index of any indicator of the effectiveness of an investment activity has a factor of unevenness in the development of different functional elements of the investment system (companies, industries, regions, and markets, etc.) and, thereby, a structural non-uniformity is introduced. The irregularity of the development of different components of the investment process inevitably leads to a change in the specific structure and technical level of capital. As such, to study the effect of factors of unevenness in the dynamics of an effective indicator is of importance. In other words, it is important to analyze the structural changes and determine their effects on the dynamics of the investment activity at all levels.

Structural changes can be estimated with the absolute changes of specific weights, expressed in percentage points. However, estimation of the intensity of structural changes and measurement of the degree of their effect on the dynamics of the average level of the studied indicator are a necessity for analyzing the effectiveness of the investment activity.

This task is performed using indices of structural change. In the simplest variant this system includes three indices: an index of the dynamics of average levels (index of a composition); an index of the average levels in an unchanged structure; and an index of structural changes. It is advisable to use this system for the estimation of the effect of a structural factor on the change in profitability of an investment portfolio or a portfolio of securities. The structure of an investment portfolio is known to determine

the type of investor it attracts and the main aim of the investment activity. For example, low-profitable securities predominating in a portfolio implies that an investor is risk-averse. On the contrary, an investor with an aggressive character and risk-taking attitude is reflected in a portfolio of securities with a high level of profitability. Furthermore, an important component of portfolio analysis is the structural changes of the investment portfolio. If an investment portfolio's structure is rather stable, this confirms a passive policy in managing it. An established tendency to change a portfolio's structure is confirmed by significant changes in the investment portfolio, implying that its management policy is active. An index system of average profitability of an investment portfolio enables the monitoring of the dynamics of its management policy. We can use both a simple and a complex system of interconnected indices of average profitability.

An index of variable composition of average profitability can be built for a securities portfolio using a simple approach to characterize a change in average profitability as a result of both the dynamics of the profitability level of different securities (i) and structural changes in the portfolio of securities (d)

$$I_{\bar{\imath}} = \frac{\sum i_1 d_1}{\sum i_0 d_0} \tag{9.17}.$$

An index of fixed composition eliminates the effect of structural change in a portfolio on changes in average profitability. When formulated, appropriate weights, in compliance with an accepted system of weighting, are incorporated at the level of the reporting period. For a profitability analysis of a securities portfolio, it is prudent to choose a base period for a weight because we capture important characteristics of average profitability for a current period in the numerator, provided the structure remains unchanged $I_i = \frac{\sum i_1 d_0}{\sum i_0 d_0}$ (9.18).

The use of the classical principle of fixation in the current period for the above index is not relevant because the value in the denominator characterizes the average profitability of a previous period under the current structure. This problem is solved by use of the principle proposed by Fisher. The main idea of the Fisher Index is to take into consideration both the base and current periods. In our case, the index is calculated as a geometric mean from indices of average profitability built on the current and base structures

$$\boldsymbol{I_i^F = \left(\frac{\sum i_1 d_1}{\sum i_0 d_1} \times \frac{\sum i_1 d_0}{\sum i_0 d_0}\right)^{1/2}} \tag{9.19}.$$

In an index of structural change, the values for profitability of different securities are classically taken to be unchangeable (fixed at the level of the base period) $I_d = \frac{\sum i_0 d_1}{\sum i_0 d_1}$ (9.20).

However, in light of earlier comments on making an index of fixed composition, it is imperative either to use fixation at the current period $I_d = \frac{\sum i_1 d_1}{\sum i_1 d_0}$ (9.21) or to apply the Fisher Index

$$\boldsymbol{I_d^F = \left(\frac{\sum i_1 d_1}{\sum i_1 d_0} \times \frac{\sum i_0 d_1}{\sum i_0 d_0}\right)^{1/2}} \quad (9.22).$$

Within this index system characterizing the dynamics of the average level of profitability of a securities portfolio, i.e. the main indicator of efficiency of financial investments, the index of structural change measures the effect of a structural change in the investment resource (the efficiency that it expresses) on the dynamics of average profitability.

However, in studying the effect of a structural change on the effectiveness of the investment activity, it is not sufficient to use a single index of structural change. This is because of the importance of identifying the occurrence of a structural change and establishing the effect of a structural change on the dynamics of the investment activity, rather than estimating the quantitative effect of a structural change.

As such, a more complicated system of interconnected indices is used to analyze structural changes in the investment portfolio, which can contain different investment instruments. It is important to consider the presence of an investment structure – by the first feature (the directions of investments D); and by the second feature (the investment tools used d). When such an index system is formulated, grouped average values, in accordance with the base and current periods, are used

$$\bar{\imath}_0 = \sum(\sum i_0 d_0)\, D_0 \quad (9.23),$$

$$\bar{\imath}_1 = \sum(\sum i_1 d_1)\, D_1 \quad (9.24).$$

The index of variable composition of the above system characterizes the dynamics of average profitability of an investment portfolio provided there is a distribution by two grouping features inside the system

$$I_{\bar{\imath}} = \frac{\sum(\sum i_1 d_1) D_1}{\sum(\sum i_0 d_0) D_0} \quad (9.25).$$

The index of fixed composition that determines a change in average profitability when there is a constant composition of an investment portfolio is to be fixed either at the base level, as in the previous index system – a deviation from a classical presentation of the index

$$I_{id} = \frac{\sum(\sum i_1 d_1)D_0}{\sum(\sum i_0 d_0)D_0} \tag{9.26}$$

or to use the Fisher Index

$$I_d^F = \left(\frac{\sum(\sum i_1 d_1)D_0}{\sum(\sum i_0 d_0)D_0} \times \frac{\sum(\sum i_1 d_1)D_1}{\sum(\sum i_0 d_0)D_1}\right)^{1/2} \tag{9.27}.$$

The index of structural change measuring the effect of changes in the distribution between the direction of investments on the dynamics of average profitability, but without consideration of the structure of investment instruments, is calculated on the conditions of current weighting

$$I_d = \frac{\sum(\sum i_1 d_1)D_1}{\sum(\sum i_1 d_1)D_0} \tag{9.28}$$

or with the Fisher Index

$$I_d^F = \left(\frac{\sum(\sum i_1 d_1)D_1}{\sum(\sum i_1 d_1)D_0} \times \frac{\sum(\sum i_0 d_0)D_1}{\sum(\sum i_0 d_0)D_0}\right)^{1/2} \tag{9.29}.$$

In addition, using this index system we can estimate the effect of a structural change inside the group $I_d = \frac{\sum(\sum i_1 d_1)D_0}{\sum(\sum i_1 d_0)D_0}$ (9.30) and also a general structural change $I_{dD} = \frac{\sum(\sum i_1 d_1)D_1}{\sum(\sum i_1 d_0)D_0}$ (9.31).

It is possible to construct the interconnection between partial indices of structural change: $I_{dD} = I_d \times I_D$ (9.32). A general interconnection is given by $I_{\bar{i}} = I_i \times I_d \times I_D$ (9.33).

The methodological uniqueness of the formulation of a system of interconnected indices of average values is the comparability of structural parts of a studied population over time. In the process of changing the structure of an investment portfolio, some instruments can be sold or disappear, while others can be bought or emerge. In such cases, the incomparability of the structural parts of a portfolio makes it impossible to analyze the index at an average level. This warrants some correction of the index system. Here, in conjunction with the indices of fixed composition, structural changes are calculated for a population with similar types of

instruments. Such an index system includes those indices that can capture the effect of new structural parts and the elements that have disappeared from the portfolio. The index system then becomes

$$I_{\bar{i}} = I_i \times I_{d^0} \times I_{d^+} \times I_{d^-} \quad (9.34).$$

All indices are calculated in consideration of the relationship between average levels of profitability for two periods.

The index of variable composition characterizes the dynamics of average profitability of a securities portfolio under the effect of all factors

$$I_{\bar{i}} = \frac{\sum i_1 d_1}{\sum i_0 d_0} \quad (9.35).$$

The index of fixed composition shows how the profitability of a portfolio with a single type composition of securities changes on average

$$I_i = \frac{\sum i_1 d_1^0}{\sum i_0 d_1^0} \quad (9.36).$$

The index of structural change characterizes the effect of a redistribution of securities in a single type composition securities portfolio

$$I_{d^0} = \frac{\sum i_1 d_1^0}{\sum i_0 d_0^0} \quad (9.37).$$

The indices of structural change, I_{d^+} and I_{d^-}, show how the average profitability of a portfolio changes with the purchase of new securities and the sale (payment) of old ones, respectively

$$I_{d^+} = \frac{\sum i_1 d_1}{\sum i_1 d_1^0} \quad (9.38),$$

$$I_{d^-} = \frac{\sum i_0 d_0^0}{\sum i_0 d_0} \quad (9.39).$$

Similarly, we can analyze the absolute change in average profitability of a securities portfolio depending on the change in the average level of an indexed indicator.

We present one more example of the practical application of an index method. In real-life experience/practice, various indices are used to estimate stock exchange activity and analyze the dynamics of exchange prices.

Among these, the most well-known indices are Dow Jones and S&P 500. The calculation is based on the average price of a stock using the unweighted arithmetic mean (Dow Jones) and the weighted arithmetic mean (S&P 500) for a comparative circle of issuers. The differences in the indexed values of weighted and unweighted average prices are explained by the effect of structural changes in the sales volume of different stocks. In the conditions of a stable stock market, the effect of structural changes on the dynamics of average prices is unstable – this can be determined in the framework of a traditional index system of average values (variable composition, fixed composition, and structural change). However, when a primary placement of securities takes place, the composition of the issuers changes. Then, to characterize a market condition, monitoring and making a comparative analysis of stock prices quoted on the stock exchange or in circulation at an over-the-counter market, it is useful to calculate indices from the data of the whole population (not a fixed one) of issuers. An average stock value at a certain bargaining point is calculated as a weighted arithmetic mean $\overline{k} = \frac{\sum_{i=1}^{n} k_i q_i}{\sum_{i=1}^{n} q_i}$, where q_i is the number of sold stocks.

The attractiveness of some stocks and the correlation in their sales volume influence the average stock value. Specified factors also have an effect on the dynamics of the average stock value. When a securities market becomes unstable, the effect of structural changes is rather significant. This can be estimated within one index system in three dimensions:

a) changes in the structure of stock sales by a comparative circle of issuers;
b) entry of new issuers into the stock market;
c) the departure of some issuers who had previously taken part in current bargaining.

Accordingly, three indices of structural changes are included in the system $I\frac{SC}{k^0}$, $I\frac{SC}{k^+}$,$I\frac{SC}{k^-}$.

The index of stock value for a comparative circle of issuers $I\frac{SC}{k^0}$, by its statistical nature, is a weighted average index of the S&P 500 type.

Then, the index system is $I\frac{VC}{k} = I\frac{\mathrm{FC}}{k^0} \times I\ \frac{SC}{k^0} \times I\ \frac{SC}{k^+} \times I\ \frac{SC}{k^-}$.

All indices are calculated through the correlation of average stock values (real or nominal) for two periods (bargaining) – current (I) and base (0).

The index of variable composition $I\frac{VC}{k}$ characterizes the dynamics of average stock value under the effect of all factors

$$I\frac{VC}{\bar{k}} = \frac{\sum k_1 q_1}{\sum q_1} \div \frac{\sum k_0 q_0}{\sum q_0} = \frac{\bar{k}_1}{\bar{k}_0}.$$

The index of fixed composition $I\frac{\mathrm{FC}}{k^0}$ shows how, on average, the stock value of a comparative composition of issuers has changed

$$I\frac{\mathrm{FC}}{k^0} = \frac{\sum k_1 q_1^0}{\sum q_1^0} \div \frac{\sum k_0 q_1^0}{\sum q_1^0}.$$

The index of structural changes $I\frac{SC}{k^0}$ characterizes the effect of the redistribution of stock sales volumes of a comparative circle of issuers

$$I\frac{SC}{k^0} = \frac{\sum k_0 q_1^0}{\sum q_1^0} \div \frac{\sum k_0 q_0^0}{\sum q_0^0}.$$

Indices of structural change, $I\frac{SC}{k^+}$ and $I\frac{SC}{k^-}$, show how an average stock value has changed due to new and departed issuers, respectively

$$I\frac{SC}{k^+} = \frac{\sum k_0 q_1^0}{\sum q_1^0} \div \frac{\sum k_1 q_1^0}{\sum q_1^0}, I\frac{SC}{k^-} = \frac{\sum k_0 q_0^0}{\sum q_0^0} \div \frac{\sum k_0 q_0^0}{\sum q_0^0}.$$

The analytical potential of such an index system is illustrated in the following example.

Example 9.7

Using the results of stock bargaining, analyze the dynamics of average stock value in general and under the effect of some factors. Make some conclusions.

Issuer	Market value, euro		Number of stocks, thousand pieces	
	k_0	k_1	q_0	q_1
A	20	25	5	10
B	50	40	2	5
C	10	-	10	-
D	-	15	-	10
E	-	20	-	40
Total	x	x	17	65

Let us build a table to calculate the indices of average value.

Issuer	Incomparable composition of issuers		Comparable composition of issuers				
	k_0q_0	k_1q_1	q_0^0	q_1^0	$k_0q_1^0$	$k_1q_1^0$	$k_0q_0^0$
A	10	25	5	10	20	25	10
B	10	20	2	5	25	20	10
C	10	-	-	-	-	-	-
D	-	15	-	-	-	-	-
E	-	80	-	-	-	-	-
Total	30	140	7	15	45	45	20

$I\frac{VC}{\bar{k}} = \frac{140}{65} \div \frac{30}{17} = \frac{2.154}{1.765} = 1.305$, i.e. an increase of 30.5 %;

$I\frac{\mathrm{FC}}{k^0} = \frac{45}{15} \div \frac{45}{15} = \frac{3}{3} = 1.0$, i.e. did not change;

$I\frac{SC}{k^0} = \frac{45}{15} \div \frac{20}{7} = \frac{3}{2.857} = 1.05$, i.e. an increase of 5 %;

$I\frac{SC}{k^+} = \frac{140}{65} \div \frac{45}{15} = \frac{2.154}{3} = 0.718$, i.e. an decrease of 28.2 %;

$I\frac{SC}{k^-} = \frac{20}{7} \div \frac{30}{17} = \frac{2.857}{1.765} = 1.619$, i.e. an increase of 61.9 %.

Let us check the interconnection of the calculated indices 1.0 × 1.05 × 0.718 × 1.619 = 1.305. Thus, the average stock value for the current period increased by 30.5 % compared to the base value – the full increase of the value was due to structural changes. The replacement of issuer B with two new issuers having low stock value (D, E) caused an increase in average stock value of 16.2 % (0.718 × 1.61). The average value increased by 5 % due to the increase in a specific weight of the stock of issuer B – the value being the highest. The index of fixed composition shows that, despite registering opposite dynamics, the stock value of issuers A and B did not, on average, change.

It appears clear that a system of indices is more informative than a single index. Using this system we can make, not just an analysis of the dynamics of stock exchange bargaining and operations at the over-the-counter securities market, but also a comparative analysis of the dynamics of the stock exchange and over-the-counter market values.

Practice Exercises

EXERCISE 9.1

Taking the data in Example 9.5 in Chapter 9, calculate summary indices using the structure of working time/hours as a weight. Explain the economic meaning of the difference between the numerator and the denominator of each index.

EXERCISE 9.2

When determining the role/effect of a structural factor on the change in average yield capacity of 2 grain crops, is there a risk that we are making a mistake in terms of how many cases out of 100 there are? What about for 10 crops?

EXERCISE 9.3

Below, we have some data on commodity sales.

Commodity	Sales, ton		Price of 1 kg, euro	
	Base period	Reporting period	Base period	Reporting period
Pears	15.0	16.2	2.5	3.0
Apples	50.0	51.0	4.5	7.0
Total	65.0	67.2	x	x

Calculate the individual and summary indices of price and physical volume of the marketed goods. Show the connection between them. Determine the general increase in turnover and the increase due to various factors. Make some conclusions.

EXERCISE 9.4

Below we have some data on commodity sales in a store.

Commodity groups	Sales for last year, million euro	Rate of increase (decrease) of physical volume of marketed goods compared to previous year, %
Foodstuffs	150	-2
Household appliances	200	+5
Clothes and footwear	30	+20
Total	380	x

Determine a summary index of physical volume and the absolute change in proceeds from sales due to physical volume. Analyze the results.

EXERCISE 9.5

Below, we have some data on commodity sales for two quarters of the current year.

Commodities	Turnover in ruling prices, euro		Change in average prices in quarter II compared to quarter I, %
	I quarter	II quarter	
Sweets	60	64	-20
Drinks	42	44	+10
Haberdashery	35	38	without changes
Total	137	146	x

Determine summary indices and the absolute general change in turnover and the change due to certain factors. Analyze the results.

EXERCISE 9.6

Below, we have some data on winter crop production in Ukraine for two years.

Crops	Yield capacity, centner/hectare		Cultivation area, thousand hectares	
	1st year	2nd year	1st year	2nd year
Wheat	38	31	7000	5200
Rye	24	23	5700	4500
Total	x	x	11270	9700

Determine: (1) a summary index of yield capacity and cultivation area; (2) the absolute change in gross yield of different crops, due to the change in cultivation area, and change under the effect of two factors; and (3) check the interconnection of the calculated indices. Present the calculations in the table and analyze the results.

EXERCISE 9.7

Below are some data on the business performance of two companies in January and February.

Company	January		February	
	Output, thousand pieces	Production expenses, million euro	Output, thousand pieces	Production expenses, million euro
A	60	24	80	20
B	60	20	120	18
Total	120	44	200	38

Calculate summary indices of production expenses, the absolute change of production expenses in general, and that due to certain factors in terms of

production cost and physical volume of output. Check the interconnection of the computed indices and make some conclusions.

EXERCISE 9.8

Below are some data on annual average milk yield per cow at farm enterprises in two regions.

Region	Productivity, kg		Livestock, thousand head	
	Previous period	Current period	Previous period	Current period
A	2060	2120	20	22
B	2040	1950	40	45
Total	x	x	60	67

Determine: (1) summary indices of production volume, annual average milk yield, and livestock number and (2) the absolute change in milk production volume in the current year compared to the previous year due to the change in milk yield and the number of livestock. Explain which of the two factors had a greater effect on volume of milk production. Present the calculations in the table and analyze the results.

EXERCISE 9.9

Below, we have some data on wage/salary and number of workers and employees doing two different types of activity for two years.

Kind of activity	Average wage of one worker, euro		Number of workers, thousand	
	Base	Fact	Base	Fact
Industry	1600	2200	150	140
Trade	1800	2400	20	50
Total	x	x	170	190

Determine: (1) summary indices of wage and number of workers; (2) a summary index of wage costs using the interconnection of indices; and (3) the absolute change in wage costs due to a change in the average wage and number of workers. Explain which of the two factors had a larger effect on the change in the fund of general wages. Present the calculations in the table and analyze the results. Make some conclusions.

EXERCISE 9.10

Below, we have some data on turnover of foodstuffs and consumer goods for two periods.

Goods	Turnover for base period, million euro	Turnover for accounting year, million euro	Price index
Foodstuffs	26.1	210	5.3
Consumer goods	17.5	130	4.8
Total	43.6	340	x

Determine: (1) the relative change in turnover in current prices and the average interest of price rise and (2) the absolute change in turnover due to the price rise. Make some conclusions.

EXERCISE 9.11

Below, we have some data on the turnover of a trading company.

Goods	Turnover in December, thousand euro	Physical sales volume, thousand kg	
		December	January
Sweets	200	40	40
Biscuits	100	50	55
Halva	25	5	6
Total	325	95	101

Determine the absolute and relative change in turnover due to the effect of the physical volume of sold goods. Present the calculations in the table and analyze the results.

EXERCISE 9.12

Below, we have some data on the turnover of a trading company.

Goods	Turnover in January, thousand euro	Price of 1 kg, euro	
		December	January
Cheese	150	6.5	6.8
Oil	120	7.0	7.5
Total	270	x	x

Determine the absolute and relative change in turnover due to the change in goods prices. Present the calculations in the table and analyze the results.

EXERCISE 9.13

Below are some data on wheat production in Ukraine for two years

Wheat type	Yield capacity, centner/hectare		Cultivation area, thousand hectares	
	1st year	2nd year	1st year	2nd year
Winter	38.0	31.0	7000	5205
Spring	27.5	25.0	20	55
Total	x	x	7020	5260

Calculate: (1) an index of variable composition for average yield capacity; (2) indexes of fixed composition and structural change for average yield capacity; and (3) the absolute change in gross yield due to the change in yield capacity of each type of wheat. Present the calculations in the table and analyze the results.

EXERCISE 9.14

Below, we have some data on available housing area in two regions for two years.

Region	General dwelling area, million m^2		Rate of increase of population, %
	Base	Fact	
V	69	76	1.3
D	96	93	0.1
Total	165	169	x

Determine: (1) the relative change in the size of the general dwelling area; (2) the relative change in dwelling area per capita (index of dwelling area supply); and (3) calculate the missing index using the interconnection of indices. Make some conclusions.

EXERCISE 9.15

Below, we have some data on the results bargaining for two stocks of A, B, and C over two days.

Issuer	Bargaining volume, thousand euro		Relative change in price of sold stock, %
	September 1	September 8	
A	25	25	+10
B	10	15	+20
C	35	60	without changes
Total	70	100	x

Calculate: (1) the index of bargaining volume; (2) the relative change in bargaining volume due to price and number of sold stock; (3) the absolute change in bargaining volume due to the change in price of each stock and its number; and (4) check the interconnection of indices. Present the calculations in the table and analyze the results.

EXERCISE 9.16

Below are some data on people's deposits in savings and commercial banks.

Bank type	Total amount of deposits in the current period, million euro	Index of average deposit
Savings	130	1.5
Commercial	80	6.5
Total	210	x

Determine: (1) the summary index of the average deposit in all banks; (2) the absolute increase in the amount of deposits due to the increase in average deposit. Present your calculations in the table and analyze the results.

EXERCISE 9.17

Below are some data on people's deposits in savings and commercial banks for two years.

Bank type	Average deposit amount, euro		Number of deposits, thousand	
	1st year	2nd year	1st year	2nd year
Savings	1550	1975	840	830
Commercial	5500	7250	45	135
Total	x	x	885	965

Calculate: (1) the index of variable composition for an average deposit amount; (2) indexes of fixed composition and structural change of the average deposit amount; and (3) the absolute change in deposit amount due to the change in average deposit amount in each type of bank. Present the calculations in the table and analyze the results.

EXERCISE 9.18

Below are some data on labor productivity and the structure of the number of workers in two regions.

Branch	Labor productivity, thousand euro		Structure of the number of workers, %		
	Region A	Region B	Region A	Region B	Average for two regions
Mining	30	35	40	20	30
Processing	60	72	60	80	70

Determine spatial indices of variable and fixed composition for average labor productivity and an index of structural change taking region A as a base for comparison. Present the calculations in the table and analyze the results.

EXERCISE 9.19

Below are some data on people's deposits in savings and commercial banks in two regions.

Bank type	Average deposit amount, euro		Deposit amount, million euro	
	Region A	Region B	Region A	Region B
Savings	1350	1240	5000	6000
Commercial	5850	4700	1450	4000
Total	x	x	6450	10000

Determine the spatial index for an average deposit amount of variable and fixed composition, taking region B as a base for comparison, and the index of structural change. Analyze the results.

EXERCISE 9.20

Below are some data on the materialization of goods.

Commodity groups	Sales in previous year, million euro	Rates of increase (decrease) in sales volume compared to last year, %
Haberdashery	150	-2
Textile	200	+5
Footwear	30	+20
Clothes	320	no change
Total	700	x

Calculate the absolute and relative change in turnover due to the change in the physical volume of goods sold. Make some conclusions.

EXERCISE 9.21

Below, we have some data on the turnover of a trading company.

Commodity	Turnover in January, thousand euro	Price of 1 kg, euro	
		December	January
Cheese	150	25.0	32.0
Oil	120	7.0	7.5
Butter	90	20.0	22.5
Total	360	x	x

Determine the absolute and relative change in turnover due to the change in the price of goods. Make some conclusions.

EXERCISE 9.22

Using the data below on commodity sales for two quarters of the current year, determine the summary indices and the absolute change in general turnover and due to particular factors. Make some conclusions.

Goods	Turnover in ruling prices, thousand euro		Price change in quarter II compared to quarter I, %
	I quarter	II quarter	
Sugar products	60	65	-20.0
Bakery products	120	140	+2.5
Drinks	40	45	+10.0
Total	220	250	x

EXERCISE 9.23

Below, we have some data on turnover in a trading company.

Commodity	Turnover in December, thousand euro	Physical sales volume, thousand kg	
		December	January
Sweets	200	40	45
Biscuits	100	50	54
Halva	25	5	6
Total	325	95	105

Determine the absolute and relative change in turnover due to the factor of physical sales volume. Make some conclusions.

EXERCISE 9.24

Below are some data on annual average milk yield in two regions.

Region	Productivity, kg		Livestock, thousand head	
	Previous period	Current period	Previous period	Current period
A	2060	2120	20	22
B	2040	1950	40	45
Total	x	x	60	67

Determine: (1) the index of average productivity of variable composition, fixed composition, and also the index of structural change; (2) the absolute change in milk production for the current period compared to the previous one due to changes in productivity and livestock number. Explain which of the two factors had a greater effect on the volume of milk production. Present the calculations in the table and analyze the results.

EXERCISE 9.25

Below are some data on real wages/salaries and the number of workers and employees by two types of activity for two years.

Activity	Average wage of one worker, euro		Number of workers, thousand	
	Base	Fact	Base	Fact
Manufacturing industry	1600	2200	150	140
Construction	1800	2400	20	50
Total	x	x	170	190

Determine: (1) the summary indices of wage and number of workers; (2) the summary index of wage costs using the interconnection of the indices; and (3) the absolute change in wage costs due to changes in average wage

and number of workers. Explain which of the two factors had a greater effect on the change in the general fund of wages.

EXERCISE 9.26

Below are some data on wheat production for two years

Wheat type	Yield capacity, centner/hectare		Structure of cultivation area, %	
	Base	Fact	Base	Fact
Winter	38	31	85	90
Spring	28	25	15	10
Total	x	x	100	100

Calculate: (1) the index of variable composition of average yield capacity; (2) the index of fixed composition and structural change for average yield capacity; and (3) the absolute change in gross yield due to a change in yield capacity of each wheat type. Present the calculations in the table and analyze the results. Make some conclusions.

EXERCISE 9.27

Below, we have some data on the results of bargaining of A, B, and C for two days.

Issuer	Stock price, euro		Number of stock sold, thousand pieces	
	January 10	January 17	January 10	January 17
A	10	12	30	80
B	50	45	70	20
C	100	100	50	50
Total	x	x	150	150

Calculate: (1) the relative change in bargaining volume in general and under the effect of two factors – price and number of stocks; (2) the absolute change due to the price factor; (3) check the influence of structural changes on bargaining volume; and (4) check the interconnection of indices. Present the calculations in the table and analyze the results.

EXERCISE 9.28

Below are some data on deposits in savings and commercial banks for two years.

Bank type	Average deposit amount, euro		Deposit structure, %	
	1st year	2nd year	1st year	2nd year
Savings	1550	1975	80	75
Commercial	5500	7250	20	25
Total	x	x	100	100

Calculate: (1) the index of variable composition for an average deposit amount; (2) the index of fixed composition and structural change for an average deposit amount; and (3) the absolute change in deposit amount due

to a change in the average deposit amount in each type of bank. Present the calculations in the table and analyze the results.

EXERCISE 9.29

Below, we have some data on deposits in savings and commercial banks in two regions.

Bank type	Average deposit amount, euro		Deposit structure, %		
	Region A	Region B	Region A	Region B	Average for two regions
Savings	550	750	95	55	70
Commercial	2800	3700	5	45	30

Determine the spatial index for the average deposit amount of variable and fixed composition, taking region A as a base for comparison; also determine the index of structural change. Make some conclusions.

APPENDIX

Table 1. Critical values for the correlation relation η^2 and the coefficient of determination R^2 (α=0,05)

k_2	k_1								
	1	2	3	4	5	6	8	10	20
3	0,771	865	903	924	938	947	959	967	983
4	658	776	832	865	887	902	924	937	967
5	569	699	764	806	835	854	885	904	948
6	500	632	704	751	785	811	847	871	928
7	444	575	651	702	739	768	810	839	908
8	399	527	604	657	697	729	775	807	887
9	362	488	563	628	659	692	742	777	867
10	332	451	527	582	624	659	711	749	847
11	306	420	495	550	593	628	682	722	828
12	283	394	466	521	564	600	655	696	809
14	247	345	417	471	514	550	607	650	773
16	219	312	378	429	477	507	564	609	740
18	197	283	348	394	435	470	527	573	709
20	179	259	318	364	404	432	495	540	680
22	164	238	294	339	377	410	466	511	653
24	151	221	273	316	353	385	440	484	628

26	140	206	256	297	332	363	417	461	605
28	130	193	240	279	314	344	396	439	583
30	122	182	227	264	297	326	373	419	563
32	115	171	214	250	282	310	360	401	544
34	108	162	203	238	268	296	344	384	526
36	102	153	102	226	256	282	329	368	509
38	097	146	184	218	245	271	316	355	493
40	093	139	176	207	234	259	304	342	479
50	075	113	143	170	194	216	254	288	416
60	063	095	121	144	165	184	218	249	368
80	047	072	093	110	127	142	170	196	298
100	038	058	075	090	103	116	140	161	251
120	032	049	063	075	087	098	119	137	217
200	019	030	038	046	053	060	073	086	139
400	010	015	019	023	027	031	038	044	074

Table 2. Critical values for the Fisher's criterion (α=0,05)

k_2	k_1								
	1	2	3	4	5	6	8	10	20
1	161,4	199,5	215,7	224,6	230,2	234,0	238,9	242,0	248,0
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,39	19,44
3	10,13	9,45	9,28	9,12	9,01	8,94	8,84	8,78	8,66
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,96	5,80
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,74	4,56
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,06	3,87
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,63	3,44
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,34	3,15
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,13	2,93
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,97	2,77
11	4,82	3,98	3,59	3,63	3,20	3,09	2,95	2,86	2,65
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,76	2,54
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,60	2,39
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,49	2,28
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,41	2,19
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,35	2,12
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,16	1,93
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,12	1,84
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	2,04	1,7Ь
120	3,92	3,07	2,68	2,45	2,29	2,17	2,02	1,90	1,65

Table 3. Quantiles of the χ^2 distribution

f	*1* - α					
	0,025	0,050	0,10	0,90	0,95	0,975
A	1	2	3	4	5	6
1	0,01	0,04	0,02	2,71	3,84	5,02
2	0,05	0,10	0,21	4,61	5,99	7,38
3	0,22	0,35	0,58	6,25	7,82	9,35
4	0,48	0,71	1,06	7,78	9,49	11,14
5	0,83	1,15	1,61	9,24	11,07	12,03
6	1,24	1,64	2,20	10,65	12,59	14,45
7	1,69	2,17	2,83	12,02	14,07	16,01
8	2,18	2,73	3,49	13,36	15,51	17,54
9	2,70	3,33	4,17	14,68	16,92	19,02
10	3,25	3,94	4,87	15,99	18,31	20,48
11	3,82	4,58	5,58	17,28	19,68	21,92
12	4,40	5,23	6,30	18,55	21,03	23,34
13	5,01	5,89	7,04	19,81	22,36	24,74
14	5,63	6,57	7,79	21,06	23,69	26,12
15	6,26	7,26	8,55	22,31	25,00	27,49
16	6,91	7,96	9,31	23,54	26,30	28,85
17	7,56	8,67	10,09	24,77	27,59	30,19
18	8,23	9,39	10,87	25,09	28,87	31,53
19	8,91	10,12	11,65	27,20	30,14	32,85
20	9,59	10,85	12,44	28,41	31,41	34,17
22	10,98	12,34	14,04	30,81	33,92	36,78
24	12,40	13,85	15,66	33,20	36,42	39,36
26	13,84	15,38	17,29	35,56	38,89	41,92

28	15,31	16,93	18,94	37,92	41,34	44,46
30	16,79	18,49	20,60	40,26	43,77	46,90
35	20,57	22,47	24,80	46,06	49,00	53,20
40	24,43	26,51	29,05	51,81	55,76	59,34
45	28,37	30,61	33,35	57,51	61,66	65,41
50	32,36	31,76	37,69	63,17	67,51	71,42

Table 4. Quantiles of normal distribution

1 - *a*	0,800	0,900	0,950	0,975
t	0,84	1,28	1,64	1,96
$\|t\|$	1,28	1,64	1,96	2,24

Table 5. Student's distribution quantile (t)

k	***1*** **- α**			*k*	***1*** **- α**		
	0,90	0,95	0,975		0,90	0,95	0,975
3	1,64	2,35	3,18	12	1,36	1,78	2,18
4	1,53	2,13	2,78	14	1,35	1,76	2,14
5	1,48	2,02	2,57	16	1,34	1,75	2,12
6	1,44	1,94	2,45	18	1,33	1,73	2,10
7	1,41	1,89	2,36	20	1,33	1,72	2,09
8	1,40	1,86	2,31	22	1,32	1,72	2,07
9	1,38	1,83	2,26	24	1,32	1,71	2,06
10	1,37	1,81	2,23	28	1,31	1,70	2,05
11	1,36	1,80	2,20	∞	1,28	1,64	1,96

Table 6. Critical values of the linear correlation coefficient (α=0,05)

Sample	5	6	7	8	9	10	12	14	16
$r_{0,95}$	0,88	0,81	0,75	0,71	0,67	0,63	0,58	0,53	0,50

Table 7. Critical values of Spearman's rank correlation coefficient (α=0,05)

Sample size, *n*	5	6	7	8	9	10	11	12
$P_{0,95}$	0,90	0,83	0,71	0,64	0,60	0,56	0,53	0,50

Table 8. Critical values of the A. Kolmogorov's function $K_{(\lambda)}$ (α=0,05)

λ	1,23	1,36	1,63	1,80	2,03
$K_{(\lambda)}$	0,9030	0,9505	0,9902	0,9970	0,9993

Table 9. Critical values of the cumulative criterion (α=0,05)

n	To check the significance of the trend		To test the hypothesis about the shape of the trend	
	T	*t*	Linear function	Parabola of the 2nd order
6	2,62	2,08	0,85	0,51
7	3,11	2,10	1,01	0,61
8	3,59	2,09	1,17	0,70
9	4,07	2,09	1,32	0,79
10	4,55	2,09	1,48	0,89
11	5,02	2,08	1,63	0,98
12	5,49	2,08	1,78	1,06
13	5,96	2,07	1,93	1,15
14	6,42	2,07	2,08	1,23
15	6,89	2,06	2,23	1,33
16	7,36	2,06	2,38	1,42
17	7,82	2,06	2,58	1,51
18	8,29	2,05	2,68	1,59
19	8,76	2,05	2,83	1,68
20	9,22	2,04	2,98	1,77
22	10,20	2,04	3,28	1,94
24	11,00	2,04	3,58	2,11
26	12,00	2,03	3,88	2,29
28	12,90	2,03	4,18	2,46
30	13,90	2,03	4,47	2,63

Table 10. Normal Distribution Functions

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0400	0080	0120	0160	0199	0239	0279	0319	0359
0,1	0398	0438	0478	0517	0557	0596	0636	0675	0711	0753
0,2	0793	0832	0871	0910	0948	0987	1026	1064	1103	1141
0,3	1179	1217	1255	1293	1331	1368	1106	1113	1480	1517
0,4	1554	1591	1628	1664	1700	1736	1772	1808	1811	1879
0,5	1915	1950	1985	2019	2054	2088	2123	2157	2190	2224
0,6	2257	2291	2324	2357	2389	2422	2454	2486	2517	2549
0,7	2580	2611	2642	2673	2704	2734	2764	2794	2823	2852
0,8	2881	2910	2939	2967	2995	3023	3051	3078	3106	3133
0,9	3159	3186	3212	3238	3264	3289	3315	3310	3365	3389
1,0	3413	3438	3461	3485	3508	3531	3554	3577	3S99	3621
1,1	3643	3665	3686	3708	3729	3749	3770	3790	3810	3830
1,2	3849	3869	3888	3907	3925	3944	3962	3980	3997	4015
1,3	4032	4049	4066	4082	4099	4115	4131	4117	4162	1177
1,4	4192	4207	4222	4236	4251	1265	4279	4292	4306	4319
1,5	4332	4345	4357	4370	4382	4394	4406	4118	4429	1141
1,6	4452	4463	4474	4484	4495	4505	4515	4525	4535	4545
1,7	4554	4564	4573	4582	1591	4599	4608	4616	4625	4633
1,8	4641	4649	4656	4664	4671	1678	4686	4693	4699	4706
1,9	4713	4719	4726	4732	4738	4744	4750	1756	4761	4767
2,0	4772	4778	4783	47R8	1793	4798	1803	4808	4812	4817
2,1	4821	4826	4830	4834	4838	4842	4846	4850	1854	4857
2,2	4861	4864	4868	4871	4875	4878	4881	4884	4887	4890
2,3	4893	4896	4898	4901	4904	4906	4909	4911	4913	4916
2,4	4918	4920	4922	4925	4927	4929	4931	4932	4934	4936
2,5	4938	4940	4941	4943	4915	4946	4948	4919	4951	4952
2,6	4953	4955	4956	4957	4959	4960	4961	4962	4963	4964
2,7	4965	4966	4967	4968	4969	4970	4971	4972	4973	4971
2,8	4974	4975	4976	4977	4977	4978	4979	4979	4980	4981
2,9	4981	4982	4982	4983	4984	4984	4985	4985	4986	4981
3,0	4987	4987	1987	4988	4988	4989	4989	4989	4990	4990