MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE

SIMON KUZNETS KHARKIV NATIONAL UNIVERSITY
OF ECONOMICS

STATISTICS

Textbook

*Under the general editorship of Professor O. Rayevnyeva,
D.Sc. (Economics)*

Kharkiv
S. Kuznets KhNUE
2020

UDC 311(075.034)
S81

Authors: O. Rayevnyeva, D.Sc. (Economics), Professor; I. Aksonova, Ph.D. (Economics), Associate Professor; O. Brovko, Ph.D. (Economics), Associate Professor; V. Derykhovska, Ph.D. (Economics), Associate Professor; H. Svydlo, Ph.D. (Economics), Associate Professor; I. Sierova, Ph.D. (Economics), Associate Professor; V. Shlykova, Ph.D. (Economics), Associate Professor.

Рецензенти: завідувач кафедри статистики, обліку та економічної інформатики Дніпровського національного університету імені Олеся Гончара, д-р екон. наук, професор *О. К. Єлісєєва*; професор кафедри економічної та управлінської аналітики Донецького національного університету імені Василя Стуса, д-р екон. наук, професор *А. В. Сидорова*; професор кафедри математичного моделювання та статистики Київського національного економічного університету імені Вадима Гетьмана, д-р екон. наук, професор *З. П. Бараник.*

Рекомендовано до видання рішенням ученої ради Харківського національного економічного університету імені Семена Кузнеця.
Протокол № 8 від 12.05.2020 р.

Самостійне електронне текстове мережеве видання

S81 Statistics [Electronic resource] : textbook / O. Rayevnyeva, I. Aksonova, O. Brovko et al. ; under the general editorship of Professor O. Rayevnyeva, D.Sc. (Economics). – Kharkiv : S. Kuznets KhNUE, 2020. – 377 p. (English)
ISBN 978-966-676-809-7

Practice-oriented educational material is presented which aims to deepen students' special knowledge of statistics and helps improve the quality of acquisition of professional competences. The material is given in the form of interrelation of the components of the triune system: applied theoretical knowledge of statistics, economic and mathematical methods of statistical information processing and information support for processing and analysis of data, which meets the modern requirements of training of competent specialists in the sectors of national economy.

For students of all specialities, graduate students, lecturers, economists, analysts who are engaged in applied statistical research.

UDC 311(075.034)

ISBN 978-966-676-809-7

Statisticians will become the heroes of the 21st century.
Open Data Institute CEO Gavin Starks, UK

Introduction

We assume that most of the students are probably asking themselves the question: "When and where will I use statistics?" If you read any newspaper, watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you read a newspaper article or watch a television news program, you are given sample information. With this information and statistical methods, you may make the most competent decision.

Statistics is not about numbers; it is about data – numbers in a context. It is the context that makes a problem meaningful and something worth considering. Statistics is a science whose focus is on data collection, analysis, and drawing conclusions. Some people are suspicious of conclusions based on statistical analysis. However, we believe that statistical methods, used intelligently, offer a set of powerful tools for gaining insight into the world around us. Statistics teaches us how to make intelligent judgments and informed decisions under uncertainty and fluctuations.

Statistics has a multidisciplinary nature. Statistical methods are used in business, medicine, agriculture, social sciences, natural sciences, and applied sciences, such as engineering and IT.

The widespread use of statistical analysis in diverse fields has led to increased recognition that statistical literacy – a familiarity with the goals and methods of statistics – should be a basic component of any well-rounded educational program.

There are **three important reasons** why statistical literacy is important:

1) *to be informed*. To be an informed consumer of reports you must be able to do the following:

- extract information from tables, charts, and graphs;
- follow numerical arguments;
- understand the basics of how data should be gathered, summarized, and analyzed to draw statistical conclusions;

2) *to understand issues and be able to make sound decisions based on data*. Throughout your personal and professional life, you will need to understand statistical information and make informed decisions using this information, i. e. you must be able to do the following:

- decide whether the available information is adequate or whether additional information is required;
- if necessary, collect more information in a reasonable and thoughtful way;
- summarize the available data in a useful and informative manner;
- analyze the available data;
- draw conclusions, make decisions and assess the risk of an incorrect decision;

3) *to be able to evaluate decisions that affect your life*. While you will need to make informed decisions based on data, it is also the case that other people will use statistical methods to make decisions that affect you as an individual. An understanding of statistical techniques will allow you to question and evaluate decisions that affect your well-being. An understanding of elementary statistical methods can help you to evaluate whether important decisions are being made in a reasonable way.

In the textbook, the basic questions of statistical methodology are considered, which will enable you:

- to form the ability of future specialists to systemly research and analyze the state of the modern national economy of the country taking into account the international standards and the trend in the development of processes and phenomena;
- to learn how to recognize the characteristic differences in economic processes;
- to build a system of indicators that will be adapted to changes in the market environment and highlight different aspects of the social and economic development of the country.

The textbook is based on the course of lectures developed at Simon Kuznets Kharkiv National University of Economics.

The authors hope that this textbook will help students to understand the logic behind statistical reasoning, to apply statistical methods appropriately and recognize faulty statistical arguments.

The authors are grateful for the careful review of the manuscript, valuable remarks and assistance in the preparation of the textbook for the publication to Professor O. Yelisyeyeva, Doctor of Sciences in Economics, Professor A. Sydorova, Doctor of Sciences in Economics, and Professor Z. Baranik, Doctor of Sciences in Economics.

Section 1
Introduction to statistics

1. Methodological principles of statistics

Basic questions:

1.1. The concept of statistics as a social science, its emergence and development.

1.2. The subject and methodological bases of statistics.

1.3. The stages of statistical research and specific methods of statistical analysis.

1.4. The concepts and categories in statistics.

1.5. Organization and tasks of statistics in modern conditions.

1.1. The concept of statistics as a social science, its emergence and development

The term "statistics" derives from the Latin words *status* (a certain state of a phenomenon, the state of affairs) and *stato* (state). It was introduced into scientific use in 1749 by the German scientist G. Achenwall in a treatise entitled "Statistics" which described the political order of the states of Europe.

Historically, the development of statistics has been linked to the development of states, with the needs of public administration. Economic and military needs in the ancient epoch of human history demanded data on the population, its composition, property status. The first works of this kind were noted even in the sacred books of different nations [2; 3].

A system of economic and administrative accounting was developed in ancient Egypt. Land, food, labor, land use, income and loss statements were monitored.

Information on the population, its composition in terms of sex and age, land profitability and changes in trade was collected in ancient China in the 23rd century BC.

Birth records were organized in ancient Athens; young people under the age of 18 were enrolled in the military service and, after reaching the age of 20 were registered as full citizens. Land cadastres were drawn up, which included information on buildings, slaves, livestock, tools, income; there were descriptions of the states. A great merit in this respect belongs to the Greek philosopher Aristotle (384 – 392 BC) – he made a description of 157 cities and states of his time.

In ancient Rome, Servius Tullius created a qualification for a census of free citizens. Censorship officials – censors – established the name, gender and age of all family members, and listed their property. The founder of the Roman Empire Octavian Augustus, according to Tacitus, directed the compilation of a special book "Breviarium", which indicated the estimates of the appropriation, information about the state of the finances of the army, the number of citizens. This demonstrates the development of state-wide budget accounting. But it was in the private farms of Rome, where accounting extensively developed. The head of each family was obliged to keep a book of household property in terms of income and expenses.

The Middle Ages left a unique artifact – "The Doomsday Book" (1086) – a summary of the census of England and its property (including data on 240 thousand households). In Rus, in the 10 – 12th centuries, various information related to the taxation of the population was collected.

The growth of social production, the expansion of trade and international relations in the period of the formation of capitalism became an impetus for the development of accounting and statistics along with simple accounting in Italy (in the early 14th century.) There is a double accounting system, under which an operation is recorded twice. The need for analysis of the economic situation is increasing, so the volume of statistical information needed is particularly dramatically increasing. It is information on the size and location of industrial and agricultural production, commodity markets, labor markets, raw materials and more. Over time, the collection of data on mass societal phenomena has become regular. Since the middle of the 18th century there have been developed rules of population censuses; and the regularity of conducting censuses in developed countries was proposed thanks to the efforts of the Belgian mathematician, astronomer and statistician A. Quetelet (1796 – 1874). To coordinate the development of statistics, on the initiative of A. Quetelet, International Statistical Congresses were held; and in 1885, the International Statistical Institute was founded, which still exists today [34].

The expanding practice of accounting and statistical work in different countries contributed to the formation of statistical science.

Statistics as a science began to develop from the middle of the 17th century. There were **two main schools** in Europe at that time: a *mathematical school of statistics (school of political arithmetic)* in England and *a descriptive school of statistics (state administration)* in Germany [2].

A striking representative of the **mathematical school of statistics** (school of political arithmetic) was W. Petty (1623 – 1687), an English economist and statistician. He is deservedly considered the founder of statistical science, as he was the first who widely used mathematics for economic analysis. W. Petty's "Political Arithmetic" was based on accurate observation, quantitative description of economic phenomena based on the calculation of numerical data. The purpose of political arithmetic was to study social phenomena through numerical characteristics. They tried to identify patterns of development and correlation of economic phenomena through mathematical calculations, realizing the need to take into account the requirements of the law of large numbers in statistical studies, since the pattern can be found only with a large volume of statistical population.

The famous representative of the **descriptive school of statistics** was the German scientist H. Conring (1606 – 1681), a well-known German polyhistor, historian and national scholar who was a professor of medicine and politics in Helmstadt. G. Achenwall (1719 – 1772) is a well-known statistician, a professor at Goettingen University and the founder of the theory of statistics, the main principles of which were outlined in 1749. It was for the first time that Achenwall began reading the course of statistics at the University of Goettingen. Representatives of the German school of descriptors considered the main purpose of statistics to be a description of the territory, population, state system, religion, foreign policy, etc. Thus, the subject of statistics is not limited to those phenomena that have a numerical characteristic. Moreover, the early representatives of this area avoided the use of numerical data, and only later (in the middle of the 18th century) the numerical data gradually gained the right to be included in descriptive statistics. The second feature of the approach was that the work of school representatives lacked an analysis of the patterns and relationships inherent in social processes. Therefore, what descriptive statistics called statistics was far from actual statistics in its current sense.

Thus, what modern statistics borrowed from the German school of descriptors, is the system of verbal description of socio-economic phenomena without numbers; as to the English school of political arithmetic, it contributed the study of social phenomena by means of numerical characteristics, a statistical generalization of the characteristics obtained in order to identify patterns of development of the studied phenomena.

In the first half of the 19th century a **third school of statistics** emerged – **the mathematical statistics** (with A. Quetelet to be one of the founders of scientific statistics), whose representatives considered the theory of probabilities as the basis of statistics. The scientists who have developed the area under consideration are: A. Quetelet, the founder of the mean values doctrine; the Englishmen F. Galton (1822 – 1911) and K. Pearson (1857 – 1936), who used mathematical methods of statistics in biology; the Americans R. Fisher (1890 – 1962), W. Mitchell (1874 – 1948) and W. Gosset (1876 – 1937, alias – Student), who used the methods of probability theory in statistical studies [34].

The methodology of statistical surveys was developed and improved by Russian and national scientists: I. F. German (1755 – 1815); D. M. Zhuravsky (1810 – 1856), who studied the interaction of statistics and political economy, production statistical observation and analysis of statistics, developed the theory of grouping; Yu. E. Yanson (1835 – 1893), the founder of statistical analysis; O. I. Chuprov (1842 – 1908), who studied the issues of communication and dependence analysis of social phenomena, investigated the problems of dynamic series stability. One of the founders of the national demographic statistics is M. V. Ptoukha (1884 – 1961). M. V. Ptoukha as a population statistics theorist developed A. Quetelet's theory of the "middle man". As a methodist practitioner he developed a scheme for studying the demographic processes and proposed a method for implementation of this scheme. His findings in the study of marriage and mortality have gained worldwide recognition. Of great importance are his methods of constructing summary mortality tables [36]. Modern statistical methodology has also been developed in the writings of such statistic scientists as V. S. Nemchinov, S. G. Strumilin, B. S. Jastremsky, A. Ya. Boyarsky, T. V. Ryabushkin, S. S. Sergeev and others.

The development of statistical science, the expansion of the scope of practical statistical work have led to a change in the meaning of the term "statistics", various interpretations of which are given in Table 1.1.

In the modern sense, **statistics** is a set of summary information that quantitatively characterizes various aspects of social life:

production, distribution and exchange of goods, politics, culture, etc.;

practical activities for the collection, processing and analysis of quantitative data on public life and their publication;

a discipline that studies the quantitative side of mass phenomena and processes in an inseparable connection with their qualitative expression for the purpose of revealing the patterns of their development.

Table 1.1

The content of the concept "statistics"

Authors	The content of the term
O. Ye. Luhinin [19]	Statistics is a field of practical activity, economic science and a discipline for studying the methods of collecting, processing and analyzing the data on mass socio-economic phenomena and processes
A. T. Marmoza [23]	The term "statistics" means three related values: 1) numbers that characterize the levels, sizes and volumes of certain social phenomena; 2) a specific area designed to collect, accumulate, process and analyze data that characterize the population, economy, culture, education and other phenomena of social life; 3) independent social science aiming to develop methods for collection, summarization, processing, analysis and theoretical generalization of digital data on the phenomena of public life
S. S. Herasymenko, A. V. Holovach, A. M. Yerina, O. V. Kozyryev, Z. O. Pal'yan, A. A. Shustikov [33]	The word "statistics" (from the Latin *status* – the state of affairs) is synonymous with a set of facts, certain information about socio-economic phenomena and processes. The defining feature of such information is its quantitative characterization. Statistics is every science that integrates the principles and methods of dealing with mass numerical data
O. G. Osaulenko, O. O. Vasechko, M. V. Pugachova (statistical dictionary) [28]	Statistics is: 1) the field of knowledge, a complex and extensive system of scientific disciplines that have a particular specificity and study the quantitative side of mass phenomena and processes in an inextricable connection with their quality side. The main branches of statistics are: statistics theory which deals with the most general categories, principles and methods of statistical science; economic statistics which studies the phenomena and processes that take place in the economy; social statistics that analyzes social phenomena and processes; mathematical statistics, with the main tasks being statistical testing of hypotheses, estimation of distribution of statistical probabilities and its parameters, studying the statistical dependence, finding the basic numerical characteristics of random samples; 2) a field of practical activity covering the collection, processing, analysis and publication of statistical information on the phenomena and processes of public life; 3) a set of digital assets that characterize the state of mass phenomena and processes of social life (or their totality)

There are *three levels* in statistics, as a discipline [4; 5]:

• *general theory of statistics*, which involves the development of the terms, concepts and system of categories of statistical science, general principles and rules for conducting statistical surveys, universal methods of information processing – that is, creating a common methodology for statistical study of mass social phenomena;

• *economic and social statistics*. The economic statistics connects the study and quantitative evaluation of economic phenomena and processes, and development of synthetic economic indicators, such as: national wealth, gross domestic product, national income, etc. Social statistics is intended for research and development of general indicators in various fields of public life: education, health care, culture, politics, science and so on;

• *the branches of economic and social statistics*. Economic statistics is subdivided into such branches as statistics of industry, agriculture, transport, construction, trade, communications, etc. Social statistics includes statistics of science, law, health care, political statistics, population statistics, statistics of living standards and more.

The tasks of sectoral statistics are studying the mass phenomena that occur in the respective branches and spheres of socio-economic life as well as development of aggregate indicators of these industries and identifying trends and patterns in the development of these areas.

Statistics develops as a unified science, and the development of each branch contributes to its improvement as a whole [22].

Statistics is of great *cognitive importance* because it:

provides numerical and meaningful coverage of the phenomena and processes under investigation, serving as a reliable way to assess reality;

gives a validity of economic conclusions, making it possible to test hypotheses and separate theoretical propositions;

reveals relationships between phenomena, shows their specific form and strength;

reveals patterns of new phenomena development, giving them quantitative and qualitative characteristics.

Thus, **statistics** as a social science develops the methodology of statistical research used by other sciences. Knowledge of statistics is necessary for a modern expert to make decisions in conditions where the analyzed phenomena are affected by chance, to analyze the elements of a market economy, to forecast and develop scenarios of behavior of economic systems by changing the conditions of their functioning.

1.2. The subject and methodological bases of statistics

The subject of the statistics history is the process of the emergence and development of statistical accounting and statistics theory in the concept of statistical science. The development of statistics is determined primarily by the development of society and the state, their socio-economic needs [34].

In the early 19th century there was a need to develop and refine statistics methods that would allow the collection and compilation of mass credentials. This required the creation of a statistics theory. The founder of statistics theory is A. Quetelet (1796 – 1874). On the one hand, he was the founder of the science of society – social physics or sociology, on the other – the founder of the statistics theory. His main works are "The Propensity to Crime" and "Letters addressed to H.R.H. the Grand Duke of Saxe Coburg and Gotha: on the Theory of Probabilities, as Applied to the Moral and Political Sciences". The success of these works with contemporaries is explained by the fact that the scientist managed to show on the mass factual material the presence of social life regularities and the example of the topical issues of society [42].

Both previous lines of statistical thought development are synthesized in Quetelet's works. As a representative of the descriptive school he defined statistics as a science that studies the state, but claimed that statistics is a science of society – "social physics", which deals with the study of the life "social body". Quetelet proved the existence of stable patterns inherent in social phenomena. In order to explain the reasons for the formation of these laws he proposed the doctrine of constant and perturbation (individual, acting unevenly, but in different directions) causes under the influence of which the phenomenon occurs. He considered the combined effect of both reasons to be the most important goal of the subject of statistical knowledge. According to Quetelet, individual differences caused by random reasons could be compensated by the mass of data [22; 38; 39].

Quetelet proposed average values as the main method of statistical analysis. In many ways, the scientist contributed to the census in terms of the organization and development of a methodology for conducting a census. The object of a census was establishing the existing population. The observation unit is a person or a family. The recommended frequency of conducting a census was once every 10 years.

A. Quetelet also substantially contributed to such principles of statistics as:

- the subject of statistics is objective laws that determine the development of society;
- all phenomena are formed under the combined operation of independent, common and individual causes;
- the basis of the methodology of statistical cognition is mass observation and generalizations that ensure cancellation of randomness;
- the use of average values is the most important technique that allows you to determine the actual types of phenomena being studied;
- probability theory is the theoretical basis of statistics. It reveals the effects of common and individual causes, allowing the assessment of reliability by generalization of statistical indicators;
- possible and necessary collection of mass statistics based on the example of the population census.

For half a century, statisticians of different countries have been divided in their attitude to the ideas of Quetelet. In Germany, the position of the anti-quetelets was stronger. For example, the famous psychologist M. Drobisch (1802 – 1896) believed that the main task of statistics is not to list the number of crimes committed, how many products are produced, but to answer the question: why crimes were committed, why the required quantity of products was not released, etc. [45].

The German scientist, a representative of G. Rümelin's descriptive school, claims that the word "statistics" has two different meanings and defines two different sciences: statistics as a science that studies quantitative methods used in any field of human activity; statistics as state studies or social science.

The English statistician A. Bowley (1869 – 1957) was formed under the influence of Quetelet's ideas. He combined the ideas of political arithmetic with the latest tendencies associated with the use of mathematical methods, especially probability theory. Statistics was interpreted by him as the science of the average [34].

The French sociologist F. Le Play (1806 – 1882) succeeded in avoiding the influence of Quetelet's ideas. For him, the purpose of statistics was to accurately observe or describe social facts, the best source of which he considered to be experts, "social authorities". Le Play introduced the monographic method of socio-economic statistics, emphasizing that the depth of the study is reached by the fact that not all phenomena are considered, but only typical cases.

The ideas of Le Play contributed to the formation of two main schools of budgetary statistics of the 19th century:

- Belgian, which focused on the pursuit of rational food consumption;
- German, which was looking for patterns in the budget structure.

E. Engel and A. Schwabe can also be distinguished among the representatives of the German school. Engel is credited with defining a unit of observation in budget statistics: since families differ in size and composition, their budgets are incomparable, and therefore a new consumer unit is needed.

A big achievement of budget statistics was the derivation of two laws:

- Engel's law – as family income increases, their spending on food increases absolutely but decreases relatively;
- Schwabe's law – with increasing family income, their housing costs rise absolutely but decrease relatively.

There were no specialized statistical institutions in almost all countries in the early 19th century. Only in Sweden, in 1748, the Reporting Commission was created – a state institution that organized statistics about population in the country. At the turn of the 18 – 19th centuries a statistical office was established in France. The first departmental statistical agency, the Central Statistical Commission, appeared in Belgium in 1841 at the insistence of Quetelet.

The bulk of the work on practical statistics was performed by administrative bodies or individual research institutions. The establishment of specialized statistical institutions testifies to the high assessment of the importance of statistics and recognition of this science as a matter of national importance. The greatest progress was made in the field of population statistics. As the census developed, the tendency towards differentiation and specialization of statistics appeared as a source of information.

By the end of the 19th century a lot of statistical material had been accumulated, following which yearly collections began to be published, including dynamic series of key indicators.

The development of international relations of statistics was facilitated by the creation of the American Statistical Association in 1839. Its purpose was to collect, store and process statistical information in different countries.

At the end of the 19th century theories of stability, correlation, and regression emerged. The theory of stability, authored by the German statistician

W. Lexis (1837 – 1914), emerged as an attempt to explain the repetitiveness of data in the dynamics, discovered by A. Quetelet. In his work "On the Theory of the Stability of Statistical Series" published in 1879, Lexis outlined the theory of stability which had given rise to many followers and opponents, and influenced the development of statistics in the 20th century.

W. Lexis understood the stability of a time series as a deviation coincidence of the level of the dynamic range from the general average. In his theory (unlike Quetelet, to whom normal distribution seemed ubiquitous), normal distribution is interpreted as the result of data classification. To eliminate the variation of data caused by external causes and to achieve normal statistics stability, the scientist proposed to use the method of least squares.

W. Lexis also introduced a distinction between the *types of dynamics*:

*evolutionar*y – the main manifestation of major trends;

undulatory – undulating development in time;

periodic – correct repetition of waves;

oscillatory – random fluctuation of levels.

The researcher developed formulas that made it possible to measure the *variance of three types*: *general, intragroup and intergroup*. Also, the scientist made a significant contribution to the development of demography methods, improvement of mortality tables. The *graphical demographic analysis method* he created was called the *Lexis demographic grid*. The followers of W. Lexis were V. I. Bortkevich, A. A. Chuprov, A. A. Markov.

In the 19th century, the main features of the statistical method were finally determined: mass observation, data generalization, analysis. In the future, the ideas of F. Galton, K. Pearson and W. Lexis received a theoretical generalization in the concept of the statistics stochastic theory.

For the first time the methods of statistics in solving problems of the new evolutionary theory were applied by F. Galton, C. Darwin's cousin. He linked correlation with the essence of evolutionary doctrine based on the example of succession of growth, and also studied the connection of anthropological variables with the intellectual abilities of the human.

In 1896, Galton defined correlation, constructing a theoretical model of change for variables, introduced the notion of the regression line and the R correlation *coefficient*. The theoretical substantiation of F. Galton's conclusions and their practical verification belong to K. Pearson.

Thus, the works of scientists of the 19th century prepared rapid development, justification and application of mathematical and statistical methods.

Modern statistics is a special science that has its *object, subject and specific research methods*.

The **object** of studying statistics is mass phenomena and processes of any nature, including the economy. This means that statistical indicators are always a result of generalization of some set of facts [22; 45; 50].

A mass phenomenon is characterized by such *attributes* as:

- availability of many individual elements, each element being a really existing material object, that have some common properties, which are referred to as statistics features;
- variation, that is, the difference in the numerical values of population individual units;
- partial or complete independence of the elements from each other.

Statistics examines mass socio-economic phenomena in specific place and time conditions.

The **subject** of statistics is the size and quantitative correlation of mass social phenomena in an inseparable connection with their qualitative aspect, in order to identify patterns of their development [22; 45; 50].

The subject of statistics has three main features:

statistics studies social phenomena, so it is a social science;

statistics examines the quantitative side of socio-economic phenomena and processes in relation to their qualitative nature. This feature distinguishes statistics from mathematics, since statistics operates on concrete rather than abstract numbers;

statistics studies mass phenomena and processes, because patterns of development can be detected only in mass observations.

Statistics studies its subject on the basis of a special methodology, which is a set of general rules (principles) and special techniques and methods of statistical research.

General rules for statistical research are based on the tenets of economic theory and political economy. They form the *theoretical basis of statistics*.

Economic theory and political economy explain the essence of phenomena or processes under research, the laws of their development in specific conditions. At the same time, statistics enriches the socio-economic sciences with the numerical data used to confirm or refute the theoretical hypotheses.

The **methodological basis of statistics** is the dialectical method of knowledge and specific methods of statistics, according to which statistics studies mass phenomena in their relationship, movement and change, revealing their quantitative and qualitative characteristics. To do this, special statistical methods and techniques as for example summarization and grouping, the index method, the correlation and regression method, etc. are to be used [3; 42; 50].

Statistics relates to mathematics. They use common methods for processing and evaluation of data, but have different subjects of knowledge. Mathematical statistics examines patterns of mass phenomena in abstract form. Statistics as a social science characterizes the size and correlation of social phenomena in specific conditions of their existence and development. A prerequisite for using statistical methods in a particular study is determination of the nature of the phenomenon under study, its essential properties. Theoretical analysis gives a comprehensive understanding of the nature and logic of the object of knowledge. This is the objective basis of methodological decisions. The features of the statistical methodology are related to accurate measurements and quantitative description of mass social phenomena, using summaries to characterize objective statistical patterns.

Statistical analysis of mass phenomena and processes is a necessary link in the system of the economy management and the state as a whole. Feedback is carried out with the help of statistics, that is, the flow of information goes from the object to the management entity, i.e. to the management of enterprises, associations, territorial, branch and central authorities. Effective management decisions are impossible without comprehensive and timely information.

1.3. The stages of statistical research and specific methods of statistical analysis

In the general sense, research refers to the development of knowledge or systematic inquiry to establish facts. A **statistical survey** is collection of information and the study of the cause and effect processes in the observed phenomena in order to obtain information for making sound management decisions. The ultimate goal of a statistical survey is to identify trends in the development of socio-economic phenomena and processes at all levels of the economy [34].

Information support for a statistical survey includes the following **stages**.

The first stage. The *information component of statistical research*. This stage envisages identification of information media within the chosen research goal in order to select sources and start collecting initial information.

The result is streams of raw information that accrue during the process of collection, creating an array of primary and secondary information: *primary information* – the raw data obtained from the researcher directly from the object of study; *secondary information* – analytical information obtained by a researcher from the work of other authors or surveys. Thus, programmatic, methodological and organizational aspects of statistical observation of the study object are developed.

The second stage. Formation of information support for statistical information processing. It includes: selection of qualified personnel for collection and further processing of statistical information; development of methodology for collecting and processing statistical information in accordance with the conducted statistical survey; adaptation of available information technologies to the objectives to be gained in the statistical survey; conducting in-depth and comprehensive analysis of statistics using application packages and generating the output of the processed data in the form of databases; systematization and formation of streams of processed information for further use.

The third stage. Evaluation and interpretation of statistical survey results. The flows of the processed statistical information are compared, verified and evaluated in accordance with the stated purpose and objectives of the statistical survey. The final results are presented in the form of analytical conclusions or reports, which analyze the main development trends of the research object. Means of information visualization of the results of statistical analysis are used for the sake of clarity of the obtained results. Informed management decisions are made based on the combination of these types of interpretation of information.

Therefore, it is advisable to provide information support for a statistical survey in stages. This will clearly state the goal and object of the study, justify the choice of methods by which the analysis will be conducted, and the results obtained will accurately show the main trends of socio-economic processes.

Based on the theoretical grounds, statistics applies *specific methods* that correspond to *three stages (phases) of statistical research*:

- statistical observation, which consists in collecting primary information about particular facts of the phenomenon under study;

- grouping and collating of the collected material, which allows it to be organized and classified. This is done on the basis of the transition from the characteristics of the individual elements of the study to the generalizing indicators in the form of absolute, relative and average values;
- processing of statistical indicators obtained as a result of grouping and summarizing, analysis of the results obtained in order to form conclusions about the state and patterns of development of the phenomenon under study. At this stage, scientifically sound conclusions are drawn, aimed at the development of measures and making effective management decisions.

The stages of a statistical study are combined for the purpose of the study. Each of them uses those methods that can give a deep and comprehensive characterization of the phenomena under study. Thus, in *the first stage*, you use the method of mass statistical observation, which ensures the comprehensiveness, completeness and representativeness of the initial information, that is, provides an information base for statistical generalizations and characterization of objective regularities. Statistics have indisputable probative force precisely because they are not based on individual facts but on their totality [4; 38; 50].

In *the second stage*, the mass observation data are summarized: the elements of the population are classified according to certain characteristics by the methods of summarization, classification and statistical grouping. For example, births can be classified based on gender and place of birth, while coal mining in terms of the mine or date. The statistical population ordered in this way is called *a statistical series.* Depending on the classification method there are distribution and dynamics series. *The distribution series* is the result of classification, grouping of a set of elements in statics (as of a certain moment or over a certain period of time). Grouping distinguishes characteristic properties and various types of phenomena. *The dynamics series* presents the value of statistical indicators in time (by periods or moments of time), describe the dynamics of mass process development.

The third stage is the analysis and interpretation of the statistical survey results: a comprehensive quantitative and qualitative analysis of mass phenomena and processes on the basis of methods of studying the variation of features, the use of dynamics indicators, the index method, establishing trends and forecasting development, identifying relationships and interdependencies between the features.

Statistical methods are adapted to the peculiarities of various social phenomena and processes to be studied. But in any study, the peculiarities of a statistical method are specific – mass of data, quantitative measurement, generalization.

Analytical potentials of statistical methods are expanded by the use of a compact and rational form of presentation of the information generalization results as well as the analysis of identified patterns. Such forms are statistical tables and graphs [1].

1.4. The concepts and categories in statistics

Statistics studies its subject through certain *categories*, that is, concepts that denote the most general and essential properties, characteristics, connections, and relationships of objects and phenomena of the objective world.

There are five basic concepts in statistics:

- a statistical population;
- a statistical unit (unit of statistical population);
- a statistical feature;
- a statistical regularity;
- a statistical indicator.

A *statistical population* is a set of elements (units of a population) that have common features or characteristics. Each element of a population is characterized by certain values of the feature that varies. This may be the population of Kharkiv, the students of Simon Kuznets Kharkiv National University of Economics and so on. A statistical population is the object of statistical study.

The proposed definition of a statistical population helps to identify its main *properties*:

- *indecomposability* – partial appearance or partial disappearance of elements of a statistical set does not destroy its qualitative basis; all its qualitative characteristics are preserved. Students of Simon Kuznets KhNUE, as a statistical population, will not change their qualitative characteristics, regardless of the fact that some students (graduates) annually leave higher education institutions, and part (freshmen) join them. This property is based on the mass of the statistical population;
- a statistical population is *formed as a set of objects, phenomena*, united by a common feature. All elements of a statistical population have this

common property. However, the general does not mean the same. The values of a common characteristic in different units of a population as a rule differ from each other. The homogeneity of a statistical population is established in each specific statistical survey according to its goals and tasks;

• *variation* – a quantitative change in the value of a statistical feature with a transition from one element to another. If the values of a feature are the same in all elements, it makes no sense to study the whole statistical population, it is enough to consider only one element of it to gain knowledge about the whole phenomenon. Variation arises under the influence of a certain set of conditions and causes. Statistics does not address these causes. This is a prerogative of a special economic discipline, while statistics quantitatively assesses the impact of each cause on the variation of a particular feature, which makes it possible to take into account a specified influence on the management decisions at different levels.

All statistical populations can be divided into *groups*:

• *created by life*, they create a unity, regardless of whether they are subject to statistical research (a statistical population of the enterprise employees, a statistical population of Kharkiv industrial enterprises, etc.). These are really existing statistical populations, they have a certain specific size;

• *created* specifically *for statistical research goals* (a population of buyers of a specific product in marketing research);

• *stochastic populations* (hypothetical sets) – imaginary, unrealistic, predictable sets (a population of celestial bodies in a galaxy; a population of infinitely large numbers of coins falling heads or tails).

A *statistical unit* or a *unit of a statistical population* is an indecomposable primary independent element of a statistical population that is a carrier of a certain statistical feature. A unit of a statistical population is the limit of fragmentation of the population which retains all the properties of the studied phenomenon or process as a special case. The choice of a population unit is determined by the goal and level of the statistical survey being conducted. For example, you can study productivity at the level of an industry, an enterprise, a workshop, a site, a team. In each case, the units of a population will be different: an enterprise of an industry, an employee of a certain enterprise; a workshop worker, a crew. Units of a population must necessarily be qualitatively homogeneous.

The total number of units of a population is called *the size of a statistical population.*

A *statistical feature* is a characteristic property, a certain qualitative nature of the units of a statistical population. For example, the statistical characteristics of enterprises may be: the type of ownership, the number of employees, the value of authorized capital, the value of assets, etc. Statistical features can be classified on a variety of grounds (Table 1.2).

Table 1.2

Classification of statistical features [34]

Grounds for classification				
The kind of expression of a statistical feature	The nature of the variation	The relation to time	The nature of the relationship	The degree of influence
Attributive (qualitative), quantitative	Alternative, discrete, continuous	Moment, interval	Factorial, effective	Essential, nonessential

The *kind of expression* implies attributive and quantitative features:

• *attributive (qualitative, descriptive) features* are expressed in terms of concepts, names, i.e. verbal means. For example, gender, nationality, education, and more. After that, you can get summary information about the number of statistical units that have a certain value of a feature;

• *quantitative features* are expressed numerically (age, length of service, sales, income, etc.). They can summarize the number of units that have a specific attributive value, and the total or average value of the attribute in the population.

According to the *nature of variation*, the features are divided into:

• *alternative* that can take only one of two possible values. These are features of owning something. For example, sex, marital status; in marketing or political science, the answer to the question is "yes" or "no";

• *discrete* – quantitative features that take only single values, with no intermediate ones; usually an integer. For example, the number of workers, the number of children in the family, the number of cars, etc.;

• *continuous* – quantitative features that take any value within a certain range. As a rule, they are rounded according to the accepted accuracy (for example: accounting profit on the balance in hryvnias, tax from tax registers – in thousand hryvnias).

Depending on the *relation to time* there are:

• *moment* features that characterize units of the population at a critical time. For example: the cost of fixed assets (CFA) is determined as of January 1 and December 31 of a certain year as the value of the CFA at the beginning and end of the reporting year;

• *interval* features that characterize a phenomenon over a period of time (year, quarter, month, etc.). For example, daily revenue, annual sales, etc. [55].

In terms of the *nature of the relationship* the features are divided into:

• *factorial features (independent variables)* that cause changes in other features or create opportunities to change the values of other features. Factorial features are subdivided respectively into *features of cause* and *features of condition*;

• *effective features* (*dependent variable*s), depending on the variation of other features. For example, the value of output is the effective feature whose value depends on the factorial features – the number of employees and productivity.

According to the *degree of influence* on the statistical unit under study, the features are divided into:

• *essential* (main), that is, features that are central to the phenomenon being studied. For example, for an enterprise it may be the number of products produced, the number of employees;

• *nonessential*, which are not directly related to the essence of the phenomenon under study. For example, the name of the enterprise, its territorial identity.

An important category of statistics is statistical regularity. As a philosophical category, any regularity is a form of manifestation of a causal relationship, expressed in the sequence, regularity, recurrence of events with a high degree of probability, if the causes giving rise to events do not change or change slightly [55; 57].

Statistical regularity is a form of manifestation of regular relationship in mass phenomena and processes that are influenced by many factors that are constantly changing. Statistical regularity is revealed at the level of the statistical population and established by statistical methods. Regularity arises as a result of the influence of a large number of permanent and accidental causes. *Permanent causes* provide regularity and repeatedness of changes in the phenomena while accidental ones cause deviations in this regularity.

At the level of statistical units, regularity is not always obvious. For example, life expectancy for women is known to be longer than for men; but this does not mean that every woman lives longer than a man (there are more long-lived men) [23; 38].

Since statistical regularity manifests itself as a result of mass statistical data, it causes its correlation with the law of large numbers – statistical regularities are a consequence of the operation of this law. *The law of large numbers*, in its simplest formulation, states that in mass phenomena and processes, accidental secondary features in the observed units are mutually cancelled, resulting in a clear manifestation of the most significant features and patterns of development of such phenomena [55; 57]. Thus, the law expresses the dialectic of the casual and the necessary. For example, 104 – 106 boys are born per 100 girls; however, in different families and even small settlements, this ratio may be completely different. In accordance with the nature of mass regularity, trends revealed by the law of large numbers are valid only as mass trends, but not as laws that concern the stable, general nature of the cause and effect of phenomena. In a statistical pattern, these relationships are less stable, not general in nature, but relate to a certain space and time, valid only for certain conditions of existence of the studied phenomenon.

The term "statistical regularity" was first used in the natural sciences (as opposed to the term "dynamic regularity") to express a form of relationship in which the specific values of any factors always strictly correspond to certain values dependent on the features of these factors. In the case of a dynamic regularity, the quantitative relationships between the values remain valid for each individual statistical unit. For example, the area of a circle changes with the change of its radius, and this dependence is expressed by the formula $S = 2\pi r^2$, which holds for any circle.

Statistical indicators are used to describe the studied mass phenomena and processes as a tool for learning these phenomena. *A statistical indicator* is a generalized numerical characteristic of any mass phenomenon (process) with its qualitative certainty in specific conditions of place and time. This differs from individual values of features (option). For example, the average wage of employees of a particular enterprise over a period of time is a statistical indicator, and the wage of a particular employee is an individual value of the features (option). In contrast to the individual value of a feature, a statistical indicator can only be obtained by calculation. This can be a simple calculation of the units of a population, summing up their values of the features or more complex calculations. For example, the number of employees at the enterprise at the beginning of the year, the quarterly volume of production, the cost, the profitability of production, etc.

According to the definition, a statistical indicator has qualitative and quantitative aspects. The qualitative side of a statistical indicator is determined by the feature to be examined and displayed in the name of the indicator, quantified in the numerical value of the indicator.

In practice, statistical indicators can be used to describe various aspects of socio-economic phenomena and processes which can be classified in a certain way (Table 1.3).

Table 1.3

Classification of statistical indicators [34]

Grounds for classification		
Function performed	Units covered	Forms of expression
Planned, accounting, prognostic	Individual, summarized	Absolute, relative, average

Planned indicators characterize the directive function, they are focused on the objectives required. *Accounting* shows the real state of the phenomenon under study. *Prognostic* is its possible status in the future.

If a statistical indicator refers to a particular phenomenon (for example, an individual enterprise), it is called *individual*. If an indicator characterizes the totality of phenomena (for example, homogeneous enterprises in the region), it is called generalized, or summarized.

Statistics also uses a system of statistical indicators – a set of interconnected summarized data arranged in a logical sequence.

An absolute indicator is the initial primary form of expression of statistical indicators. *Relative indicators* are derivative, secondary indicators relative to the absolute ones, expressing certain relationships between the quantitative characteristics of statistical populations. *Averages* are the most common form of statistics that characterize the typical level of a phenomenon. Average indicators are calculated per unit of a statistical population.

1.5. Organization and tasks of statistics in modern conditions

The beginning of organization of international statistics was initiated by a group of European statisticians who, in 1853, proposed to hold an International Statistical Congress in order to develop, on a scientific basis, uniform methods, common rules and programs for joint work. For the organization of the convening of the Congress much effort was made by the outstanding statisticians of that time A. Quetelet and E. Engel. In order to develop this international forum, the International Statistical Institute (ISI) was established in 1885. According to the Statute, ISI's main task is to develop and improve statistical methods in different countries of the world. In fact, the ISI is pursuing a course for achieving international comparability of statistical indicators by developing identical methods for calculating them and developing common classifications of such indicators in different fields of statistics in order to exchange views between scientists and practitioners of statistics and their organizations [34].

ISI is an autonomous international organization, but it has consultative status under UN ECOSOC and UNESCO. It is not part of the UN specialized agencies, but it regularly participates in the work of the UN Statistical Commission without being a member. The ISI's practice is coordinated by its executive body, the Permanent Bureau, located in Voorburg (Netherlands).

The modern mechanism of international statistics traces a sort of division of labor: some statistical organizations develop statistical methodology, observation programs; others directly collect, process, and publish statistics. The first type of organization includes those that have previously held International Statistical Congresses (ISCs) and the current International Statistical Institute. The second type is the UN Statistical Office and its specialized agencies and statistical offices of other numerous international organizations.

The main practical activity in the field of international statistics is carried out by specialized UN services. The largest volume of work is carried out by the United Nations Statistics Commission and the Statistical Office of the UN Secretariat, which serve as the secretariat of the Commission. The UN Statistical Commission (established in 1946) is one of the so-called functional commissions of the United Nations Economic and Social Council (ECOSOC), which directs and controls the activities of all UN departments in the field of statistics.

Today, the Statistical Commission operates in the following areas:

- analysis of individual countries' experience in the development of statistical methodology;
- improvement of data comparability and development of international standards;
- the use of computers in international statistics and the development of integrated systems for collection and processing of international information;
- development of measures to assist developing countries in the development of national statistics.

The ultimate goal of the work of the Statistical Commission is to form a unified system for collecting, processing and disseminating international statistical information by UN bodies and agencies.

The functions of the Statistical Office of the UN Secretariat are:

- preparation of all activities for the sessions of the Statistical Commission;
- preparation and final processing of methodological reports and findings from countries and various statistical organizations; collecting, analyzing, publishing and updating statistics obtained from Member States and UN specialized agencies;
- improvement of comparability of data through organization of additional studies and calculations; coordination of UN statistical activities, specialized agencies, national statistical and international economic organizations;
- promoting the development and improvement of statistics as a whole and above all in developing countries; preparation of necessary materials and holding of sessions of the Statistical Commission.

Statistical bodies of several international organizations produce *statistical collections*. Considerable work has been done to build a system of key interrelated indicators of the national economic process as a whole – a system of national accounts – and to compare its elements with the balance of the national economy of individual countries. The recommendations of the UN Statistics Commission on the standards of National Accounts show the impact of national best practices on the balance sheet development. This is evidenced in particular by attempts to combine national accounts and financial statements, proposals to include national wealth and labor indicators in the System of National Accounts, some changes in the methodology for calculation of the national income, etc.

The most important statistical collections of the UN and its specialized agencies since 1946 have been "Statistical Yearbook", "Demographic Yearbook", "Yearbook of International Trade Statistics", "Yearbook of the United

Nations" and others. The Food and Agriculture Organization of the United Nations (FAO) publishes the "Yearbook of Food and Agricultural Statistics" and the Fisheries and Forestry Statistics; the United Nations Educational, Scientific and Cultural Organization (UNESCO) publishes "International Yearbook of Education" and the summary "Statistical Yearbook" [34].

The efforts of many countries have created a *Global Statistical System*, which includes:

- the UN Statistics Commission;
- sectoral UN statistical units;
- the system of statistical publications of the UN and other international organizations;
- the United Nations special agencies: FAO – the UN Food and Agriculture Organization; UNESCO – the United Nations Educational, Scientific and Cultural Organization; WHO – the World Health Organization; WB – the World Bank, IMF – the International Monetary Fund; WTO – the World Trade Organization; ILO – the International Labor Organization;
- regional statistical organizations, such as the Statistical Office of the European Communities (Eurostat), the Statistical Committee of the Commonwealth of Independent States (CIS).

There is no strict subordination between the statistical authorities, but the UN Statistical Commission has an official status as the "first among equals". It coordinates the general list of international standards and classifications and is responsible for their transmission to a number of countries.

The main purpose of the Global Statistical System is the effective use of resources for statistical activities at national and international levels. The websites of international statistical organizations are listed in Table 1.4.

Table 1.4

The website addresses of international statistical organizations

Name of the organization	Website address
1	2
Statistical Office of the European Union	https://ec.europa.eu/eurostat
Statistical Committee of CIS	http://www.cisstat.com
International Labor Organization	https://www.ilo.org
Organization for Economic Cooperation and Development	http://www.oecd.org/statistics

Table 1.4 (the end)

1	2
Database of the Food Organization at the United Nations	http://www.fao.org/statistics
International Monetary Fund	http://www.imf.org
UNESCO Institute for Statistics	http://data.uis.unesco.org
Economic Commission for Latin America and the Caribbean at the United Nations	https://www.cepal.org
United Nations Statistical Website	http://unstats.un.org/unsd
United Nations	http://www.un.org
Economic Commission for the Asia Pacific Region of the United Nations	http://www.unescap.org/stat
United Nations Industrial Development Organization	http://www.unido.org
World Health Organization	http://www.who.int/whosis/menu.cfm
World Bank	http://www.worldbank.org
World Trade Organization	http://www.wto.org

The system of statistical bodies of Ukraine corresponds to the state system and administrative and territorial division of the country.

Organization of statistical work in various fields is carried out by the State Statistics Service of Ukraine (SSSU), headed by the Chair, who is appointed to this position and dismissed by the Cabinet of Minister of Ukraine upon the recommendation of the Prime Minister of Ukraine put forward on the basis of proposals by the Minister of Economic Development and Trade.

The activity of the Service is regulated by the Regulation on the State Statistics Service of Ukraine, approved by the Cabinet of Ministers of Ukraine. Other permanent or temporary consultative, advisory and other subsidiary bodies may be formed to consider scientific advice and conduct professional consultations on major issues of activity in the State Statistics Service. The decisions on setting up or liquidation of the board, other permanent or temporary consultative, advisory and other auxiliary bodies, their quantitative and personnel composition, regulation are approved by the SSSU Chair [95].

The subdivisions of the State Statistics Service of Ukraine are:

the Department of National Accounts and Macroeconomic Statistics;

the Department of Production Statistics;

the Department of Structural and Enterprise Finance Statistics;

the Department of Service Statistics;

the Department of Agriculture and Environment Statistics;

the Department of Household Surveys;

the Department of Trade Statistics;

the Department of Labor Statistics;

the Department of Population and Regional Statistics;

the Department of Price Statistics;

the Department of Information Technology;

the Statistical Planning and Coordination Department;

the Department of Statistical Infrastructure;

the Department of Dissemination of Information and Communications;

the Department of Financial and Economic Support;

the Department of Personnel Management and Organizational Support;

the Office for International Cooperation and European Integration;

the Accounting and Reporting Department;

the Administration Department;

the Legal Support Department;

the Security Department;

the Internal Audit Sector;

the Chief Specialist on Prevention and Detection of Corruption.

The State Statistics Service of Ukraine conducts its work through the relevant territorial bodies – Main Departments of Statistics in each region. The lower level organizations of national statistics, which are under the control of the statistical department, are district and city departments of statistics.

One of the most important institutions in shaping the official statistics of the country is the *National Bank of Ukraine*, which ensures the collection, compilation and dissemination of data of the financial and external sectors of the economy. These data include: monetary and financial statistics (reviews of financial corporations, deposits, loans, interest rates of deposit-taking corporations, securities, financial accounts of the financial corporations sector); balance of payments statistics, international investment position and external debt; international reserves and exchange rates; statistics of financial sustainability indicators [97].

The National Academy of Statistics, Accounting and Audit and the state enterprise *Information Analytical Agency (Gosanalitinform)* are functional bodies of State Statistics. The purpose of the Gosanalitinform activity is creation of conditions for providing statistical information for citizens, state agencies, enterprises and organizations, other juridical entities and individuals of Ukraine and abroad in the way of realization of works and performing of services on the payment basis.

The main task of all statistical authorities is to collect, validate, develop, summarize, analyze and timely submit scientifically substantiated statistics to legislative, executive, management and economic bodies. Such information characterizes the economic and social development of the country, the implementation of national and regional programs, the increase in the efficiency of social production, economic reforms, the use of natural, labor and material resources, the dynamics of the living standards of the population and so on.

The principles of organization of statistics according to the Constitution of Ukraine are determined by the laws of Ukraine. The Law of Ukraine "On State Statistics" [99] regulates legal relations in the field of state statistics, defines the rights and functions of state statistics bodies, organizational principles of carrying out state statistical activities in order to obtain comprehensive and objective statistical information on economic, social, demographic and environmental situation in Ukraine and its regions and the protection of the state and society. The latest revision of this Law was held on April 19, 2014 in connection with the adoption of the Laws of Ukraine "On Information" and "On Access to Public Information" [98].

Thanks to the legal framework created, it has become possible to approximate statistics to international standards and world practice. The most significant steps in this direction were: implementation of the National Accounts System; indicators of the standard of living of the population, foreign economic activity, functioning of the labor market, new structures of entrepreneurial activity; consumer price index calculation (inflation).

The first stage of reform, which was largely completed in 1997, was characterized by a concentration of resources in the most important areas of statistics or in areas that were not in line with international practice, in accordance with the State Program of Ukraine's Transition to the International System of Accounting and Statistics. During 1993 – 1997, the basic elements of the system of national accounts and balance of payments, foreign trade statistics, monetary and banking statistics, new labor market statistics were created. In accordance with international standards, the calculation of inflation indices has been introduced, industry statistics have been significantly improved, the financial condition of enterprises and organizations (including such new entities as trusts, insurance companies, extra-budgetary funds, etc.)

has been monitored. Stock market statistical survey was organized, the Unified State Register of Enterprises and Organizations of Ukraine (EDRPOU) was created, the main provisions of the State system of classification and coding of technical and economic and social information were developed, work was started to create a bank of statistics and computer network of state statistics.

Today, national statistics of Ukraine is in the next phase of its reforms. On February 27, 2019, the Cabinet of Ministers of Ukraine approved the Program for the Development of State Statistics, developed by the Ministry of Economic Development and Trade of Ukraine jointly with the State Statistics Service by 2023 [95, 96]. The Program defines the strategic directions and tasks of the development of state statistics, as well as the expected results. It provides for the modernization of national statistics through the implementation of the EU international standards and regulations and the full transition to a process-oriented statistical information production system (GSBPM), as well as the use of information technology (special software for submission and processing of statistical reporting, electronic reporting designers).

Another important area of national statistics reform is to ensure the transparency and accessibility of statistical information. In particular, it means creating a modern, more user-friendly State Statistics Service web-portal and its mobile version, providing a wide access of users to microdata (regarding employment, household status, etc.), activating communication on the results of the activities of the State Statistics Service and its research data, as well as in general raising the level of statistical literacy of the society through relevant educational programs in secondary and higher education institutions, information in the media, etc.

The program envisages reducing the reporting burden on respondents who provide statistics for further analysis. This can be achieved through increased number of inconsistent statistical observations (sampling) and the introduction of modeling methods, the use of modern methods and technologies of data collection, the expansion of the use of administrative data (including the regional level), the introduction of free reporting by respondents to the state statistics authorities in electronic form.

Important innovations of the State Statistics Development Program include the launch of two new sample surveys of households: income and living conditions (EU-SILC) and household budget surveys (HBS), as well as the modernization of calculations of national accounts, including the calculation of

GDP based on purchasing power parity (PPP). This will ensure harmonization and will make it possible to compare the physical volumes of gross domestic product by end-use categories of EU countries and EU candidate countries.

An important component of the implementation of the Program is also the development of human capital and the enhancement of the professional competence of the staff of national statistics authorities. In particular, educational programs for training specialists in statistics have been harmonized with European standards, especially with the Master's Program in Official Statistics of the EU (EMOS), the staff qualification system has been improved and the study of the languages of the Council of Europe has been provided [95].

The tasks of statistics at the present stage of the national economy development are inextricably linked with the transformation processes in the country. The main tasks of statistics in Ukraine are:

a comprehensive study of the profound transformations of social and economic processes in society, based on scientifically sound indicators;

generalization and prediction of state development trends;

identification of reserves of social production efficiency;

timely provision of statistical information to the governing bodies;

development of designed measures to bring the national methodology of statistical research closer to the methodology and standards of international statistics;

formation and operation of a monitoring system (specially organized observations);

development and introduction of the national automated system of collection, accumulation, processing and analysis of statistical information;

improvement of statistical methodology and the system of statistical indicators, collection and processing of statistical information, economic and statistical analysis of socio-economic phenomena and processes.

Important concepts

Statistics is a social science that develops a methodology for statistical research used by other sciences.

The object of statistical study is mass phenomena and processes of any nature.

The subject of statistics is the size and quantitative relationships of mass social phenomena in an inseparable connection with their qualitative aspect in order to identify regularities of their development.

The theoretical basis of statistics is conceptual issues of economic theory and political economy.

The methodological basis of statistics is specific methods of statistics and the dialectical method of cognition.

Statistical population is a set of elements (units of a set) that have common features or characteristics. Each element of the set is characterized by certain values of the feature that varies.

Statistical unit, or unit of statistical population, is an indecomposable primary independent element of a statistical population that is a carrier of a certain statistical feature.

Statistical feature is a characteristic feature, a certain qualitative characteristic of a statistical population unit.

Statistical regularity is a form of manifestation of a regular relationship of mass phenomena and processes that are influenced by many constantly changing factors.

Statistical indicator is a generalized numerical characteristic of any mass phenomenon (process) with its qualitative certainty in specific conditions of place and time.

Questions for discussion at the seminar

1. Describe the main stages in the statistics development and explain what made the transformation of accounting to statistics as a science.

2. Create a list of indicators that can characterize: the country's population, consumer market, industry, banking institution.

3. Explain the essence of the stages of a statistical research.

4. Describe the thematic focus of online resources on international statistics.

Reference to laboratory work

The guidelines for carrying out the laboratory work on the topic "The subject, methods and tasks of statistics" are presented in [100]. The laboratory work is aimed at the formation of theoretical knowledge and understanding of methodological foundations of statistics through detailed consideration of the seminar problem.

Questions for self-assessment

1. What is the origin of the term "statistics", by whom and when was it introduced into scientific use?

2. Describe the stages of development of statistics.

3. What is statistics in today's context?

4. State the object and subject of a statistics study.

5. Name the main categories of statistics.

6. Define a statistical population and a unit of a statistical population. Give examples.

7. Describe the stages (steps) of a statistical survey.

8. On what grounds can statistical features be classified?

9. Give examples of moment and interval features.

10. What are the classes of statistics?

11. Describe the system of state statistics bodies of Ukraine and list the tasks of statistics at the current stage of development.

Questions for critical rethinking (essays)

1. Specify which populations can be chosen for statistical study at an enterprise or organization. Name the unit of the population. Justify your answer.

2. Specify the quantitative and attributive features that can characterize a population of students of a higher education institution, a population of insurance organizations, a population of high qualification workers, a population of countries – importers of fuel.

3. Specify the main feature that determine the variation of student's success, food industry profits, wages for workers in different sectors of the economy.

4. To what kind (quantitative or qualitative, discrete or continuous) do the following features belong: a) the number of a firm's employees; b) family ties of family members; c) sex and age of a person; d) social status of bank depositors; e) the number of children in a family; e) retail turnover of stores. Justify your answer.

2. Statistical observation

Basic questions:

2.1. The essence and tasks of statistical observation.

2.2. Forms, types and methods of statistical observation.

2.3. The plan of statistical observation.

2.4. Statistical observation errors.

2.1. The essence and tasks of statistical observation

In order to study socio-economic phenomena and processes, it is necessary to collect primary statistical data (information), which implies the accuracy of characteristics of mass phenomena and processes obtained as a result of statistical observation. These data are the starting material for generating summary indicators and conclusions on trends in their development.

The information collected about the phenomenon being studied is not always considered statistical. *Statistical information* must meet certain **requirements** [6; 17; 34]:

Completeness which means that it must:

cover all units of a statistical population or a part of them which makes it possible to draw conclusions about the population;

cover all significant aspects of the phenomenon, its properties, internal and external relations;

be collected for the longest possible time, which will help to reduce the influence of random factors and identify patterns of development;

reliability – the data collected in the process of observation should correspond to the actual state of the phenomenon nature;

comparability – in order to ensure further comparison of these data, they should be collected in a fixed time, according to a single program, using the same methods;

timeliness of submission (especially for making management decisions);

accessibility, which is particularly relevant in a market economy where non-governmental entities' low responsibility for reporting causes a breach of the requirements of completeness and data recording timeliness.

The main statistical information **properties** are **mass nature** and **stability**. The first property is related to the peculiarities of the statistics subject, the second one – to the invariability of the collected information, its ability to become obsolete and the need to obtain new information to form rational management decisions. Practical management requires constant updating of statistics. Information that is reliable, complete, but not timely becomes practically unnecessary for decision-making at any level of government.

Getting good statistical information depends a lot on what level it is collected. There are two **systems of collection of information in Ukraine:**

centralized (nationwide),

decentralized (departmental, provided by separate economic structures).

A centralized collection system has greater possibilities for quality observation, scientific methods, skilled personnel, technical security and more. However, a decentralized system is more operational due to the shorter time span between data collection and the use of ready statistical information. For a decentralized system, the problem of the scientific validity of statistical observation methods and their practical application is more acute.

The composition of statistical information is determined by the needs of social development. In the current economic conditions, the consumers of statistical information are both public authorities and various non-governmental structures. The main source of statistical information is the publications of national statistics bodies.

Provision of statistical information is the main task of the bodies of national statistics, while statistical information itself is the product of their activity, which, like any other, has its value. The highest value is given to information beyond the scope of the program of national statistics bodies' work.

The main source of statistical information and the first stage of statistical research is statistical observation.

Statistical observation is the systematic, scientifically organized collection of data or information about mass phenomena and processes, which consists in the registration of selected features for each population unit [34; 44; 75].

Statistics includes observations that study the statistical patterns that manifest themselves in mass processes, in a large number of units of a population by registering (recording) the relevant features of phenomena (processes). Thus, an observation is considered to be statistical if it provides for the recording of established facts in accounting documents for further generalization and meets such conditions as [34; 64; 77]:

planned nature;

mass nature;

regularity.

The *planned nature of a statistical observation* implies that it is carried out in accordance with a pre-designed plan. Such a plan includes issues of methodology, organization, technique of gathering information, quality control and reliability, registration of final results.

The *mass nature of a statistical observation* means that it covers so many cases of the phenomenon under study that is sufficient to obtain reliable statistics that characterize the population as a whole.

Systematic observation provides for its regularity. This makes it possible to study trends and patterns of socio-economic phenomena and processes.

Depending on the **level of registration**, statistical observations may be primary or secondary.

Primary observation is the registration of initial data coming from the object that produces them (current records of the number of registered marriages and divorces at the appropriate institution; population surveys on the process of property privatization).

Secondary observation is the collection of previously recorded and processed data (bank statements, audit results, stock exchange results).

The statistical observation **process** consists of the following **stages:**

Stage 1 – preparation of observation;

Stage 2 – mass data collection;

Stage 3 – preparation of the observation data for processing;

Stage 4 – development of proposals for improvement of the statistical observation.

In the *first stage*, the program-and-methodological and organizational issues of the statistical observation plan, the organization of the observation are dealt with.

The *second stage* is related to direct observation and includes such activities as: sending out forms, questionnaires, forms of statistical reporting, census letters; filling them in and submitting to the bodies conducting the observation.

In the *third stage*, the collected information is checked for completeness, subjected to arithmetic and logical control in order to detect and eliminate errors.

In the *fourth stage*, the causes of errors during filling in the statistical forms are analyzed, and proposals for improvement of the statistical observation are developed.

2.2. Forms, types and methods of statistical observation

Statistical observation as a method of research differs from other methods of collecting information by the nature and mass of data and the methods of obtaining them.

The information on the essence of the process of observation is given in Table 2.1 [25; 29; 34].

Table 2.1

Forms, types and methods of statistical observation

Forms of statistical observation	Types of statistical observation		Methods of statistical observation
	depending on the time of data registration	depending on the extent of coverage of units	
1. Statistical reporting. 2. Specially organized observation: a) censuses; b) recording; c) special surveys; d) interviews. 3. Register	1. Current (continuous). 2. Discrete: a) periodic; b) one-time	1. Complete. 2. Incomplete: a) selective; b) inspection of the main array; c) monographic; d) questionnaire; e) monitoring	1. Direct observation. 2. Documentary records. 3. Surveys: a) expeditionary; b) self-registration; c) correspondence method; d) questionnaire method

There are three organizational **forms** of statistical observation: statistical reporting, specially organized observation and registration [34; 46; 52; 75; 77].

Statistical reporting is the main form of statistical observation through which statistical authorities obtain the required data in a timely manner in the form of statutory reporting documents.

Reporting is characterized by such *properties* as obligatory nature, regularity, reliability.

Obligatory nature means that reporting is to be submitted by all registered entities, following the standardized form and the approved list of indicators and specifying their details: name, address, surname and signature of the responsible person, date of reporting.

Regularity involves regular, timely preparation and submission of reports within approved deadlines.

Reliability means that the data reported should be consistent with the facts and exclude any distortions. Subjects of activity are administratively and judicially responsible for the accuracy of the above data. It helps to prevent non-submission or misrepresentation.

Financial statements are prepared on the basis of initial operational and accounting data.

The reporting classification features are as follows:

in terms of approval and purpose:

external – approved and collected by the bodies of state statistics, ministries and departments;

internal – developed by the entities themselves for their own operational, analytical and management purposes. For example: analysis of market situation, evaluation of their own resources, forecasting of activity. Internal reporting is done based on the internal reporting data of multistructured entities, joint stock companies, concerns, associations;

depending on the frequency of presentation:

periodic – monthly, quarterly covering the indicators of the current activity of subjects;

annual – summarizing the main results of financial and production activities of economic entities over a year;

according to the urgency of delivery:

postal – transmitted by mail;

electronic – transmitted by email.

Specially organized statistical observations are organized and conducted to obtain information that is not available in the accounts or to verify the data. For example: an all-Ukrainian population census requires specially trained people, a specially designed program; it is time consuming and expensive.

Specially organized observations can be conducted in the form of censuses, records, special surveys.

A census is a continuous or selective observation of mass phenomena designed to determine their size and composition on a specific date. Censuses are carried out periodically (usually at regular intervals) or once. In most countries of the world a census is conducted every 10 years. The peculiarity of the census is that it is carried out at the same time throughout the territory with a single program for all units. Thus, the program of the last census of Ukraine (December 5, 2001) included issues pertaining to an individual (age, gender, nationality, mother tongue, marital status, education, citizenship, sources of livelihood), and households as a whole (household composition, employment of its members, living conditions, etc.).

Recording is a continuous observation of mass phenomena based on the survey data, interviews, and records. For example, recording of the land fund according to the type of land, quality of soil, categories of state gifts.

Special surveys are incomplete observations of particular mass phenomena in accordance with a specific topic that goes beyond reporting. Such surveys may be periodic or one-time. For example: surveys on informal employment, budgetary household surveys, marketing surveys.

Interviews are often incomplete observations of opinions, evaluations, motives of a certain part of respondents, which are recorded according to their words. A referendum, which is a complete interrogation of the whole population on various socio-economic or political issues, is quite an exception.

Register is a form of continuous statistical observation of long-term processes. A register is a system that constantly monitors the status of the observation unit and evaluates the intensity of the impact of various factors on the indicators being studied. In a register, each observation unit is characterized by a set of indicators: those that remain unchanged during the observation time and are recorded once; those indicators whose frequency of change is unknown and which are updated in case of change; the third ones that are dynamic series of indicators with a known upgrade period.

The most famous registers include: *population registers* as well as *registers of enterprises and organizations*.

A *population register* is a named and regularly updated list of the country residents. Observations are conducted according to the following features: gender, date and place of birth (constant data), marital status (variable feature). Information about each person born or those who came from abroad is recorded in the register. In the event of death or departure to a permanent residence in another country, the registry data are removed.

The *register of enterprises and organizations* is a list of subjects of all economic activity types, indicating their details and main indicators. Ukraine has adopted the Unified State Register of Enterprises and Organizations (EDRPOU). This Register makes it possible to establish a common information space which includes market players.

The EDRPOU Information Fund contains:

the subject's registration code;

information about the branch, territorial identity of the subject, its subordination;

types of property;

organizational forms of management;

background information (executives' names, addresses, telephone numbers, faxes, founders, etc.);

economic indicators.

The types of statistical observation differ in:

the nature of time logging:

current – when changes are recorded as they are received (birth, marriage registration). Such observation is conducted to study the dynamics of any phenomenon;

periodic – observations made on a similar program at regular intervals (population census, registration of prices for certain types of goods, a record of success);

one-time – information about the phenomenon being studied received once at the time of observation as the need arises (marketing research of the local goods market);

extent of coverage:

complete observation covers all units of the population (reporting, population census);

incomplete observation assumes that only part of the units of the population under investigation is subject to inspection.

Incomplete observation, in turn, is divided into:

selective – random selection of the population units to be investigated (small and medium-sized business surveys);

main array method – a survey of the most significant, largest units of the population, which are crucial for the characteristics of the observation object (financial rating of the most influential banks);

monographic – individual population units, usually representatives of any new types of phenomena, subjected to a detailed survey for the purpose of a thorough study (IMF activity survey);

questionnaire – collecting information through questionnaires (a survey of a campaign participants);

monitoring – the process of systematic or continuous collection of information about the parameters of a complex object (activity) in order to determine the trends of change in it (weather monitoring, stock exchange prices).

According to the **methods of obtaining information** there are the following types of observations:

direct observation – information is collected by specially trained people, whose task is to obtain information through a personal record of population units (counting, weighing, measuring the value of a feature). This is the most accurate and reliable, but also the most time consuming and costly way to get data;

documentary observation (recording) is used by enterprises, organizations, institutions to prepare reports on the basis of primary accounting documents; it is a fairly reliable source of information;

a survey is a method of observation in which information is obtained from the accounts of the respondent (interviewee). A survey is used to generate information about processes and phenomena that are directly observable. In statistical practice, the following main survey methods are used:

an oral (expeditionary) survey – specially prepared registrars fill out census sheets (record facts) based on the survey of the person being surveyed. The registrar simultaneously controls the accuracy of the information received. An oral survey provides fairly accurate results, but it is a costly way of obtaining data;

the way of self-registration is the process when respondents fill in the forms themselves; registrars distribute questionnaires, instruct respondents, collect the completed forms, controlling the correctness of completed information;

the *correspondence method* is the process when interviewed people fill in the forms and send them to the research organization without the participation of the registrars based on the instructions of how to complete them. This method of interviewing requires the least cost but does not provide high quality of the received data as it is not always possible to check the accuracy of the received information;

the questionnaire method involves collecting information in the form of questionnaires. Questionnaires are distributed in various ways and are filled out on a voluntary basis, usually anonymously. The number of questionnaires received by the survey organization is always much smaller than the number of questionnaires sent. Such a survey provides approximate, indicative data that is commonly used to study public opinion on various issues.

2.3. The plan of statistical observation

Any statistical observation is conducted according to a predesigned plan. A **statistical observation plan** consists of two groups of questions: programmatic-and-methodological and organizational ones [6; 17; 34; 46].

The development of programmatic-and-methodological issues of the observation plan consists in the scientific and practical substantiation and highlighting of the phenomenon essence, the conditions of its formation and manifestation. In addition, a system of features characterizing the phenomenon is selected, a possibility of quantitative processing and checking the features for accuracy is taken into account.

A set of programmatic and methodological issues can be submitted in the order of their emergence and resolution (Fig. 2.1).

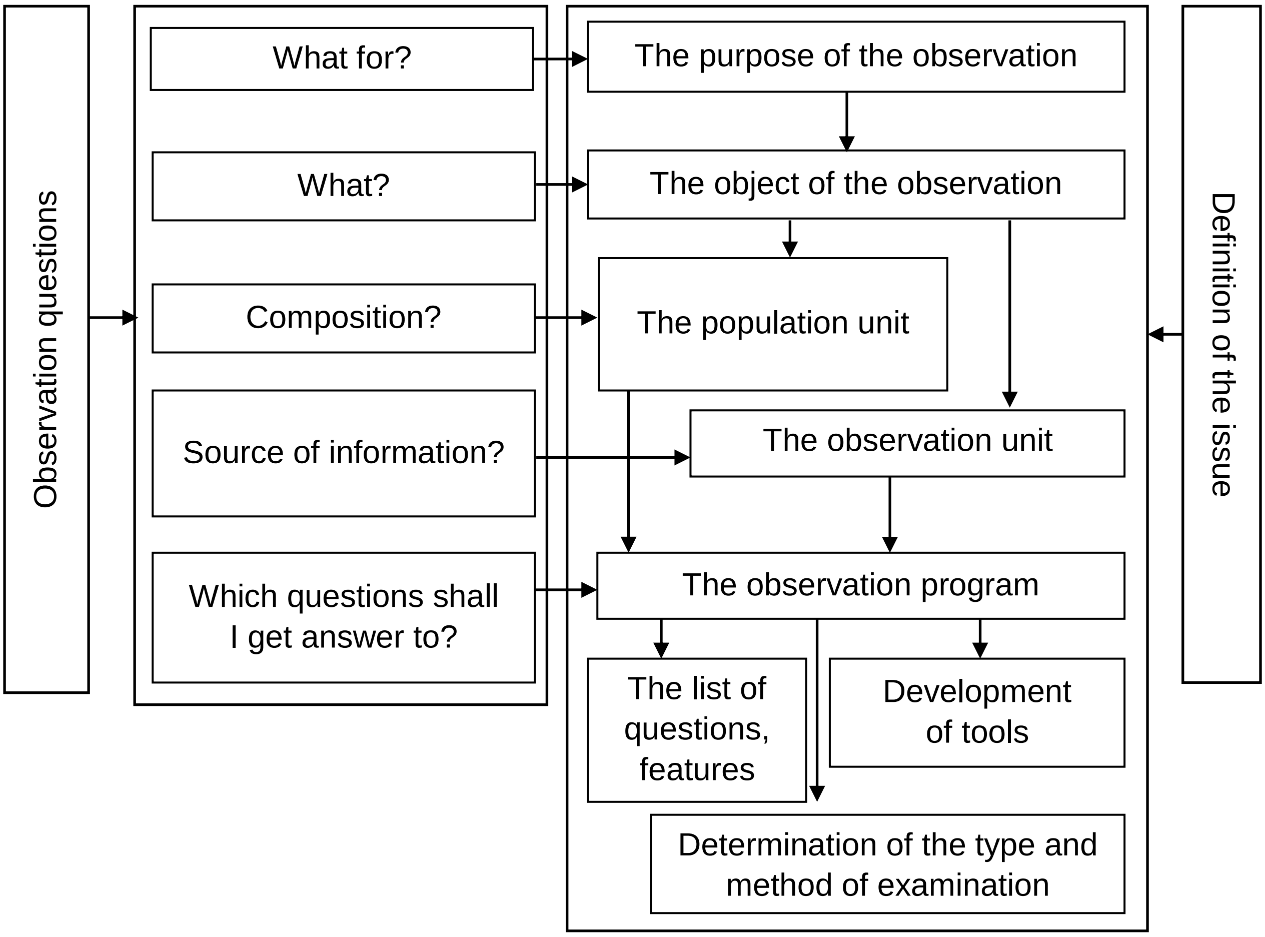

Observation questions
What for?
What?
Composition?
Source of information?
Which questions shall I get answer to?
The purpose of the observation
The object of the observation
The population unit
The observation unit
The observation program
The list of questions, features
Development of tools
Determination of the type and method of examination
Definition of the issue

Fig. 2.1. **The programmatic and methodological issues of the plan of statistical observation**

Programmatic-and-methodological issues of the statistical observation plan **include** [34; 64; 65; 77]:

determining the purpose and objectives of the observation;

choosing the object, observation units and units of the population;

developing an observation program;

choosing the form, type and method of observation.

Each statistical observation is conducted in order to obtain reliable data on the processes and phenomena under study. The *purpose of a statistical* *observation* should be specific and clearly formulated, based on the overall objectives of the statistical survey. According to the principles of the systematic approach, the tasks of the observation must be consistent with the stated goal, and proceed from it.

The purpose and objectives determine the ways in which the program is developed as well as the choice of observation forms, while the specificity of the formulation of the purpose and objectives ensures the completeness of data collection.

Depending on the purpose and objectives the *object of observation* that is a statistical set of socio-economic phenomena or processes to be surveyed is determined. Establishing an object of observation means determining the exact boundaries and composition of the population. For example, during a census, it is necessary to determine which population is to be registered: current (available at the time of the census in the surveyed area) or permanent (population permanently living in the territory, but possibly absent at the time of the census).

Sometimes the concept of qualification is used to limit the object of observation. *Qualification* is a series of features with their quantitative value in conducting a statistical observation being the basis for inclusion (or non-inclusion) of a unit into the studied population. For example, if the object of observation is the work of small businesses, then under the current law, they include businesses with no more than 50 employees.

If the object of observation is not clearly defined, the observation may cover phenomena that do not belong to the population under consideration, and vice versa, some units will fall out of the survey.

Any object of observation as a statistical population consists of units of the population. A *population unit* is an element of a statistical population that bears the features to be registered. For example, in demographic surveys, both a person and a family can be a unit of the population, while in budget surveys it can be a household or a family. The number of population units is called the **size of a population**. For example, in a population census, the *statistical population* is the country' whole population, while the *population unit* is a separate person; in trade, a *statistical population* is made up of trading enterprises, and the *population unit* is a separate trading enterprise; if water quality is to be investigated, all water sources are the *population*, and the *population unit* is a separate water source.

Units of a population have many different features. A *statistical feature* is a specific property, quality, a distinctive feature of a population unit. For example, an employee of an enterprise has age, gender, education, family status, etc. Researchers are interested in certain features, so the observation program provides a list or formulation of features that require data to be recorded.

Clarification and formulation of features of a population unit is carried out according to the following *rules*:

features are selected according to the objectives of the study, the possibility of processing and analysis of the data obtained;

the selected features should not be numerous;

the features must be combined to complement each other;

the selected features should take into account the capabilities of the researcher.

However, not every unit of a population can provide information about itself. Therefore, a *unit of observation (*a *reporting unit)* is used during a survey*.* An *observation unit* is a primary unit from which information is obtained. Thus, at the time of a census, a household as well as each of its members is the unit of observation. When it comes to registering sold apartments on the real estate exchanges, the unit of observation is the exchange. Therefore, *the unit of a population and the unit of observation may or may not coincide.*

When observation involves reporting, the *reporting unit* (observation unit) may coincide with the unit of a population. In budget surveys of households, such units are households; in the course of studying the public opinion of the population each respondent, i.e. the person who expresses his opinion is a unit. However, they may not match. Thus, in the study of solvent demand, the population unit is the buyer; when a researcher asks him a question, the buyer acts as a reporting unit, but as far as the sales information is concerned it can be obtained from the sales manager or cashier.

Proceeding from the set goals, defined tasks and the selected object, an observation program is developed.

An *observation program* is a list of features to be recorded (in direct observation) or a list of questions used to collect information (in surveys). The development of an observation program is a complex and responsible process that ensures the quality of the information collected. The composition and content of the observation program are determined by the objectives of the study and the characteristics of the socio-economic phenomenon under study. Any phenomenon has many features. It is inappropriate and impossible to collect information about everyone. Therefore, it is necessary to select the most important features that meet the objectives and the purpose of the observation.

An observation program must meet the following *requirements* [34, 52, 65]:

inclusion of only essential features that directly characterize the phenomenon under study, its type, main features and properties (minor issues that are not related to the task, should not be included in the program);

omission of questions that may have deliberately inaccurate answers;

inclusion of control questions to verify the information being collected (logically related questions about age, marital status, education, having children);

orientation of all questions of the program to a specific form of answer: *digital, alternative, multivariate.* With a *digital* question the answer is given in quantitative form (about age, seniority, earnings); an *alternative* question requires a yes/no answer (gender, marital status); a *multiple choice* question provides a choice of one or more options from the proposed menu (the question about the marital status may include the following options: married; never married; widower; divorced).

numerical coding of the response to facilitate processing;

determining the composition and sequence of questions (the logic of the location of the questions contributes to obtaining more reliable data: the year of birth and the number of complete years of the interviewee age).

At the same time, the monitoring program develops *tools* in the form of *statistical forms* and *instructions* for completing them.

The statistical toolkit provides not only the input but also the output part of the observation information base. That is, when defining features, writing blocks of questions, they simultaneously prepare layouts of source tables where no digital information takes place. The layout of the tables can help to determine how each question (feature) is consistent with others, to predict how one or another question (feature) works, to leave out the informative ones, to justify the method of further statistical processing of other questions.

The *statistical form* is a single sample document containing the program and the results of the observation. It can have different names: a survey form, a census, a questionnaire, a report.

The required *elements of the statistical form* are the *title* and *address* parts. The *title section* shall indicate: the name of the statistical survey and the body conducting it, the form number, as well as by whom and when it was approved. The *address* is the address of the reporting unit, its subordination.

The *instruction* provides the procedure for observing or completing the form. Depending on the complexity of the observation program, this may be a separate brochure or a response card, or an explanation on the back of the form.

The *observation program* also provides a definition of the type and method of data logging. Typically, the type and method of observation depend on its purpose, the nature of the object of observation, the extent and degree of accuracy of the expected results.

The general scheme of organizational issues of a statistical observation plan is presented in Fig. 2.2.

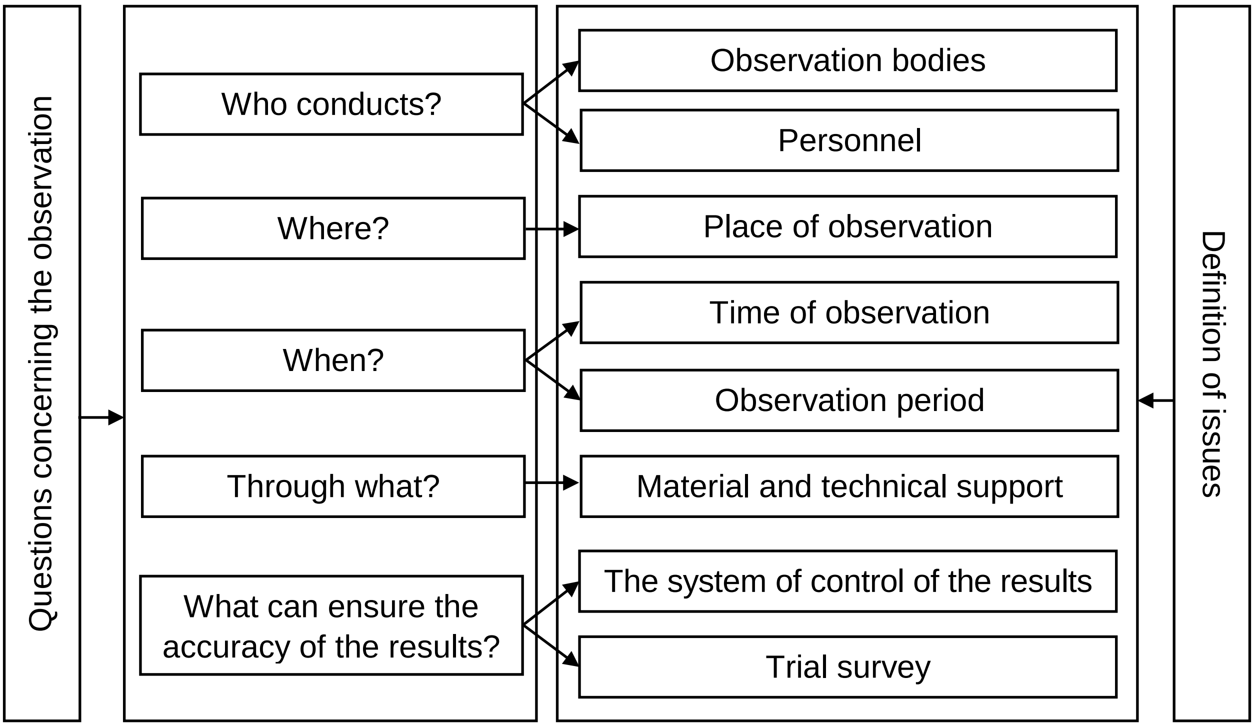


Fig. 2.2. **Organizational issues of a statistical observation plan**

The second part of the statistical observation plan is a set of **organizational issues** which include [33; 34; 44; 75]:

determining the body (executor) of the observation – the subject of the observation (the observation can be carried out by their own forces or by organizations specialized in conducting the observations);

determining the *observation time*: the start and the end date of the observation, the critical date. The *term (period)* of the observation is set based on the volume of work and the number of staff involved in data collection. The *critical date (moment)* is considered to be a specific day of the year (time of day), as of which the features of each unit of a statistical population are recorded;

determining the *location (territory)* of the observation. For example, if the cost of a consumer basket in Kharkiv is to be determined, the territory of the city will be the place of the observation. The choice of the observation location is determined by its purpose.

Depending on the scale of the object of observation and the interest in its results, the observation *subjects* may be:

central bodies of state statistics: the State Statistics Service of Ukraine and its regional offices that perform state surveys at the macro level (population census, land fund; survey of migration flows of population, informal employment, household budgets, activities of business entities);

statistical departments of ministries and departments carrying out state surveys of a local nature on a specific topic (surveys of the State Fiscal Service, the State Border Service, etc.);

specialized institutes, agencies, international organizations that conduct surveys based on the study of: public opinion or motivation, behavior and evaluations of certain subjects of socio-economic life (International Institute of Sociology, International Labor Organization – ILO, General Agency for Economics and Finance of the EU Committee, etc.);

analytical departments of separate economic structures (enterprises, organizations, firms, banks, exchanges, insurance companies, etc.), conducting micro-level surveys of marketing or control nature (marketing surveys of tourism firms which aim to determine the volume of the consumer market; surveys of actual and potential clients of commercial banks aiming to analyze the need for different types of banking services).

The rationale for the *observation location* is determining the place where the observation unit is located and the data are recorded. For example, during a telephone survey of viewers, the respondents' place of observation is the place of residence. In the case of a survey of people arriving from abroad (the purpose of their arrival being the subject of observation), the place of the survey is the district office of visas and registrations (VVIR).

For the observation to provide plausible and timely data, the time and period of observation are of great importance.

Observation time (objective time) is the time to which the observation data belong. When the object of observation is a *process*, the time interval is selected over which the data is accumulated. If the object of observation is a *certain state*, they choose a *critical moment* – a moment of time, as of which the data is recorded. For example: the number of transactions concluded on the universal exchange can be registered both over a certain period (during a month) and at certain times – the days of exchange trades, which usually take place once a week. In this case, the transactions at a certain time are recorded as a primary observation, while over a period – as a secondary one.

The critical moment is used when conducting a census (when it is necessary to accurately register the status of the population at a certain "moment"). Thus, the time of the last all-Ukrainian census was 24:00 of February 4 to 5, 2001. It is clear that a census of the entire population of a country can't be taken in an instant, so in addition to the observation time, an *observation period (subjective time)* is established during which the data are recorded. If the census period is for example 10 days (from February 5 to February 14 inclusive), for example, on February 10, all data on the members of a particular household are recorded as of the beginning of the census (February 5). If a baby is born in a household in the period of registration, it will not be included in the census sheet because at the critical moment there was no baby to be recorded.

The observation time is chosen in the best or neutral period for the object of observation. For example: the census of fruit and berry plantings is carried out during their flowering and availability of fruit. The population census is conducted in the period of the lowest migration activity (for Ukraine it is winter months).

For any statistical survey, appropriate *material and technical support* is required: print media, computing and reproduction equipment, vehicles, statistical tools and advertising media. The latter is important when preparing macro-observations that require clarification and communication for a wide range of citizens.

At the same time, a methodological (as a means of preventing errors) and organizational (as a means of detecting and correcting errors) issue of the observation plan is the *control of the observation data*.

Another element of the organizational objectives of the statistical observation plan is covering its results in the media.

2.4. Statistical observation errors

During statistical observation, errors may occur, i.e. discrepancies between the actual situation and the results obtained. Errors are corrected through control.

Control means checking the survey data for completeness and accuracy. The completeness of the data is usually controlled visually: they are checked for all units and positions. The accuracy of the data is checked by means of logical and arithmetic control [34; 46; 52].

Logical control is a test for the data compatibility, which means comparing interdependent characteristics. For example, control of the respondents' answers received during a survey includes: comparison of the age of the respondent with his marital status, type of activity with the source of livelihood, financial position of the economy with the structure of budget expenditures. Logical control determines only the presence of an error, but not its size. But sometimes it is possible to determine the approximate limits of an error by comparing observational data with similar data of other observations or data in dynamics (an error in the stock price quotation data is determined by knowing the price rate and its fluctuation limits; an error in the credit interest rates is found based on the NBU discount and lending rates statements).

The size of the error can be determined and corrected by means of *arithmetic control*: direct or indirect conversion of the registered data (the amount of total household income can be checked against all its expenses as to major items; the size of the share of the company capital can be controlled through the availability of information on the number of shareholders and the size of the average share). If the errors go beyond logical and arithmetic control, it is impossible to establish the accuracy of the data and it is advisable to take into account the nature of the errors.

The classification of statistical observation errors is performed based on the following features:

- *the causes of* errors:

registration errors may occur in any observation due to the distortion or incorrect recording of facts;

representativity errors arise during sampling because of incompleteness of data registration and violation of the principles of random selection;

- *the nature of errors*:

accidental errors which occur as a result of a coincidence of random circumstances (due to the registrar's negligence or the respondent's lack of attention). They distort the observation data, but due to the mass of cases, the effects of such errors are counterbalanced, and they do not significantly affect the result;

systematic errors result from constant distortions in one direction (the tendency to round off the amounts of revenues and expenses to integers: revenues in the direction of decrease and expenditures in the direction of increase). Such errors significantly shift the observation results to one direction (increase or decrease).

Systematic errors, in turn, are subdivided into:

unintentional errors which arise from the unreasonableness of the observation program, the incompetence of the registrars. For example: in the observation of the impact of the Chornobyl accident on the health of the population, the results significantly increased the incidence rate, as all the population in the medical institutions was surveyed; a classic example is female flirting (decreasing the age), aging flirting (increasing the age) or rounding the age to numbers multiple of 5 or 10 – a so-called age accumulation;

malicious errors that arise from deliberate misrepresentation of facts in order to obscure or adorn reality. For example, mass facts of concealment of tax revenues; notes in the reports; concealment of proceeds from commercial sales network, catering and so on.

Significant accuracy of statistics is conditioned by an effective system of measures aiming to reduce or avoid errors. They include: high-quality primary accounting, development of scientific and methodological recommendations on data validation, selection of qualified statisticians and analysts, computer processing of information and automation of statistical work.

Important concepts

Object of observation is a statistical set of socio-economic phenomena or processes that are subject to observation.

Unit of a population is an element of a statistical population that is the bearer of the characteristics to be registered.

Observation unit is the primary unit from which information is obtained.

Statistical observation plan is a set of programmatic, methodological and organizational issues.

Observation error is the discrepancy between the real situation and the results of the observation.

Observation program is a list of features to be registered (in direct observation) or a list of issues on which information is collected (in surveys).

Registry is a form of continuous statistical observation of long-term processes.

Specially organized observation is an observation that is organized and conducted to obtain information that is not present in the report or to verify the information.

Statistical reporting is the main form of statistical observation through which statistical authorities obtain the required data in the form of statutory reporting documents in due time.

Statistical observation is systematic, scientifically organized collection of data or information about mass phenomena and processes, which consists in recording the selected characteristics for each unit of a population.

Subjective time is the time during which data is recorded.

Observation time (objective time) is the time to which observation data belongs.

Typical tasks

Task 1. A draft statistical observation program is to be prepared to determine the trend in the changes in the number of thefts in the region during 2010 – 2018 and to draft a statistical form.

The solution.

According to the purpose of the observation, it is necessary to draw up table forms which will contain the collected information (Table 2.2).

Table 2.2

The program of statistical observation of dynamics of the number of thefts registered in the region in the period 2010 – 2018

The number of thefts	2010	…	2018	…	Changes	
	event, units	deviation,%	event, units	deviation,%	event, units	deviation,%
Auto						
Housing						
Office						
In public transport						
On rail way						
In urban markets						
Robbery with theft						

Analysis of dynamics can be performed according to the results of the observation (Table 2.3).

Table 2.3

A draft statistical form of dynamics of the number of thefts registered in the region in the period 2010 – 2018

Years	Number of thefts, units	Absolute increase, units		Relative increase, %	
		over the years	compared to 2010	over the years	compared to 2010
2010					
2011					
…					
2018					

Task 2. The following application form is submitted for obtaining the position of workshop foreman at an enterprise:

Filing date: May 25, 2018.

First name: Petro Ivanov.

Birthday: February 30, 1976.

Marital status: married.

Children: 2,

including minors: no.

Education: higher technical.

Work experience: 32 years.

Place of previous work: the galvanic shop area foreman.

Perform logical and arithmetic control of the given data.

The solution.

Logical control is performed for the content and relevance of each piece of the information received.

The date of birth fails to pass logical control:

February 30, 1976 – there may not be more than 29 days in February, so there is an error in the date or month of birth.

Arithmetic control is performed by calculations of the related data.

The work experience record fails arithmetic control.

At the date of submission, Ivanov was 42 years of age. The mentioned experience is 32 years, so it turns out that officially P. Ivanov started working at the age of 10 (42 − 32 = 10), which is contrary to the current legislation.

Task 3. Determine the objective and subjective observation time:

a) bank accounts payable at the beginning of the quarter should be submitted within ten days from the beginning of the next quarter;

b) the deadline for submission of annual export reports about export-oriented enterprises' products is no later than January 15 of the following year.

The solution.

Based on the definition of the objective and subjective time of observation (which means, respectively, the date of observation and the time during which the data is recorded) we have:

a) the objective time is the quarter, since information on payables should be submitted at the beginning of the quarter, that is, for the previous quarter; the subjective time is 10 days, since during this time the information must be processed and submitted to the relevant authorities;

b) the objective time is a year, since information on export volumes of products must be submitted for a period of one year (annual report); the subjective time is 15 days, as information must be submitted within 15 days.

Task 4. Define the organizational form and method of observation:

a) a list of all registered economic structures, indicating their details, type and field of activity;

b) a survey of customer feedback at the service stations.

The solution.

a) the organizational form of observation is the register, because it is a list of units of a particular observation object. In this case it is a set of registered economic structures, indicating the necessary features (requisites), which are constantly updated, such as the name, the address, the telephone, etc. The observation method is documentary recording, since registration or updating of the features of economic structures occurs according to the data given in certain documents;

b) the organizational form is specially organized observation, since such form of observation covers the spheres of life that are not presented in reporting. There are no reporting forms available for customer service reviews, so this is a specially organized survey. The observation method is a survey that is an incomplete observation of opinions, assessments of the service quality at service stations, which are recorded according to the customers of these stations.

Reference to laboratory work

The guidelines for performing laboratory work on the topic "An overview of Excel capabilities" are presented in [100]. The laboratory work is designed to master and consolidate the skills in the use of the basic methods of work with spreadsheets: data entry and editing, formatting, designing tables.

Questions for self-assessment

1. What is statistical observation, what is its essence and main tasks?

2. What are the main programmatic, methodological and organizational issues of statistical observation?

3. What is an entity, an observation unit and a population unit?

4. What is the critical moment and the time of observation?

5. What are the requirements to be met when designing a statistical observation program?

6. Identify the object, the unit and the purpose of statistical observation and develop a program of: a) a survey of the activities of export-oriented enterprises; b) analysis of the enterprise's advertising costs; c) analysis of the use of labor and working time at an enterprise.

7. What are the main forms of statistical observation?

8. Describe the types of statistical observation according to:

a) the extent of coverage of statistical units; b) the time of data registration.

9. What are statistical observation errors and what are their types?

10. How is the monitoring of results carried out? What are the methods of correcting statistical observation errors?

Questions for critical rethinking (essays)

1. Make a list of questions included in the observation program:

a) interviewing graduate students regarding the practical orientation of the disciplines taught;

b) a sample survey of the budget of young families in the region;

c) recording of bank credit operations.

2. How does the creation of automated databases affect the development of an observation program?

3. Develop a program of statistical observation and control of the results obtained in the survey of production volumes and factors of production of industrial enterprises in the region.

4. List the advantages and disadvantages of the forms of statistical observation. Justify the answer.

5. Describe any three different statistical surveys in terms of appropriate types and methods.

3. Presentation of statistical data: tables, graphs, charts and maps

Basic questions:

3.1. The role and value of the graphical method.

3.2. The main elements of graphs. The rules of plotting statistical graphs.

3.3. The types of statistical graphs and tables. The ways to build them.

3.1. The role and value of the graphical method

Visualization of information emerged as a result of studies of human-computer interaction, computer science, graphics, design, psychology, and economics. It is increasingly being used as the most important component in research, digital libraries, information analytics, and so on. Visualization of information is designed to create new and clearer approaches to data transmission in intuitive ways.

Among the fundamental approaches to data analysis *visualization of statistical data* takes a special place in visual data analysis, which relies on the cognitive skills of analysts. *The graphical format* of statistics makes it easy to understand a vast amount of information and relationships between different data sets. That is why an analyst must use different methods and tools to interpret any socio-economic process by visualization of statistics.

Graphs are important means for expressing and analyzing statistics because visual representation facilitates the perception of information. Graphs allow you to cover and instantly comprehend a set of indicators – identify their most typical relationships, trends, characterize the structure, the degree of the plan implementation, evaluate the geographical location of objects. This explains the use of graphs for the promotion of statistical information which characterizes the development results of various spheres of the national economy and social relations.

A statistical graph is a drawing in which statistical populations are characterized by certain indices described by conventional geometric images or features [10; 21; 28; 58]. Presentation of tabular data in the form of a graph makes a stronger impression as compared to the digital form, allows you to better understand the results of statistical observation, interpret them correctly, facilitates the understanding of statistical material, makes it visual and accessible. However, graphs are not only illustrative. They provide new knowledge about the subject matter of the study, as a method of generalization and visualization of the initial information.

Due to the introduction into the analytical work of new economic and mathematical methods, modern computer technology and packages of useful computer graphics programs, it is impossible to imagine scientific research without the use of the graphical method with its wide arsenal of visualization tools. Especially relevant and diverse is the use of the graphical method in statistical surveys aiming to identify complex relationships between the studied indicators, trends and patterns of mass socio-economic phenomena and processes in dynamics and space. Graphs help to visualize the structure of phenomena, study the results of comparison, control the implementation of the plan, investigate the extent of prevalence of certain processes and phenomena in the territory. Moreover, the graphical method is widely used in international comparisons, promotion of best practices, visualization of advanced technologies and scientific achievements and so on.

Therefore, the value of the graphical method in the analysis and generalization of the data is quite significant. Graphic representation makes it possible to check the reliability of statistical indicators because presented on the graph, they more clearly highlight the inaccuracies associated with either the presence of observation errors or the nature of the phenomenon under study. By means of a graphical image it is possible to study the patterns of the phenomenon development, to establish the existing relationships. Simple comparison of data does not always make it possible to grasp the structure of causal dependencies, whereas their graphical representation helps to identify causal relationships, especially in the case of establishing primary hypotheses to be further investigated. Graphs are widely used to study the structure of phenomena, their changes in time and placement in space. They identify more clearly the comparative characteristics and express the main trends in the development and interconnection inherent in the process or phenomenon under study.

3.2. The main elements of graphs. The rules of plotting the statistical graphs

A number of *requirements* must be met when plotting a graph. First of all, a graph should be sufficiently visual, since the whole point of a graphic image as a method of analysis is to provide a clear presentation of statistical indicators. In addition, a graph should be expressive, understandable and comprehensible. To meet these requirements, each graph should include a number of *basic elements*:

1) a graphic image;

2) a graph field;

3) spatial landmarks;

4) scale benchmarks;

5) explication of the graph [16; 34; 41; 75].

Let us take a closer look at each of these elements.

A graphic image (the basis of a graph) is a set of geometric signs, that is points, lines, figures, by which statistical data are depicted [34]. It is important to choose the right graphic image that should fit the purpose of the graph and help maximize the clarity of the statistical data display. Graphs are only images in which the properties of geometric signs – the figure, the line size, the arrangement of parts – are essential for expressing the content of figurative statistic quantities, and every change in the depicted content corresponds to a change in the graphic image.

A graph field is the part of the plane where the graphic images are located. The graph field has certain dimensions that depend on its purpose [34; 41].

Spatial landmarks of a graph are given as a grid system. A coordinate system is required to place geometric signs in the graph field. The most common one is *the rectangular coordinate system*.

In the polar coordinate system, one of the rays (usually the right horizontal one) is taken as the coordinate axis, for which the angle of the rays is determined. The second coordinate is considered to be its distance from the center of the grid, called *the radius*. In radial graphs, the rays denote the moments of time while the circles denote the magnitudes of the phenomenon being studied [16]. On statistical maps, spatial reference points are given by the contour grid (the contours of rivers, the coastline of seas and oceans, the borders of states) and they determine those territories to which statistical values are assigned.

Scale benchmarks of a statistical graph are determined by scale and the system of scales. *The scale of a statistical graph* is a measure of the translation of a numerical value into a graphical one [28; 34].

A scale is a line whose individual points can be read as certain numbers [16; 75]. The scale is very important in the graph. It includes three elements: a line (scale carrier); a certain number of dots marked on the carrier scale in a specific order; digital notation of numbers corresponding to the individual marked dots. As a rule, not all marked points are provided with a digital mark, but only some of them are arranged in a certain order. According to the rules, the numerical value should be placed strictly opposite the corresponding dots, not between them.

The scale carrier can be either a straight or a curved line. Therefore, the following kinds of scales are distinguished: *straight-line* (for example, a millimeter ruler) and *curvilinear* – arcs and circles (a clock face).

Graphic and numeric *intervals are equal or unequal*. If, over the whole scale, equal graphical intervals correspond to equal numerical intervals, such a scale is *uniform*. When unequal graphical intervals correspond to equal numerical intervals, and vice versa, the scale is nonuniform.

A uniform scale caliber is *a cut-off length* (graphical interval), taken as a unit and measured in some units of measurement. The smaller the scale, the denser the points with the same value are placed on the scale [41; 51]. Scale building means placement of dots on a given scale carrier and marking them with corresponding numbers according to the terms of the task. As a rule, the scale is determined by the approximate estimate of the possible length of the scale and its boundaries. For example, on a field of 20 cells, it is necessary to construct a scale from 0 to 850. Since 850 is not divisible by 20, we round the number 850 to the nearest convenient number. Among the nonuniform scales, the logarithmic scale is the most widespread. The method of making this type of scale is somewhat different, since on this scale the segments are proportional to value logarithms rather than values.

The last element of a graph is *explication*. Each graph should have a verbal description of its contents. The description includes the name of the graph, which transmits its contents in a short form; signatures along the scale and explanations of individual parts of the graph.

Various graphs are used for graphical representation of statistical data: their classification is shown in Fig. 3.1.

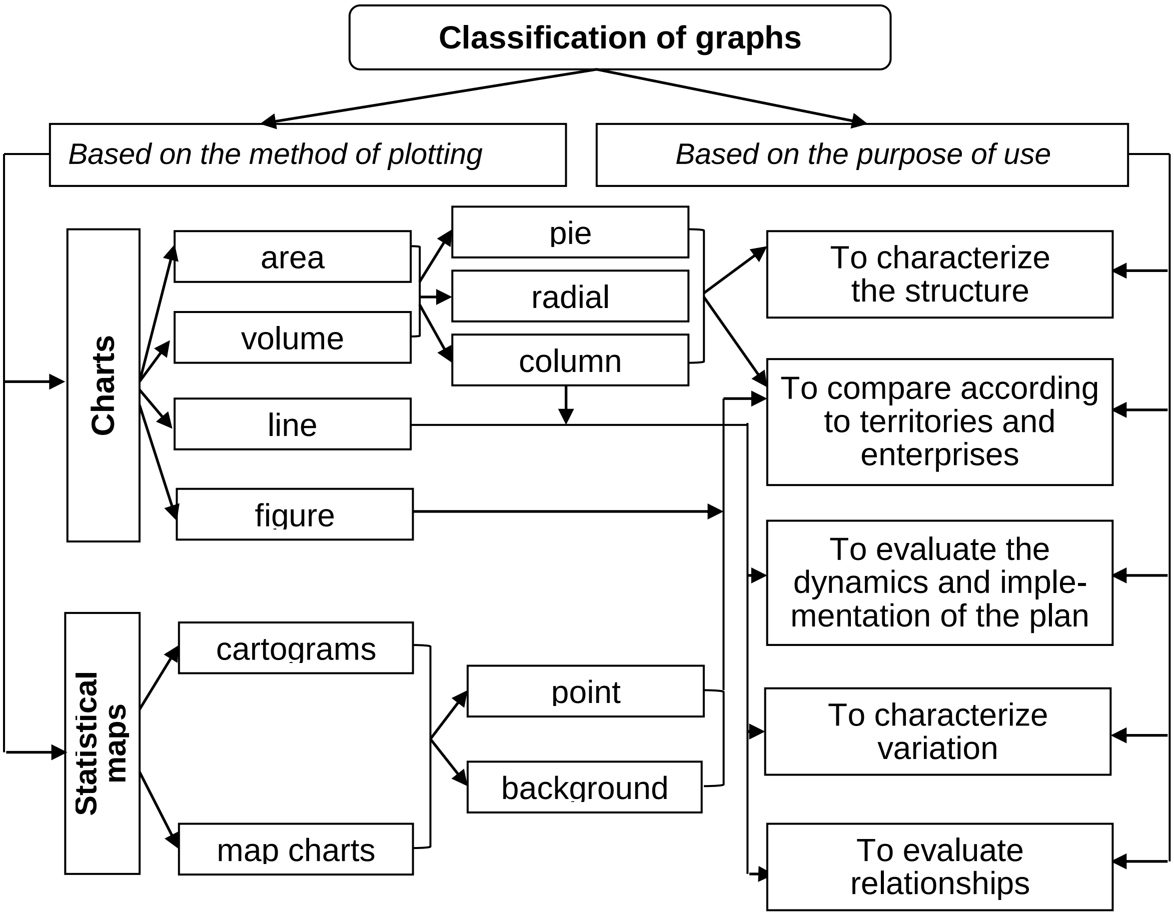

Classification of graphs
Based on the method of plotting
Based on the purpose of use
Charts
area
volume
line
figure
pie
radial
column
Statistical maps
cartograms
map charts
point
background
To characterize the structure
To compare according to territories and enterprises
To evaluate the dynamics and implementation of the plan
To characterize variation
To evaluate relationships

Fig. 3.1. **Classification of graphs** [34; 66; 72]

Today, computer graphics application packages, web applications, and libraries have been developed and are widely used to facilitate the analyst's practical application of graphing. The most common application programme packages are "Harvard graphics", "Statgraf", "STATISTICA", "Excel", "SPSS", "R", "Tableau", "Project BI", "Plotly", "Chart.js", "Raw", "Dygraphs", "FusionCharts" and others [91; 92; 102].

It is advisable to note that classification shown in Fig. 3.1 is not exhaustive, as libraries of modern data visualization software allow you to use a wide range of infographics according to the purpose of the analysis and the available statistics.

Despite the variety of graphic image types, the following general rules are used for plotting of graphs:

an appropriate graph type is selected according to the purpose of use;

there must be a field of the graph, some space in which the geometric signs are placed;

spatial landmarks using scales (uniform or nonuniform) are set;

a coordinate system required to place geometric signs in the graph field is selected [34; 51; 75].

Complying with these rules provides quality visualization.

3.3. The types of statistical graphs and tables. The ways to build them

Let's consider the most common types of graphs.

One way to represent statistical data graphically is a chart. **A chart** is a way to visualize the information given in the form of a table of numbers [8; 62]. The graphical format of the chart makes it easy to understand a vast amount of information and relationships between different data sets. The chart provides an overview of the situation, allowing you to group data and identify important trends.

Line charts are used to characterize dynamics, that is, to estimate changes in phenomena over time, to characterize variation series, to evaluate the implementation of the plan, to evaluate the relationship between phenomena [7; 21; 85]. They are built on a rectangular coordinate system, where the abscissa shows the intervals of time, and the axis of ordinates presents the levels of a time series or the rate of the indicator change. The points obtained are connected by broken lines. Multiple charts can be placed on the same graph to compare the dynamics of different indicators. In variation series, such charts are called *distribution polygons*.

You can use *column charts* for the same purpose. The columns have the same base and their height corresponds to the numeric characteristic of the feature values. The height of the columns determines the relationship between the levels of the studied indicators. In variation series, these graphs are called *histograms*.

Column charts can also be used for straightforward comparisons: comparing across territories, countries, businesses. In addition, they are widely used to study the structure of phenomena. The rules for the construction of column charts allow simultaneous positioning of several indicators on the horizontal axis of images. In this case, the columns have groups, for each of which different dimensions of the varying features can be used.

Varieties of column charts are so-called *ribbon* or *bar charts*. They differ in the location of the scale – horizontally above or below; it determines the length of the strips. The use of the ribbon and bar charts is the same, since the rules for their construction are identical. The unidimensionality of the depicted statistical indicators and the same scale for different columns and strips require the observance of uniformity of positioning: compliance (columns placed according to height, strips according to length) and proportionality to the magnitudes. To fulfill this requirement, it is necessary that the scale on which the size of the column (strip) is set should start from zero; this scale must be continuous, that is, cover all numbers of a certain statistical series; scale breaks and accordingly columns (bars) are not allowed. Failure to comply with the listed above rules leads to a distorted graphic representation of the analyzed statistical material [6; 10; 34].

Directional charts are a variety of bar charts. They differ from the usual two-sided arrangement of columns or strips and have a scale beginning in the middle. Typically, such diagrams are used to represent values of the opposite qualitative value. Comparison of columns (strips) of different directions is less efficient than unidirectional columns. Nevertheless, the analysis of directional charts makes it possible to draw sufficient meaningful conclusions, since the special layout gives a vivid image to the graph. The bilateral group includes charts of numerical deviations. In them, the bands are oriented on either side of the vertical zero line: right – for growth; left – for decrease. Such charts can conveniently represent deviations from a plan or some level taken as the basis of comparison. An important advantage of these charts is the ability to see the magnitude of the variation of the studied statistical feature, which is of great importance for economic analysis.

Charts can be used for simple comparison of independent indicators. They are constructed so that comparable values are presented as correct geometric figures, placed in such a way that their areas are correlated as quantities. In other words, *these charts express the magnitude of a phenomenon by the size of their depicted area*. Various geometric shapes are used to obtain charts of this type: a square, a circle, and rarely a rectangle. It is known that the area of the square is equal to the square of its side, and the area of the circle is determined in proportion to the square of its radius. Therefore, first a square root from the comparable values must be obtained to construct the charts. Then, based on the results obtained, the side of the square or the radius of the circle is determined according to the accepted scale.

For correct construction of charts of a square or a circle, it is necessary to place them at the same distance from each other, and to indicate the numerical value in each figure that depicts it without giving the scale of measurement. This type of diagrams include *a graphic image* obtained by constructing squares, circles, or rectangles one in another with different hatching or coloring. Such diagrams also help to compare a number of studied values. The most expressive and easily comprehensible way is to construct *comparison charts* in the form of figures. In this case, the statistical collections are not represented by geometric figures, but by symbols or signs that in some way reproduce the external image of the statistics [72; 74; 83]. The advantage of this method of graphic representation is a high degree of clarity, obtaining a picture which presents the content of the compared sets.

The most important feature of any chart is the scale. Therefore, in order to construct a figure chart, it is necessary to define *the unit of measurement.* It is taken as a separate figure (symbol), which is conventionally assigned a specific numerical value, and the studied statistical value is represented by a separate number of figures identical in size and sequentially located in the figure chart [19]. However, in most cases it is not possible to depict a statistic with a whole number of figures. The last of them has to be divided into parts, since the scale of one character is a very large unit of measurement. Usually this part is defined "by eye". The difficulty of determining it accurately is a drawback of the figure diagrams. However, if the high accuracy of statistical reporting is not pursued, the results obtained will be quite satisfactory.

Pie charts are used to characterize the structure of socio-economic phenomena. The structure analysis is performed on the basis of comparison of different parts of the whole with the help of the areas formed by sectors of the circle [62].

Radial (radar) charts are used to represent phenomena that change periodically over time (for example, depending on seasonal variations). For construction of this kind of charts they use the polar coordinate system. The circle is divided into 12 equal parts, each denoting a month. On the radius, starting from the center, the scale defines segments that show the month's even features. Their ends connect the segment with each other, resulting in a dodecahedron, which characterizes the seasonality of phenomena.

Table 3.1 shows the classification of chart types according to the available data and the purpose of visualization [21; 25; 62; 85].

Table 3.1

Diagram types according to the purpose and type of data available

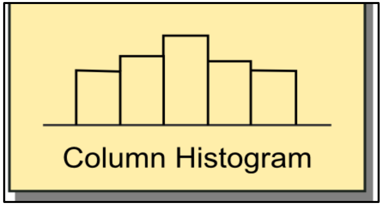
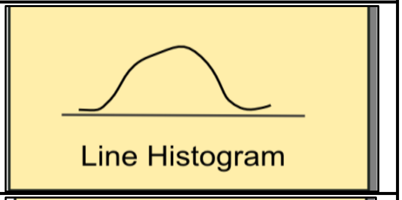
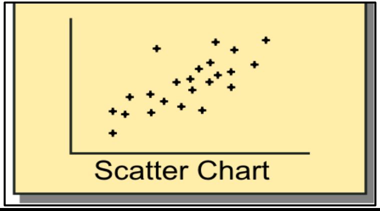
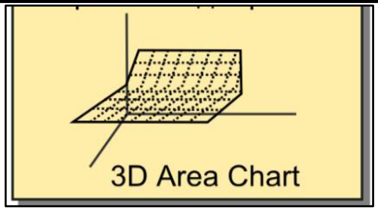
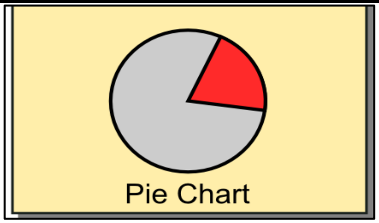
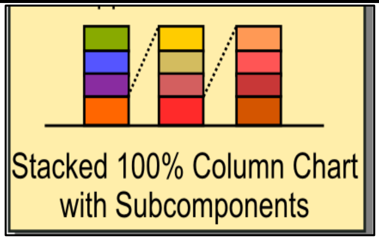
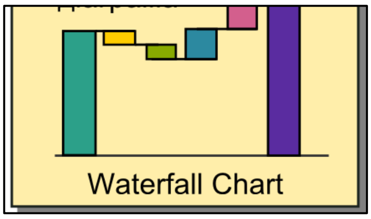
The purpose of statistical data visualization		Characteristics of available data	Recommended chart type
1	2	3	4
Distribution	One variable	Few data points / categories / intervals	
		Many data points	
Distribution. Connection or dependency	Two variables	Does not matter	
Distribution	Three variables	Does not matter	
Composition (structure of the phenomenon or process under study)	Static	A simple part of the whole	
		Parts of components	
Composition (structure of the phenomenon or process under study)	Static	Accumulation of additions and subtractions in general	

Table 3.1 (continuation)

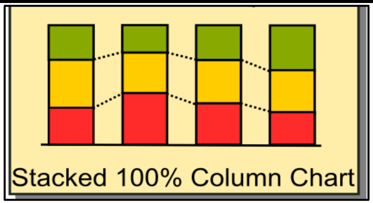
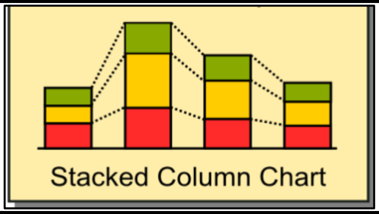
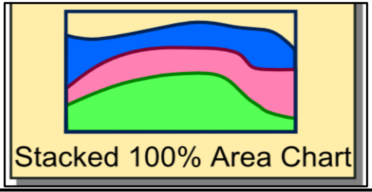
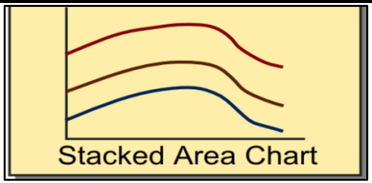
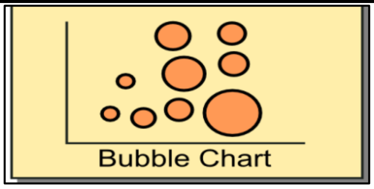
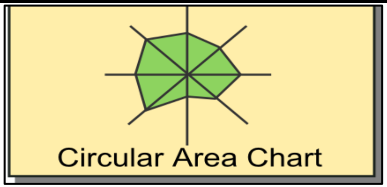
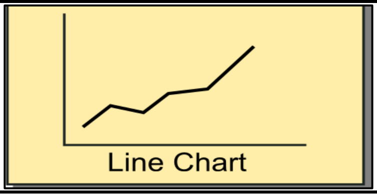
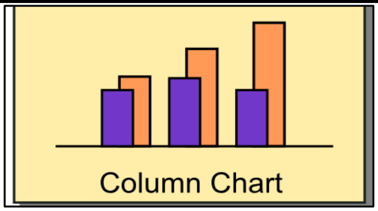
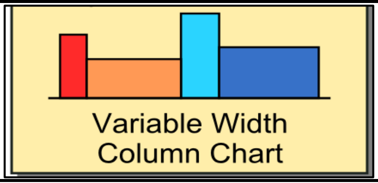
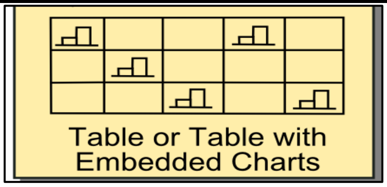
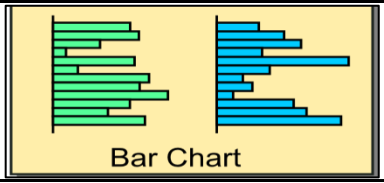
1	2	3	4
The intensity of changes in the phenomenon (process) over time	Not many periods	Only relative differences are important	
		The absolute and relative differences are important	
	Many periods	Only relative differences are important	
		Both absolute and relative differences are important	
Connection or dependency	Three variables	Does not matter	
Comparison	In time	Many periods are cyclical data	
	In time	Many periods are cyclical data	
		Few periods, many categories	
	In time and between objects	Few periods, one or more categories	
		One variable per object, few categories, few objects	
	Between objects	Two variables per object	

Table 3.1 (the end)

1	2	3	4
		One variable per object, many categories	
		One variable per object, few categories, many objects	

Statistical maps are used to evaluate the geographical location of a phenomena and conduct comparative analysis of territories. They are quite often used in UN publications.

Statistical maps are a kind of graphical representation of statistical data on a schematic map; they characterize the extent of the spread of a phenomenon in a particular territory [8; 34; 66; 102]. Territorial placement tools are hatching, background coloring, or geometric shapes. There are *cartograms and map charts*.

A cartogram shows the territorial distribution of the phenomenon under study in different regions. It is used to detect the non-uniformity of this distribution. **A cartogram** is a schematic geographical map showing the relative intensity of any indicator within each unit of the mapped territorial division (for example, population density across districts or regions, distribution of regions based on the GDP volume, etc.). Maps are divided into *background and point types.*

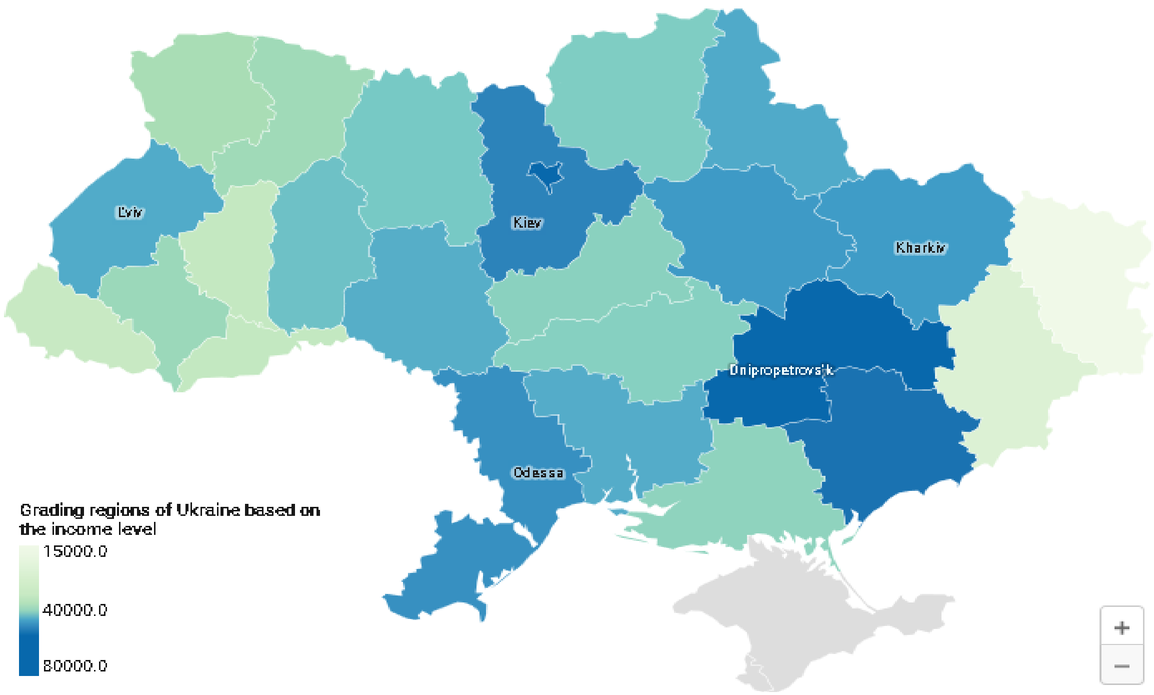
A background cartogram is a type of cartogram on which hatching of different density or color of a certain degree of saturation show the intensity of any indicator within a territorial unit.

A point cartogram is a type of cartogram where the level of the selected phenomenon is represented by dots [28; 34]. A point represents one unit or a number of them, showing the density or frequency of a particular feature on the map.

Background cartograms are usually used to represent average or relative indicators, while *point cartograms* show volume (quantitative) indicators (population, livestock, etc.).

Fig. 3.2 shows an example of a background cartogram*.*

Population income of Ukraine on the regional bases in 2018



Map: Derykhovskaya Viktoriia · Source: State Statistics Service of Ukraine · Created with Datawrapper

Fig. 3.2. **The cartogram of Ukraine's regions based on the income level in 2018** (developed by the author)

Another large group of statistical maps is **map charts**, which is a combination of charts and maps. Diagrammatic figures (columns, squares, circles, shapes, stripes) that are placed on the contour of a geographical map are used as depictive symbols in map charts. Map charts make it possible to display geographically more complex statistical and geographical constructions than on cartograms.

Map charts are divided into simple comparison charts, graphs of spatial displacements, isolines [7; 21; 74]. In a *simple comparison chart* (as opposed to a regular chart), chart figures depicting the values of the studied indicator are not arranged in a row, as in a regular chart, but are spread throughout the map according to the area, region or country they represent. Elements of a simple map chart can be found on a political map, where cities differ in different geometric shapes depending on the population sparseness.

Spatial displacement graphs are used to map various spatial displacements on a map: natural (wind activity, sea currents over a period of time) and socio-economic (population migration, capital movements, freight transport). Using different graphical tools (arrows of different colors, patterns and width), on the graphs of spatial displacements it is possible to represent the direction and speed of displacements, quantity, quality and other characteristics of statistical phenomena and processes that are studied (displaced) in space.

Isolines are lines of equal importance of some magnitude in its distribution on the surface, in particular on a map or graph [34]. Isolines show a continuous change in the magnitude of the studied value depending on the other two variables; they are used in the mapping of natural and socio-economic phenomena. Isolines are used to obtain quantitative characteristics of the studied values and to analyze correlation relationships between them. These types of charts are not exhaustive but they are most commonly used.

The results of aggregations and groupings are recorded in statistical tables, which vividly represent the results of the study. *A statistical table* is a form of rational and visual presentation of statistics on the phenomena studied [10; 34; 51]. The basis of a statistical table is a grid with vertical columns called graphs, and horizontal columns called rows. The rows and graphs which have names make up the layout of the table (Fig. 3.3).

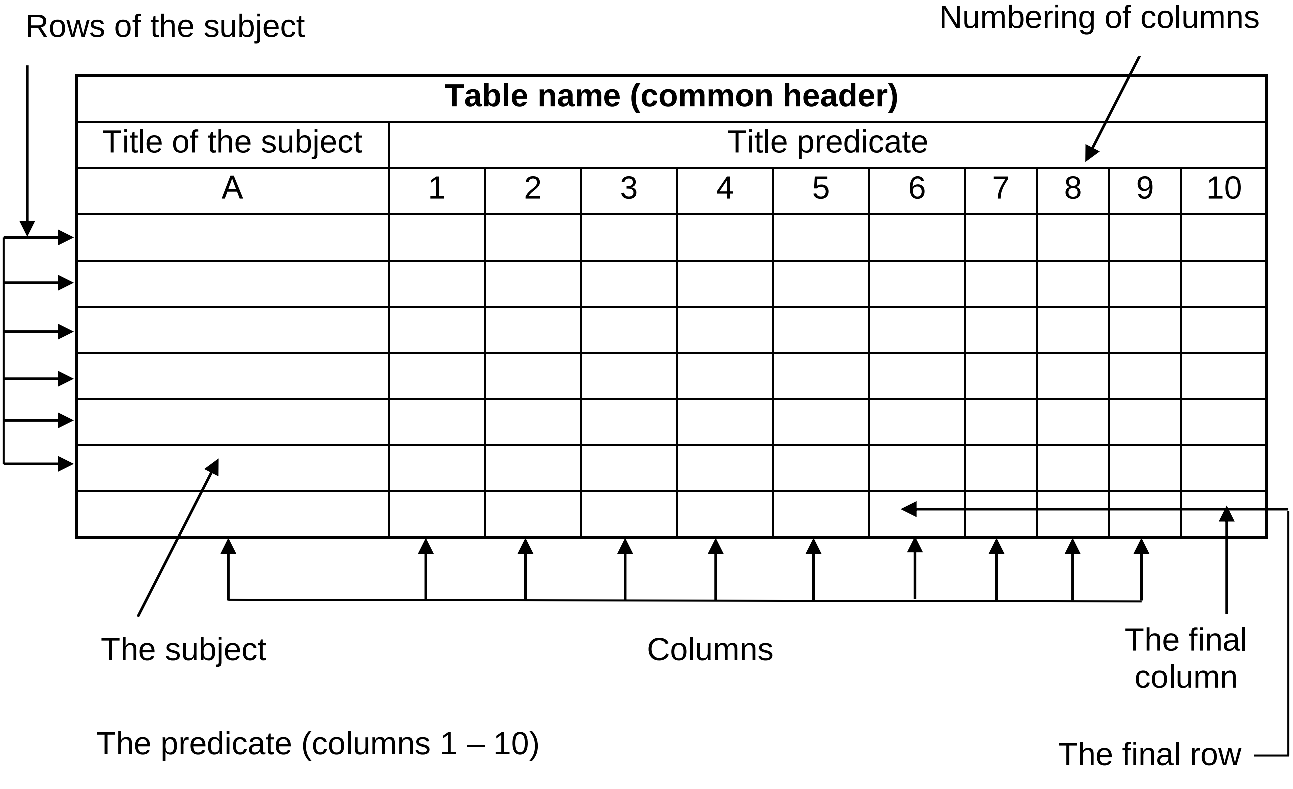


Table name (common header)										
Title of the subject	Title predicate									
A	1	2	3	4	5	6	7	8	9	10

Fig. 3.3. **The layout of a statistical table**

The table contains the title, the subject and the predicate.

The table *header* presents the table content, the place and time with which its data correlate, the units of measurement, if they are common to the data reported.

The subject of the table is a list of units of a group, that is, the object of study; *the predicate* is the digital data that characterize the subject, that is, the results of the summary [34; 41; 72].

Usually, *the subject* is placed on the left in the form of row names, and *the predicate* – on the top in the form of column names.

According to *the content of the subject*, all statistical tables can be divided into the following groups:

simple tables that do not have groupings. They include summarized indicators pertaining to the list: aggregate units (enumeration tables); chronological dates (chronological tables); territories (territorial tables);

group tables in which the object under study is subdivided into groups according to a specific attribute. Each group can be characterized by a number of indicators;

combinational tables in which subgroups are given in accordance with two or more features taken in a combination [34].

The table predicate can be *simple or complex*. A simple predicate implies a parallel arrangement of indicators, (columns 1, 2 in the layout of the table) while a complex one is combined (column 10).

One of the statistical table varieties is *the conjunction table* (interconnectedness table, factorial table), which is widely used as a universal means of studying the statistical relationships between variables. This kind of table shows the main picture of the relationship between two variables and helps to find interaction between them. It is also used in business intelligence, research, and engineering.

A table of interconnectedness is a table that contains a summarized numerical characteristic of a population that is studied according to two or more attributive features or a combination of quantitative and attributive features. *The table of interconnectedness* is a matrix showing the common frequency distribution of two variables [28; 74].

The layout of a table of interconnectedness with i x j dimension, where i = 1,2… k is the number of value variants of one feature (A); j = 1,2… n is the number of variants of values of another feature (B), is given in Table 3.2.

Table 3.2

The general scheme of a table of interconnectedness

	B_1	B_2	…	B_i	Total
A_1	f_{11}	f_{12}	…	f_{1j}	f_{10}
A_2	f_{21}	f_{22}	…	f_{2j}	f_{20}
…	…	…	…	…	…
A_i	f_{i1}	f_{i2}	…		
Total	f_{01}	f_{02}	…	f_{0j}	f_{00}

When building statistical tables, you must follow certain *rules* of design:

1) the table should be compact, easily accessible for viewing. It should not be loaded with unnecessary details that make analysis difficult;

2) the title of the table should express its contents clearly and concisely. The headings of the subject rows and the predicate columns must also be formulated accurately and concisely;

3) it is desirable to give the numbering of columns in the table. This makes it easier to use the table that shows how to calculate numbers in cells. Columns containing the subject are denoted by capital letters of the alphabet, columns that contain the predicate are numbered in Arabic numerals. Clipping is not allowed in the headings of the subject and the predicate;

4) if the units of measurement are different, they are indicated in the names of rows and columns;

5) the features given in the subject and the predicate must be arranged in a logical manner, taking into account the need for their joint consideration. Information is placed from partial to general, that is, at first they show the elements, and at the end they summarize;

6) if the table does not show all the data, but only the most important of them, first they show the summary, and then select the most important parts with the help of expressions "including", "out of them";

7) one should see the difference between "Subtotal" and "Total". "Subtotal" is the result for a certain part of the population, while "Total" is the result for the whole population;

8) the following designations are used in the layout of the table:

dash (−) if the phenomenon is absent;

symbol "×" if the phenomenon has no substantial meaning;

three dots (...) if there is no information (or the entry "no information");

if information is available but its numerical value is less than the one accepted in the accuracy table, it is expressed in the decimal number 0.0;

9) the rounding of the numbers given in the table must be carried out with the same degree of accuracy;

10) if one value exceeds the other one, the obtained relative indicators are expressed in the number of times rather than in percentage [34].

Compliance with the rules of construction and design of statistical tables makes them the main means of submission, processing and synthesis of statistical information.

Important concepts

Graphic image (the basis of the graph) is geometric signs, that is, a set of points, lines, figures, by which statistical indicators are depicted.

Cutting length (graphical interval) is the caliber of a uniform scale, taken as a unit and measured in some appropriate units.

Explication is a verbal description of the graph content.

Table header is a concept that presents the table content, the place and time with which its data correlates, and the units of measurement if they are common to the data provided.

Isolines are lines of equal magnitude in a certain value in its distribution on a surface, on maps or graphs in particular.

Cartogram is a chart showing the territorial distribution of the phenomenon under study in different areas; it is used to identify patterns of this distribution.

Map chart is a chart that combines a chart with a geographical map*.*

Line charts are used to characterize dynamics, that is, to evaluate changes in phenomena over time, to characterize variation series, to evaluate the implementation of a plan, to evaluate the relationship between phenomena.

The scale of a statistical graph is a measure of the conversion of a numerical value into a graphical one.

The scale is a line whose individual points can be interpreted as numbers.

Table subject is a list of units of a group, that is, the object of study; the predicate is the digital data that characterizes the subject, that is, the results of the summary.

Graph field is the part of the plane where the graphic images are located.

Radar chart is a chart used to represent phenomena and change periodically over time*.*

Pie chart is a chart that can be used to characterize the structure of socio-economic phenomena.

Statistical table is a form of rational and visual presentation of statistics on the phenomena under study.

Statistical map is a kind of graphic representations of statistical data on a schematic geographical map, which characterize the level or degree of distribution of a phenomenon in a certain territory.

Bar charts are charts that can be used for spatial comparisons: comparing across territories, countries, firms, and to study the structure of phenomena.

Table of interconnectedness is a table that contains a summary numerical characteristic of a population that is studied based on two or more attributive features or a combination of quantitative and attributive features.

Typical tasks

Task 1. There are data about migration movement of the population of Ukraine in the period 2013 – 2018 (Table 3.3). Using the application Microsoft Excel, visualize the dynamics of changes in the number of people inflow and outflow. Justify the choice of the chart type.

Table 3.3

Migration movement of the population of Ukraine

Years	The inflow, people	The outflow, people
2013	761 842	621 842
2014	622 506	519 914
2015	613 278	519 045
2016	556 808	506 188
2017	442 287	430 290
2018	595 333	561 656

The solution.

Microsoft Excel is a multifunctional program that is used in any field of activity. Various *MS Excel* calculation and analysis tools help to process a large amount of data and visualize the results. In particular, *MS Excel* makes it possible to create many different graphs. Graphs are created based on the data contained in the worksheets of the *MS Excel* workbook. Moreover, all types of graphs in *MS Excel* are dynamic, that is, when you change the source data on which the chart is built, they are automatically updated on the graph that is already built.

Follow these steps to build a standard chart in *MS Excel*:

1) select the output information space, i.e. set the data range (the table with the output data on a *MS Excel* worksheet);

2) launch the *Chart Wizard* or open the *Insert* tab of the toolbar; select the *Diagrams* menu and select the required type of chart (histogram, graph, pie, line, sector, bar, area, radar, bubble, scatter);

3) set the necessary chart parameters (chart title, axes titles, grid, data series, legend);

4) create the desired design (choose the background, the color fill). Modification, formatting and editing of the chart is done using the *Context* menu of the elements, the command *Charts* or the toolbar tab called *Design*.

To visualize migration flows, we choose a line chart (Fig. 3.4), which presents graphically the general trend of the process over time (on a yearly basis) and compare the values according to pre-traced categories.

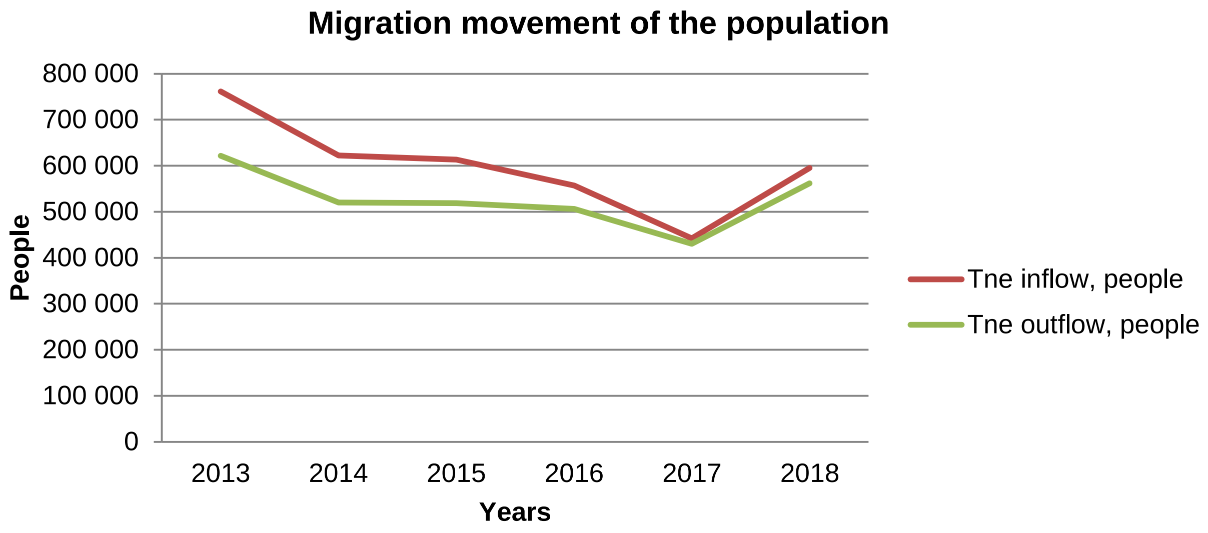


Fig. 3.4. **The dynamics of the population migration flow in the period 2013 – 2018**

Moreover, there are several curves on the same line chart that will allow you to compare the dynamics of different indicators or the same indicator across regions, industries, and more. This type is appropriate if there is a dynamic row.

Task 2. Build a combination chart based on the data in Table 3.4.

Table 3.4

Monthly average expenses per household

Indicators	Years						
	2013	2014	2015	2016	2017	2018	2019
Average monthly expenses per household, UAH	3073.3	3458	3592.1	3820.3	4048.9	4952	5720.4
Food and non-alcoholic beverages	1585.8	1774	1799.6	1914	2101.4	2629.5	2848.8
Alcoholic beverages, tobacco products	104.49	117.57	125.72	133.71	137.66	163.42	165.89
Nonfood products and services	1072.6	1224.1	1336.3	1398.2	1469.8	1807.5	2316.8
Nonconsumptive total expenses	310.4	342.34	330.47	374.39	340.11	351.59	388.99

Visualize the average monthly expenses per household. To visualize the total expenses of one family in the period 2012 – 2018, build a combination chart that will consist of a bar chart and a chart with markers (to visualize the types of spending per household).

The solution.

We use the algorithm described in Task 1 to build a combination chart using MS Excel.

In the first stage, we select the entire array of the output data and choose the type of chart – a bar chart. Then we select each row of data that contains information about a particular type of household expenses and through the context menu we change the type of chart for a series of data to "Chart with markers" (Fig. 3.5). For editing and formatting the design of the created combined chart we use the *Design, Layout, Format* MS Excel tabs.

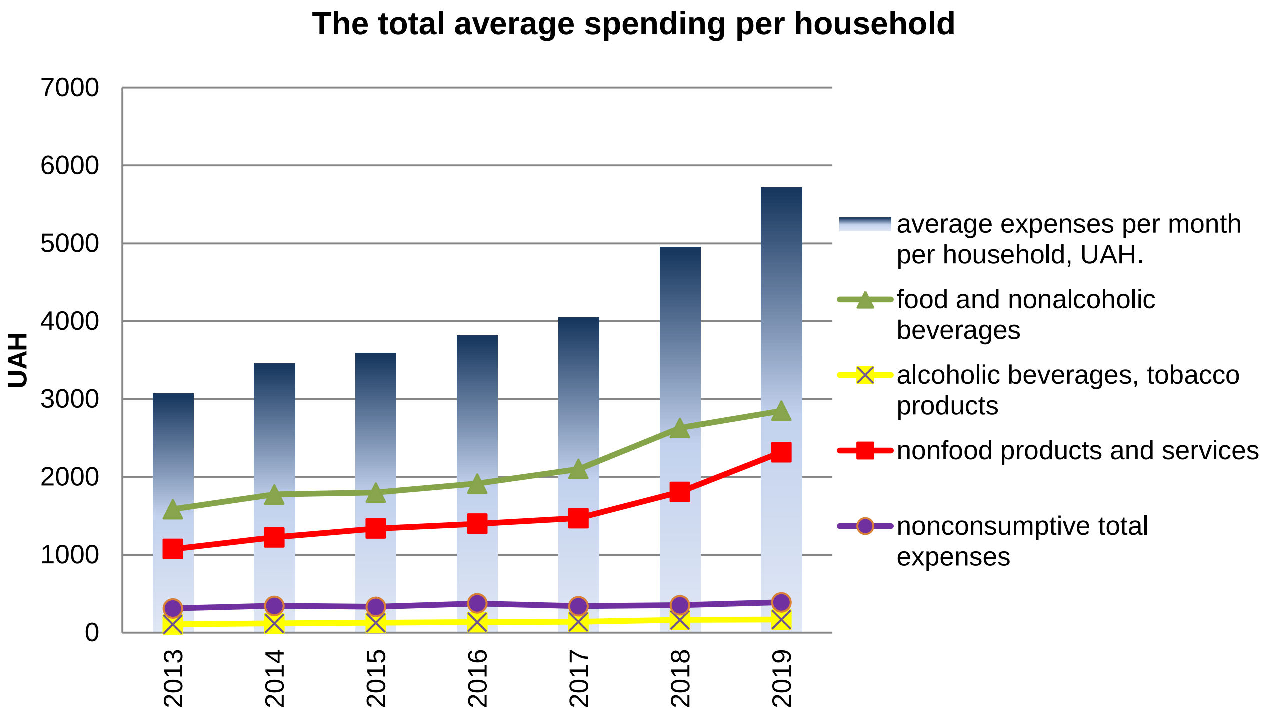


Fig. 3.5. **The combination chart of average monthly expenditures per household in the period 2013 – 2019**

Task 3. There are some data on the income of a certain agro-industrial enterprise specializing in growing and selling of agricultural products (Table 3.5).

Table 3.5

The revenue from agricultural sales in 2019

Types of agricultural products	The amount of revenue from the sale, mln US dollars
Bulk sunflower oil	1067.5
Bottled sunflower oil	140.7
Farm products	381.3
Grain, including:	877.14
barley	129.26
wheat	360.09
maize	387.79

It is necessary to construct a secondary pie chart that shows the income structure of the agro-industrial enterprise.

The solution.

A pie chart shows both the absolute value and the contribution (share) of each element of the data series to the total. A pie chart is used to depict graphically the structure and proportionality of the phenomenon or process under study. However, only one row (indicator) of data can be placed on a pie chart. A secondary pie chart is used to detail one of the sectors of a pie chart. This method of visualization helps to simplify and refine the view of small areas (sectors) of a pie chart.

When building a secondary pie chart, it is important to follow some rules:

1) the main and secondary diagrams are parts of one data series. They cannot be edited independently;

2) the sectors on the secondary circle also show the shares as in a usual chart. However, the sum of percentages is not equal to 100, but is the total value of the sector in the main pie chart (from which the secondary is separated);

3) by default, the last third of the data is displayed on the secondary circle. If for example there are 7 rows (7 sectors for the diagram) in the original table, the last three values will be displayed on the secondary diagram (Fig. 3.6).

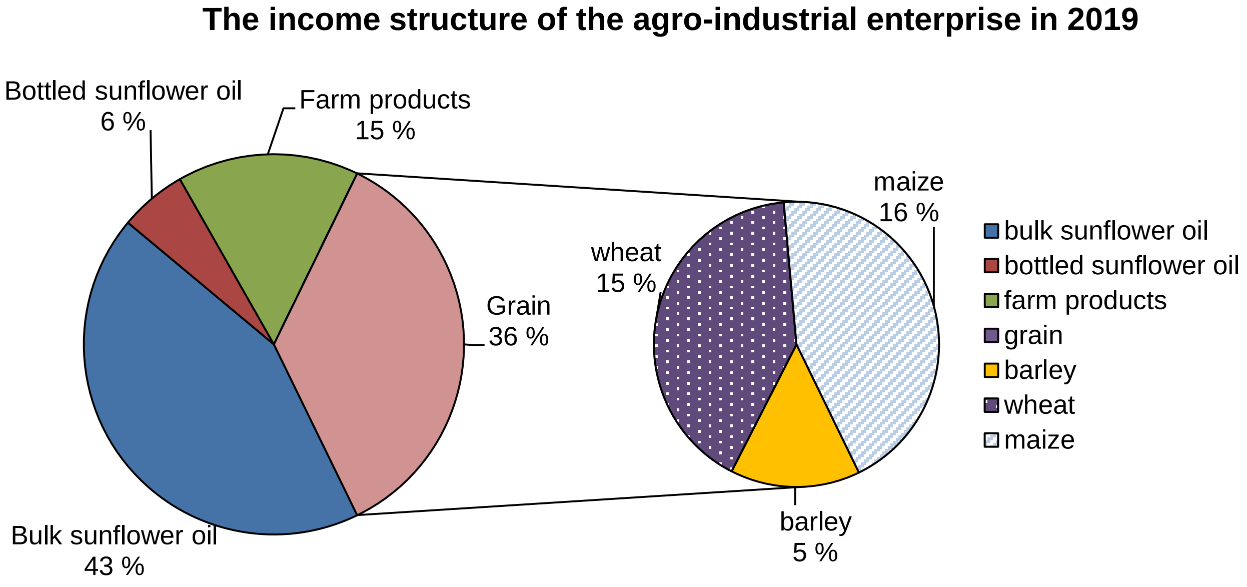


Fig. 3.6. **The secondary pie chart of the enterprise income structure**

Task 4. Build a radar chart based on the data in Table 3.6.

Table 3.6

The sales of the garden center, thousand UAH

Months	Type of goods			
	bulbs	grain	flowers	trees and bushes
January	–	140	–	–
February	–	294	–	52
March	–	462	71	110
April	–	313	84	195
May	–	178	286	189
June	–	52	381	90
July	–	–	412	22
August	–	–	345	–
September	280	–	157	121
October	420	–	87	302
November	193	–	–	297
December	–	–	–	178

The solution.

Let us represent the seasonality of sales of the garden center with the help of a radar chart (Fig. 3.7).

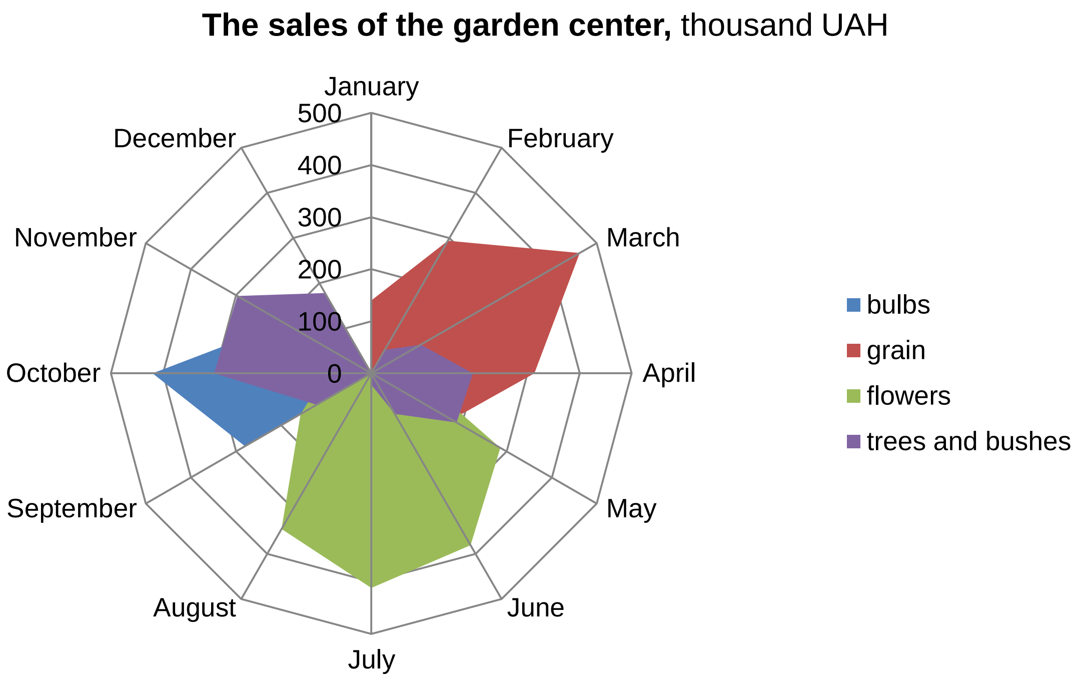


Fig. 3.7. **The radar chart of the distribution of the garden crops sales**

A radar chart is analogous to a graph depicted in the polar coordinate system and displays the distribution of values relative to their origin. The axes are pointing outwards from the center of the chart. The value of each point is indicated on the corresponding axis.

A radar chart is used if the categories cannot be directly aligned. So, the radar chart, shown in Fig. 3.7, contains 12 months of the year as categories, and seasonal sales of garden crops as values.

Thus, a complete radar chart was constructed using MS Excel to show the seasonal nature of garden crops sales.

Reference to laboratory work

The guidelines for performing the laboratory work on the topic "Visualization of socio-economic information: construction and analysis of different types of graphs" are presented in [100]. The laboratory work aims to get students acquainted with the types of graphs and gain practical skills in visualization and analysis of socio-economic information using the graphical editor of MS Excel.

Questions for self-assessment

1. What is called a statistical table? What are the functions of statistical tables?

2. What is a table layout? What are the components of a table layout?

3. What is a conjunction table? What is it used for?

4. What are the rules for building statistical tables?

5. What are the basic requirements for the construction of statistical graphs?

6. What are the main elements of a graph?

7. What is a chart? What types of charts do you know?

8. How is a radar chart built? What does a radar chart represent?

9. How are pie charts and secondary pie charts built?

10. What are cartograms and map charts? What is the difference between them?

Questions for critical rethinking (essays)

1. Explain the essence of the graphical method and its value in statistical data analysis.

2. For what purpose do we use line, column and sector charts? Give examples of their practical use in economics.

3. What current software packages, editors or statistics visualization applications do you know? Describe their functions and peculiarities.

4. Provide a detailed classification of chart types according to the purpose of the statistics visualization and the available source information space.

5. Make a comparative analysis of the practical application of statistical map types in the economy – cartograms and map charts.

4. Statistical data summarization and grouping

Basic questions:

4.1. The essence of statistical summarization and its types.

4.2. Statistical groupings and their types.

4.1. The essence of statistical summarization and its types

The statistics collected in the process of observation do not allow us to obtain generalizations of the characteristics of the studied population, to reveal patterns of its development, because in the process of observation, the characteristics of only certain units of a population are recorded. In order to obtain general characteristics, the collected information must be systematized and transformed into an orderly system of statistical indicators. The systematization of the information obtained and the generalization of the observed factors are the *content of the second stage* of a statistical study, called *summarization* and *grouping* [1; 3; 6].

Statistics is a series of sequential operations aiming to generalize specific single facts that form a set, to identify the typical features and patterns inherent in the phenomenon under study. *Summarization* is a set of generalization techniques and processing of statistical observation data designed to obtain statistical indicators and their subsequent analysis. On the basis of these indicators they characterize the size of the population, the scope of features inherent in it, the structure of the population and its qualitative composition, establish specific features and regularities of the studied phenomenon, the relationship between the characteristics, etc. [16; 17; 24; 27].

The purpose of summarization is to obtain summary data by counting single data.

The components of statistical summarization are:

1) development of a systematization and grouping program;

2) justification of the system of indicators for the characteristics of groups and the population as a whole;

3) designing table layouts in which the results are summarized;

4) definition of technological schemes of information processing, software;

5) preparation of data for processing on a computer, formation of automated databases;

6) direct summarization, generalization, calculation of indicators [32; 37; 40; 41; 43].

The types of summarization are given in Table 4.1.

Table 4.1

Classification of statistical summarization [89; 90; 101; 103; 104]

Classification feature	Type of summarization
According to the depth of material processing	Simple, complex
According to the form of information processing	Centralized and decentralized
According to the technique of execution	Automated, manual

According to the depth of processing of material summaries are divided into *simple* and *complex* [7; 28; 32; 37; 40; 41; 43; 88].

A simple summary is the operation of counting totals of the units of observation, that is, determining the size of the phenomenon under study.

A complex summary is a set of operations that include grouping of observation units, calculating the totals for each group and the population on the whole as well as presenting the grouping results in tabular form [7; 28; 32; 37; 40; 41; 43; 88].

According to the form of material processing, summaries are divided into centralized and decentralized ones.

In *a centralized summary*, all primary material is fed into one organization, where it is processed according to the adopted program, in accordance with a single method (for example, in the State Committee for Statistics of Ukraine or territorial offices of statistics).

In *decentralized summaries*, the development of statistical material is carried out in a hierarchical management system, subjected to appropriate processing at each level. For example, businesses submit reports to district statistical offices which compile information about their area and send aggregated information to regional offices or committees. They send their reports to the State Committee for Statistics of Ukraine, where the indicators for the national economy as a whole are determined.

Data processing is either *automated*, that is carried out using a PC, or *manual* [7; 28; 32; 37; 40; 41; 43; 88].

The statistical summary is organized following a special program, which must consist of the development plan and the observation program.

A statistical summary program includes [1; 3; 6; 12; 16; 17; 24; 27; 28]:

- selection of grouping features;
- establishing the order of formation of groups;
- developing a system of statistics to characterize the selected groups and the population as a whole;
- developing table layouts to provide summary results.

The statistical summary plan provides guidance on the timing and sequence of execution of the summary individual stages, its executors, and the sequence of presentation of its results [1; 3; 6; 12; 16; 17; 24; 27; 28].

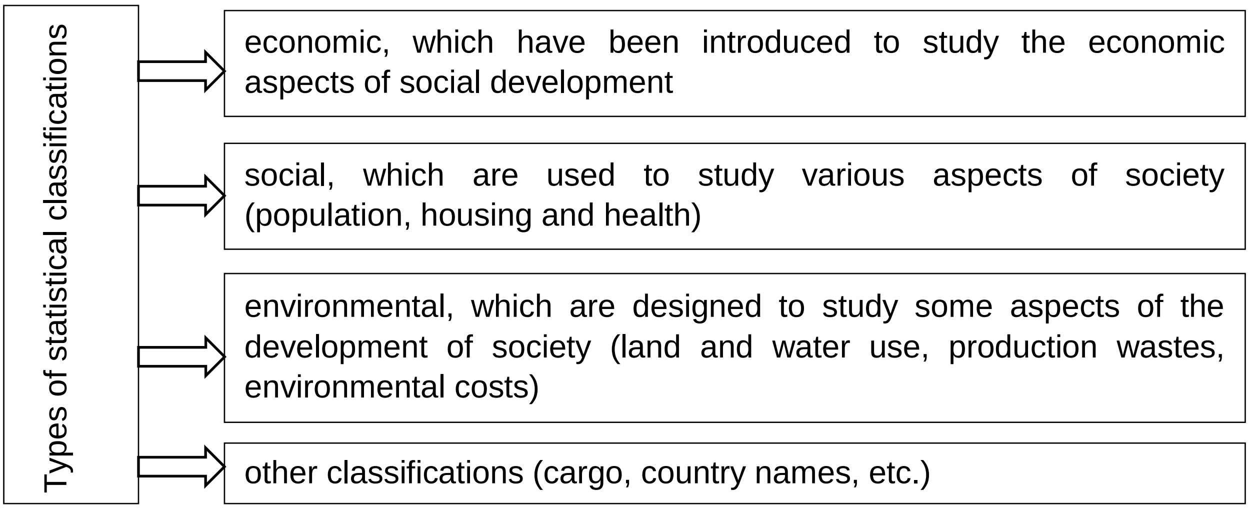
In the first stage of a statistical survey, statistical information is obtained, which is then systematized, summarized, processed, analyzed and generalized.

In the second stage of a statistical study, the population is divided into groups, homogeneous in one sense or another. To do this, the most important principles of this kind of distribution are used: in one group the elements of sets, to some extent similar to each other, are combined; the degree of similarity between the elements of one group is much higher than that of the elements belonging to different groups. In each specific statistical survey, the following problems are solved: what should be used as a basis for grouping; how many groups should be formed; how these groups should be differentiated.

The basis for dividing the elements of a set into groups can be any feature (attributive or quantitative), having a qualitatively different characteristic. This find of feature is called a *grouping* one. There may be several grouping features depending on the complexity of the phenomenon and the purpose of the study [1; 3; 6; 12; 16; 17; 24; 27; 28].

If the division of elements into groups is carried out according to attributive features, this type of grouping is called *classification* or *nomenclature*. They are developed by international and national statistical authorities and are recommended as a statistical standard.

Classification in statistics is a systematic division of phenomena and objects into specific groups, classes, categories based on their similarities or differences. A variety of classifications is the product nomenclature as a standardized list of objects and groups. The types of statistical classifications are presented in Fig. 4.1.



Types of statistical classifications

economic, which have been introduced to study the economic aspects of social development

social, which are used to study various aspects of society (population, housing and health)

environmental, which are designed to study some aspects of the development of society (land and water use, production wastes, environmental costs)

other classifications (cargo, country names, etc.)

Fig. 4.1. **The types of statistical classifications** [24; 27; 28]

Examples of modern national level classifications are those that are fully aligned with international standards [88; 89; 101; 103; 104]:

The Statistical Classification of Economic Activities in the European Community (NACE) which contains three features accepted for classification: the purpose of manufactured products; the unity of production technology; the uniformity of raw materials used;

Classification of Forms of Ownership in which the objects of classification are grouped in accordance with the established forms of ownership under the current legislation of Ukraine (state, collective, private property, etc.);

The Ukrainian Classification of Goods of Foreign Economic Activity which meets the needs of statistical services, customs bodies of foreign economic activity;

The Classification of Organizational and Legal Forms of Business – the classification of business entities (state, collective, private enterprise, etc.), organizations carrying out business activities (institution, etc.), branch offices (a subsidiary, a representative office).

A common feature of national and international classifications is that the variation of their attributes is fixed in a certain systematic form by means of codes of the classified items.

The following classification systems (nomenclatures) are used today for classification of goods and services and for characterization of economic activity:

The International Standard Industrial Classification of All Economic Activities (ISIC);

The United Nations Standard International Trade Classification (SITC);

The Harmonized Commodity Description and Coding System (HS);

The Major Product Classifier (MPC).

4.2. Statistical groupings and their types

The scientific basis of a summary is *statistical grouping* as a process of formation of homogeneous groups on the basis of division of a statistical population into parts or combining of the studied statistical units into partial populations based on their essential features [7; 37; 40; 41; 43; 88; 89; 90; 101; 103; 104].

The grouping method is the basis for the application of other methods of statistical analysis in order to distinguish the main aspects and characteristics of the phenomena under study. In the process of research, the method of grouping performs functions similar to those of the experiment in the natural sciences: by grouping based on particular features and combination of the features themselves, statistics helps to identify patterns and relationships of phenomena in the conditions that to some extent determine it. Using this method, it is possible to trace the relationship of various factors and determine the extent of their influence on the effective feature.

Statistical grouping is carried out in several successive *stages* [40; 41; 43; 88; 89; 90; 101; 103; 104]:

stage 1 – theoretical analysis of the phenomenon or process under study;

stage 2 – selection of a grouping feature;

stage 3 – determining the number of groups and the size of the interval; construction of an interval row of distribution of the population units under study based on the grouping feature (features);

stage 4 – definition and justification of the system of statistical indicators for the selection and characterization of typical groups; compiling the spreadsheets;

stage 5 – calculation of absolute, relative and average indicators;

stage 6 – tabular and graphic presentation of group results;

stage 7 – analysis of the obtained results; formulating conclusions and proposals.

Depending on the degree of complexity of the phenomenon being studied, statistical grouping may be performed based on one or more *grouping features*. Grouping is called *simple* (one-dimensional) if homogeneous groups are formed on the same basis at the same time. If homogeneous groups are formed on two or more grounds, the groupings are *complex* (Fig. 4.2).

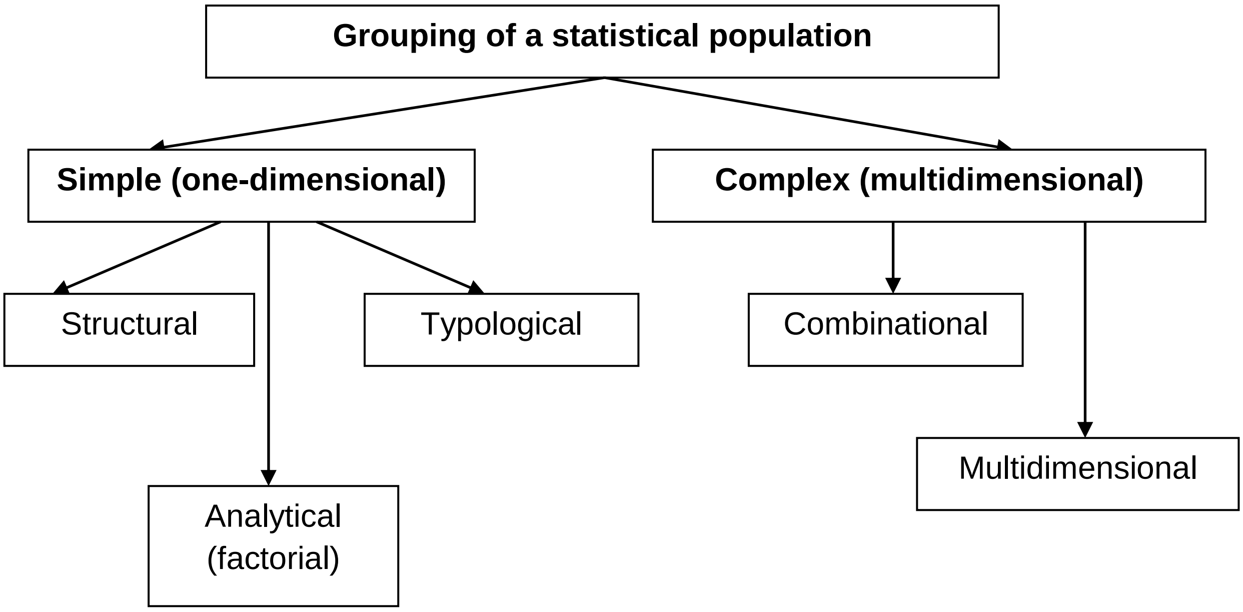


Fig. 4.2. **The types of groupings** [1; 3; 6; 7; 12; 16; 17; 37; 40; 41; 88; 90]

The types of simple (one-dimensional) grouping [1; 3; 6; 7; 12; 16; 17; 37; 40; 41; 88; 90] are here described as follows.

Structural groupings are used to study the internal structure of a statistical population and the characteristics of structural shifts. They provide information about the current state of mass phenomena and are used for operational management. Structural grouping is performed in several stages [1; 3; 6; 7; 12; 16; 17; 37; 40; 41; 88; 90]:

- selection of a grouping feature;
- identification of the required number of groups;
- determination of group parameters;
- allocation of observation units to the selected groups;
- calculation of structural characteristics;
- formulation of conclusions.

Typological grouping aims to study the prevalence of different types of economic phenomena in a statistical population. Typological groupings are usually applied to a heterogeneous population and are eliminated by means of complex multi-interval grouping. The result of typological grouping is the division into classes, socio-economic types, homogeneous groups of units [1; 3; 6; 7; 12; 16; 17; 37; 40; 41; 88; 90].

In its essence, typological grouping is a grouping classifier. Such groupings are often based on a stable list of groups that do not change or change slightly over time. An example would be the grouping of enterprises depending on the type of ownership (public, municipal, private, mixed) or economic sectors.

When performing typological groupings, it is important to correctly select the reason for grouping. For this purpose it is necessary to identify the possible type of phenomenon beforehand on the basis of analysis of the essence and regularities of its development. The number of groups and their parameters are set informally based on the selected qualitative patterns, often involving quantitative features. For example, there are 4 age groups for age grouping: preschool age up to 7 years; school age from 7 to 17 years; working age from 17 to 55 (60) years; retirement from 55 (60) years [88; 89; 101; 103; 104].

According to the technique of implementation, typological grouping is similar to the structural one, with the exception of the first stages – the feature of grouping, the number of groups, their parameters are determined on the basis of qualitative analysis. Specialized intervals are often used in such groups. Typological groupings are presented in tabular form, the object of analysis is the indicators of the structure.

Analytical groupings are intended to identify the relationship between the features under study. They enable the researcher to detect the presence and types of relations, as well as to measure their tightness and strength. All the features studied in this case are divided into two groups: factorial (independent) and effective (dependent). The relationship between them manifests itself in the fact that with the change in the average value of a factorial feature the average value of the effective feature systematically changes [1; 3; 6; 7; 12; 16; 17; 37; 40; 41; 88; 90].

Analytical groupings differ from structural and typological ones by the technique of execution, that is:

the grouping of units of the population is carried out based on the factorial feature, it is performed as a structural one;

in each selected group, the corresponding values of the effective feature are selected and on the basis of them some generalized indicator of the average value is calculated;

changes in the generic index, i.e. the average value of the effective feature across groups, are analyzed; a conclusion is drawn about the presence or absence of the relationship and its type. If with the change of the values of the factorial feature underlying the grouping the value of the effective one changes, there is a relationship between the features. If with the increase of the values of the factorial feature the value of the effective one increases, the relationship is direct; otherwise, it is inverse.

The *complex (multidimensional) grouping* implies simple and combinational groupings. *Simple* grouping is carried out on one basis. *Combinational grouping* is performed on several grounds in succession. The sequence is established based on the logic of the relationship of indicators. Typically, grouping begins with an attributive feature. In combinational grouping, the population is logically and sequentially divided into homogeneous parts on separate grounds: into groups on one basis, then within each group into subgroups based on another feature and so on [1; 3; 6; 7; 12; 16; 17; 37; 40; 41; 88; 90].

Such groupings are intended for a deeper analysis of the phenomenon under study. They enable the researcher to identify and compare differences and relationships between the features that cannot be identified on the basis of isolated groups for each of them. However, it should be borne in mind that when studying a large number of features the use of the combinational grouping is impossible: it leads to the fragmentation of information, and therefore to obscuring the manifestations of patterns. Even in the presence of large amounts of information, we have to confine ourselves to two or four features.

Combinational grouping based on two features (X, Y) is performed in the form of a chess table, in which the values of one X-feature are placed in rows while the values of the other Y-feature in columns. There are frequencies of joint manifestation of the Y-feature value in the j-th column and the value of the X-feature in the i-th row at the intersection of the j-th column and the i-th row of the matrix [1; 3; 6; 7; 12; 16; 17; 37; 40; 41; 88; 90].

Multivariate groupings. These groupings include those formed based on multiple features at the same time. The purpose of multivariate groupings is to classify data on the basis of multiple features, that is, to isolate groups of statistical units that are homogeneous in terms of several features at a time. In the process of such grouping, for example, the tasks of typing are solved, and the economic or social types of phenomena are distinguished. Thus, using the techniques of multivariate classification a whole set of industrial enterprises can be divided into into small, medium and large, using the following characteristics: the number of the industrial and production personnel, production volume, consumption of material resources and so on. It is possible to distinguish between the types of enterprises according to their financial status on the basis of such indicators as: the size of profit, the level of production profitability, the level of capitalization, the level of liquidity of securities, etc. In psychology, multivariate groupings are used to distinguish between the types of people as to their professional suitability, and in medicine to diagnose diseases based on many symptoms.

Two main approaches can be used to perform multivariate grouping [1; 3; 7; 27; 28; 32; 37; 40; 41; 43; 89; 90; 101; 103; 104]:

- with the first one a summarizing index of the population features of the grouping is calculated and a simple grouping is performed;
- the second approach uses the cluster analysis method.

The first approach is represented by the *method of a multivariate average.* The grouping algorithm according to this method involves performing some successive stages [1; 3; 7; 27; 28; 32; 37; 40; 41; 43; 89; 90; 101; 103; 104].

The first stage is the use of a matrix of the features' absolute values for all statistical units (matrix x_{ij}, where i = n; 1 are statistical units; j = k, 1 are features).

The second stage. The absolute values of the features are replaced by the levels normalized based on the average value:

$$P_{ij} = \frac{x_{ij}}{x_j}, \qquad (4.1)$$

where P_{ij} is the normalized value of the j-th feature in the i-th statistical unit;

$\overline{x}_j$ is the average value of the j-th feature, $\overline{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}$.

The third stage: a multivariate statistical average is calculated for each unit:

$$\overline{P}_i = \frac{\sum_{j=1}^{k} P_{ij}}{k}, \qquad (4.2)$$

where k is the number of grouping bases.

The fourth stage: according to the values of the multivariate mean, the population is divided into homogeneous groups, i.e. simple grouping is performed based on the multivariate mean.

Along with the primary grouping, which is performed in stages, a secondary one is used in statistics, which is carried out on the basis of what was previously done. Such grouping is used to improve the characteristics of the phenomenon being investigated if the primary grouping does not give a clear characteristic of distribution of the population units.

The following problems in statistics are solved with the help of groups [28; 32]:

- studying the composition of statistical populations;
- selection of individual types of phenomena within the population;
- identification of cause and effect relationships between different features within the population;
- classification of the population units according to many features.

The main categories of the grouping method are the grouping feature (the basis for grouping) and the interval. Of fundamental importance for the formation of groupings is the choice of the grouping feature, on the basis of which different types, groups and subgroups are distinguished. The most substantial features are accepted for grouping. Grouping can be based on an attributive (qualitative) or a quantitative feature.

The grouping feature (the basis for grouping) is the feature according to which the formation of homogeneous groups is performed. The selection of the grouping feature is made according to the objectives of the statistical survey. The grouping feature is usually an essential one. A prerequisite for performing any grouping is the ordering of the statistical population based on the values of the grouping feature.

The choice of the grouping feature is always based on the analysis of the qualitative nature of the phenomenon under study. Grouping is usually performed based on one of essential, easily recognizable features that have both attributive (qualitative) and quantitative nature.

For an *attributive* feature, the number of groups corresponds to the number of the feature varieties. Thus, for the division of enterprises depending on the ownership forms the following forms are applied: state, communal, private, collective, international organizations and legal entities of other states.

Quantitative (variational) grouping raises questions about the number of groups and grouping intervals.

An *interval* is a set of values of a feature that vary in a group. It defines the quantitative boundaries of groups, and its width is the interval between the maximum and minimum values of a feature in a group.

The following types of intervals are used to *perform grouping* [28; 32]:

• *equal* (the width of the interval is the same in all selected groups);

• *unequal* (in each group, the width of the interval is different; it can change naturally (for example, grow evenly), or arbitrarily);

• *closed* if the upper and lower limits of the intervals are known (the maximum and the minimum value of the feature in groups);

• *open* if only one upper or lower limit of the interval is known.

In the grouping process, the following principles should be taken into account: the differences between the observations units assigned to one group should be smaller than between the units assigned to different groups. Determining the required number of groups is done in accordance with such a rule that the number of groups should be sufficient to objectively represent the population under study. In a large number of groups, the differences between them become inconspicuous, and in the groups themselves due to their low content, the law of large numbers ceases to exist and manifestations of accidentality may happen. In an independent number, statistical units with features that differ significantly in values may be included in one group.

The number of selected groups is influenced by the following *factors* [28; 32]:

• the level of fluctuation of the grouping feature – the greater the variation of the grouping, the greater the number of groups to be formed under other equal conditions;

• the size of the statistical population under study – the larger is the size of the population under study, the larger is the number of groups to be formed.

The selected groups should be sufficiently filled. The presence of empty groups or the small number of statistical units in them indicate that their number is incorrect. Approximately the number of groups can be determined using the empirical dependence of the Sturges formula [6; 7; 12; 16; 17; 24; 27; 28; 32; 37; 40; 41; 43; 88; 89; 90; 101; 103; 104]:

$$m = 1 + 3.322 \times \lg N, \tag{4.3}$$

where m is the number of groups;

N is the number of units of the statistical population.

The Sturges dependence produces good results if the population consists of a large number of units, the distribution is close to normal and equal intervals are applied. In practical calculations, one can use the ratios obtained on the basis of the Sturges formula (Table 4.2).

Table 4.2

The ratios obtained on the basis of the Sturges formula [28; 32]

No.	15 – 24	25 – 44	45 – 89	90 – 179	180 – 359	360 and more
m	5	6	7	8	9	10

There is another way to determine the number of the formed groups. It is related to the application of the standard deviation equal or unequal to σ: if the width of the interval is 0.5 σ, there are 12 groups, if it is 2/3σ, there are 9 groups, if it is equal to σ, there are 6 groups.

Definition of group parameters. In each selected group, the following parameters are calculated [1; 3; 6; 12; 24; 27; 28; 32; 37; 40; 41; 43; 89; 90; 101; 103; 104]:

- the upper limit of the minimum interval and the lower limit of the minimum interval;
- the width of the interval h;
- the middle of the interval x_i.

The lower limit of the interval is the lowest value of the feature in the group. *The upper limit* of the minimum interval is called the highest value of the feature in the group.

Grouping intervals may be equal and unequal (progressively increasing, progressively decreasing, arbitrary, specialized). If the variation of a feature manifests itself within relatively narrow boundaries, and the distribution of statistic units is sufficiently uniform, groupings are formed at regular intervals [32; 37; 40; 41; 43; 89; 90; 101; 103; 104]:

$$h = \frac{x_{max} - x_{min}}{m}. \tag{4.4}$$

On the basis of the calculated width of the interval h, the boundaries of the intervals are consistently determined. The definition of boundaries begins with the first group. The lower boundary of the interval is taken to be equal to the minimum value of the feature in the population, $x_i^{lower} = x_{min}$. The upper boundary of the first interval is calculated as $x_i^{upper} = x_i^{lower} + h$. Further the intervals are calculated using the formulas [32; 41; 43; 89; 90; 101; 103; 104]:

$$x_i^{lower} = x_{i-1}^{upper};$$
$$x_i^{upper} = x_{i-1}^{upper} + h. \tag{4.5}$$

The middle of the interval (the central variant) is defined as the half-sum of the upper and lower boundaries of the interval. The mid-interval parameter is used to calculate the generalized characteristics of the population under study.

Distribution of units of the population among groups. The main task of this stage is to calculate the number of units that fall into each of the selected n_i groups. There is uncertainty in the allocation of observation units to the selected groups (especially if the grouping feature is continuous): which group should include the units with the values of the features that are the same as the interval limits? To eliminate uncertainty, the principle of equality is used. Such units are included in the group in which the lower boundary coincides with the value of the feature. For example, there are groups of enterprises formed based on the volume of production, million UAH: 400 – 450; 450 – 500; 500 – 550; 550 – 600; 600 – 650. Which group should include an enterprise with a production volume of 500 million UAH? In accordance with the principle of equality, it is the second group.

Calculation of the structural characteristics. The calculation of the parameters is done by determining the proportion of each group units in the total volume of the statistical population. Like any relative value, this indicator can be defined as coefficients [89; 90; 101; 104]:

$$d_i = \frac{n_i}{N} \text{ or percentage } d_i = \frac{n_i}{N} \times 100. \tag{4.6}$$

Having calculated such shares for all groups, we obtain the structure of the studied statistical population which is equal to the complete set of shares: $\sum d_i = 1$.

Formulation of conclusions about the composition of the population. As for the structure of the groups, the conclusions hold two positions:

• what values of the feature in the population occur most often, what values are most rare;

• what is the nature of change in the structure depending on the change in the value of the feature. With x increasing, the proportion may increase or decrease. This is fairly typical of economic indicators.

The conclusions are obligatory, otherwise the sense of grouping is offset. Structural group data are usually presented in the form of a corresponding table.

Example 4.1. Performing structural grouping. Groupings should be made at regular intervals according to the length of service of the employees of the ABC company. The head office employs 50 employees with work experience from 0 to 26 years.

According to the Sturges formula, we calculate the optimal number of groups [32; 37; 40; 41; 43; 89; 90; 101; 103; 104]:

$$m = 1 + 3.322 \times \lg 50 \approx 7.$$

The constant width of the intervals is then calculated [32; 37; 40; 41; 43; 89; 90; 101; 103; 104]:

$$h = \frac{x_{max} - x_{min}}{m} = \frac{26 - 0}{7} \approx 4.$$

Structural grouping is performed according to the parameters calculated in Table 4.3.

Table 4.3

Grouping of the ABC company employees based on the length of service
[32; 37; 40; 41; 43; 89; 90; 101; 103; 104]

Group No.	Group parameters, years				Number of employees in the group	Share of employees in groups,%
	Interval limits		interval width	the middle of the interval		
	lower	upper				
1	2	3	4	5	6	7
1	0	4	4	2	6	12
2	4	8	4	6	8	16

Table 4.3 (the end)

1	2	3	4	5	6	7
3	8	12	4	10	11	22
4	12	16	14	14	13	26
5	16	20	4	18	6	12
6	20	24	4	22	4	8
7	24	28	4	26	2	4
Total	0	28	28	14	50	100

Based on the grouping one can draw the following conclusions:

1. The largest number of employees of the company in the head office have the length of service from 12 to 16 years, they represent 26 % of the total number of the company employees; the smallest number is with experience from 24 to 28 years, their share is 4 %.

2. As the length of service increases, the number of company employees increases, reaching its maximum in group 4, and then decreases.

An example of typological grouping is given in Table 4.4.

Table 4.4

Grouping of the territorial community population based on the age categories

Age categories of the population	Interval limits		The number of people in the group	Structure indicator in %
	lower limit	upper limit		
Preschool	up to	7	192	14.3
School	7	17	218	16.3
Working	17	55 (60)	574	42.8
Retirement	55 (60)	and more	357	26.6
Total			1340	100.0

Example 4.2. Analytical grouping. It is necessary to establish the relationship between the length of service and the size of salary of the site employees. The length of service and monthly salary of every employee is known. Here the factorial feature is the length of service, the effective feature is the salary. On a factor basis, structural grouping was previously conducted.

The company's employees were divided into 7 homogeneous groups. Additionally, in each group, the total salary-per-employee pay per month was calculated and its average value was calculated using the formula of a simple arithmetic mean [1; 3; 6; 7; 32; 37; 40; 41; 43; 88; 89; 90;101; 103; 104]:

$$\bar{y}_i = \frac{\sum_{i=1}^{n} y_i}{n}, \tag{4.7}$$

where y_i is the salary of the i-th employee.

The results of the calculations are given in Table 4.5. The object of analysis in Table 4.5 is the mean value of the effective feature – the average monthly salary of the employees in groups and the middle of the interval of the average length of service.

If the mean of the effective feature across the groups has some difference, the relationship between the features can be considered established. If the average result with the transition from one group to another is virtually unchanged, there is no relationship between the features.

Table 4.5

Research into the company employee' salary dependence on the length of service

[1; 3; 6; 12; 16; 17; 24; 27; 28; 32; 40; 41; 43; 89; 90; 101; 103; 104]

Group No.	Groups of employees of the company based on experience, years			The number of the company employees in a group, people	The total salary of the company employees in a group, UAH	The average salary in a group, UAH
	Interval limits		The middle of the interval			
	lower	upper				
1	0	4	2	6	18 000	3 000
2	4	8	6	8	28 000	3 500
3	8	12	10	11	41 800	3 800
4	12	16	14	13	59 800	4 600
5	16	20	18	6	34 800	5 800
6	20	24	22	4	27 200	6 800
7	24	28	26	2	14 800	7 400
Total	0	28	14	50	224 400	4 488

In the example considered, a change in work experience leads to a change in salaries. Thus, by means of analytical grouping, it is possible to establish the relationship between the features, but not to describe it. For the description it is necessary to use the correlation and regression analysis.

Example 4.3. Perform grouping on the basis of a multivariate mean: it is necessary to form homogeneous groups of statistical units on three grounds. The size of the statistical population is 10 objects each of which is characterized by conditional values of the features. The initial data and the calculation of the multivariate mean are presented in Table 4.6.

Thus, a conditional characteristic is calculated for each of the 10 objects, which is a multivariate mean that replaces the three primary features. Based on the multivariate mean as a grouping feature, one-dimensional structural grouping must be performed.

For this purpose the following parameters are determined: the number of homogeneous groups $m = 1 + 3.322 \lg 10 \approx 4$, the width of the intervals $h = \frac{x_{max} - x_{min}}{m} = \frac{1.7 - 0.3}{4} = 0.33$.

Table 4.6

Calculation of the multivariate mean [101; 103; 104]

Object number	The absolute values of the features			Normalized feature values			Calculation of the multivariate mean	
	X_{11}	X_{12}	X_{13}	P_{11}	P_{12}	P_{13}	$\sum_{j=1}^{k} P_{ij}$	$\bar{P}_i$
1	2	18	62	0.57	1.67	1.0	3.24	1.08
2	1	5	40	0.29	0.46	0.64	1.39	0.46
3	2	7	40	0.57	0.65	0.64	1.86	0.62
4	6	15	77	1.71	1.39	1.24	4.34	1.4
5	1	9	43	0.29	0.83	0.69	1.81	0.60
6	6	20	95	1.711	1.85	1.53	5.1	1.7
7	5	9	62	1.42	0.83	1.0	3.25	1.08
8	1	1	46	0.29	0.10	0.74	1.13	10.38
9	8	15	84	2.29	1.39	1.35	5.03	1.68
10	3	9	72	0.86	0.83	1.16	2.85	2.95
Total	35	108	621	10.00	10.00	10.00	10.00	10.00
$\bar{x}_j$	3.5	10.8	62.1	1.0	1.0	1.0	1.0	1.0

Structural grouping is presented in Table 4.7.

Table 4.7

Grouping of objects based on the multivariate mean [101; 103]

Group No.	Group parameters			Object number	The number of objects in a group, n_i
	lower limit	upper limit	interval width		
1	0.38	0.71	0.33	2, 3, 5, 8	4
2	0.71	1.04	0.33	10	1
3	1.04	1.37	0.33	1, 7	2
4	1.37	1.7	0.33	4, 6, 9	3
Total	0.38	1.7	1.32	–	10

When multivariate grouping is performed, each unit of a population having a set of k features is treated as a point in the k-dimensional space of the feature space; each feature is treated as a coordinate. The task of classification in this case is to show the density of objects in this space. Different algorithms are used for this purpose, but homogeneous groups are always distinguished on the basis of the approximation of objects according to a set of features. The closeness of objects, that is, the degree of similarity of the units of a population may be measured based on different criteria.

There are three types of similarity measures [1; 3; 6; 7; 32; 37; 40; 41; 43]:

- similarity coefficients;
- correlation coefficients;
- distance indicators.

Similarity coefficients are used to measure the closeness between a pair of objects, each of which takes on the value of 0 or 1.

Correlation coefficients are used as a measure of the bond strength between statistical units or between features. The linear correlation coefficients are used to measure the bond closeness of quantitative features.

In a cluster analysis, the measure of similarity is the distance between two objects i and j. Euclidean distance is used for quantitative features.

Sometimes, depending on the goals of economic research, it is necessary to regroup the data to ensure that the structures of two populations are comparable with one feature. The result of regrouping is called a *secondary grouping*. The regrouping is performed either by combining or splitting of the intervals of the primary grouping.

If the limits of the intervals of the primary and secondary groups coincide, the frequencies (shares) of the combined intervals are simply summed up. When splitting of the primary grouping interval is performed, the frequencies are distributed between the newly formed groups in proportion to the ratio of parts of the length of the original interval.

Within the interval, the distribution is assumed to be uniform. The technique of the data regrouping should be considered based on the example of the distribution of employees according to the size of the average monthly salary in two industrial sectors (Table 4.8).

Table 4.8

Distribution of employees based on the average annual salary

[1; 3; 6; 7; 12; 16; 17; 24; 27; 28; 32; 37; 40; 41; 43; 89; 90; 101; 103]

Sector A		Sector B	
Salary, UAH	Share of employees, %	Salary, UAH	Share of employees, %
up to 160	15	up to 160	12
160 – 180	20	160 – 190	30
180 – 200	26	190 – 220	21
200 – 220	23	220 – 250	18
220 – 240	9	250 – 280	13
240 and more	7	280 and more	6
Total	100	Total	100

The results of the primary grouping cannot be directly compared, since the grouping intervals are different: in sector A, the width of the interval is 20, in sector B it is 30 UAH. We regroup the data to form five groups with an interval of h = 40 UAH. Obviously, the separation intervals in sector A should be merged while in sector B they are to be broken.

The results of the secondary grouping are given in Table 4.9.

Table 4.9

Secondary grouping of employees based on the average monthly salary

[1; 3; 6; 7; 12; 16; 17; 24; 27; 28; 32; 37; 40; 41; 43; 90]

Salary, UAH	Share of employees, %	
	Sector A	Sector B
1	2	3
Up to 160	15	12

Table 4.9 (the end)

1	2	3
160 – 200	20 + 26 = 46	$30+\frac{1}{3}21=37$
200 – 240	23 + 9 = 32	$\frac{2}{3}21+\frac{2}{3}18=26$
240 – 280	7	$\frac{1}{3}18+13=19$
240 and more	–	6
Total	100	100

Comparing the shares of the secondary grouping, we can see that in sector B, the population of employees in terms of salaries is more differentiated. By regrouping the data, you can proceed from structural to typological grouping.

Important concepts

Simple summary is an operation of calculating the total sum of the observation units, that is, determining the size of the phenomenon under study.

Complex summary is a set of operations that include grouping of observation units, calculating the totals for each group and the population as a whole, as well as presenting the grouping results in tabular form.

Centralized summary is a process where the entire primary material is concentrated in one organization where it is processed according to the adopted program using a single method.

Decentralized summary is the development of statistical material carried out according to a hierarchical management system and subjected to appropriate processing at each level.

Statistical grouping is the process of forming homogeneous groups on the basis of the breakdown (division) of a statistical population into parts or combining the studied statistical units into a part of the population based on essential features.

Classification is a more stable division of units of a population (observation) as compared to grouping. Classifications have been in use for a long time, though they may undergo more or less significant changes over time. Classifications are approved in the form of both national and international standards.

Similarity coefficients are coefficients used to measure the degree of approximation between a pair of objects, each of which takes the value of 0 or 1.

Correlation coefficients are coefficients that are used as a measure of the strength of the relationship between statistical units or features. Linear correlation coefficients are used to measure the closeness of the quantitative feature relationship.

The measure of distance between two objects i and j is the similarity measure used in the cluster analysis. Euclidean distance is used for quantitative features.

Specialized intervals are the intervals used to separate the same types of the same feature from a population for phenomena that occur under different conditions.

Typological grouping intervals are intervals that are formed based on socio-economic meaning rather than mathematical principles. The interval limit is regarded as a conditional limit on the transition of quantity to new quality.

Typical tasks

Task 1. There is some information about the number of books received by students by subscription during the last academic year. It is necessary to construct a ranked and discrete variational distribution series, marking the elements of the series (Table 4.10).

Table 4.10

The input data

2	4	4	7	6	5	2	2	3	4
4	3	6	5	4	7	6	6	5	3
2	4	2	3	5	7	4	3	3	2
4	5	6	6	10	4	3	3	2	3

The solution.

The considered population is a set of variants of the number of books received by students. It is necessary to calculate the number of such variants and arrange them in the form of variational ranked and variational discrete distribution series (Tables 4.11, 4.12).

Table 4.11

The ranked distribution series

Number of books received	Number of students who received the books
2	7
3	9
4	9
5	5
6	6
7	3
10	1
Total	40

Table 4.12

The discrete series

Number of books received	The proportion of students who received the books in the whole population
2	7/40 = 0.175
3	9/40 = 0.225
4	9/40 = 0.225
5	5/40 = 0.125
6	6/40 = 0.150
7	3/40 = 0.075
10	1/40 = 0.025
Total	1

Task 2. There are some data on the value of fixed assets at 50 enterprises (thousand UAH). It is necessary to build a distribution series, forming 5 groups of enterprises (at regular intervals). The input data are given in Table 4.13.

Table 4.13

The input data

18.8	16.0	12.6	20.0	30.0	16.4	14.6	18.4	11.6	17.4
10.4	26.4	16.2	15.0	23.6	29.2	17.0	15.6	21.0	12.0
10.2	13.6	16.6	15.4	15.8	18.0	20.2	16.0	24.0	28.0
16.4	19.6	27.0	24.8	11.0	15.8	18.4	21.6	24.2	24.8
25.8	25.2	13.4	19.4	16.6	21.6	30.0	14.0	26.0	19.0

The solution.

For the calculation, we choose the largest and smallest value of the enterprises' fixed assets. This is 30.0 and 10.2 thousand UAH.

Find the size of the interval: h = (30.0 − 10.2): 5 = 3.96 thousand UAH.

Then the first group will consist of enterprises whose size of fixed assets is from 10.2 thousand UAH to 10.2 + 3.96 = 14.16 thousand UAH. There will be 9 such enterprises. The second group will include enterprises whose size of fixed assets will be from 14.16 thousand UAH. to 14.16 + 3.96 = = 18.12 thousand UAH. There will be 16 such enterprises. Similarly, we find the number of enterprises in the third, fourth and fifth groups.

The obtained series of distribution is given in Table 4.14.

Table 4.14

Grouping of enterprises based on the value of fixed assets

Groups of enterprises based on the size of fixed assets, thousand UAH	Number of enterprises
10.2 – 14.16	9
14.16 – 18.12	16
18.12 – 22.08	11
22.08 – 26.04	8
26.04 – 30.0	6

Task 3. We have some information about light industry enterprises (Table 4.15).

Table 4.15

The input data

No.	Average number of workers, people	Fixed assets, thousand UAH	Production volume, million UAH	No.	Average number of workers, people	Fixed assets, thousand UAH	Production volume, million UAH
1	2	3	4	5	6	7	8
1	100	369	560	16	102	256	410
2	140	473	760	17	96	220	370
3	94	251	440	18	98	240	330
4	83	280	520	19	84	106	210

Table 4.15 (the end)

1	2	3	4	5	6	7	8
5	157	590	800	20	76	180	160
6	195	1200	960	21	96	250	300
7	54	160	310	22	85	230	240
8	120	480	570	23	110	370	240
9	180	970	820	24	112	350	230
10	125	400	440	25	67	125	150
11	43	120	100	26	63	140	130
12	256	900	990	27	250	1150	920
13	182	670	870	28	212	790	650
14	124	500	410	29	184	290	340
15	110	379	310	30	137	275	320

Conduct grouping of enterprises based on the number of workers, form 6 groups at regular intervals. Calculate for each group: the number of enterprises, the number of workers, the volume of production per year, the average actual productivity per worker, the volume of fixed assets, the average size of fixed assets of one enterprise, the average value of products produced by one enterprise. The results of the calculations should be presented as a table. Draw conclusions.

The solution.

For the calculations, we choose the largest and smallest values of the average number of workers at the enterprise. It is 43 and 256.

The size of the interval is to be found: h = (256 − 43): 6 = 35.5.

Then the first group will consist of enterprises with an average number of employees ranging from 43 to 43 + 35.5 = 78.5 people. There will be 5 such enterprises. The second group will include enterprises with an average number of employees ranging from 78.5 to 78.5 + 35.5 = 114 people. There will be 12 such enterprises. Similarly, we find the number of enterprises in the third, fourth, fifth and sixth groups.

Put the obtained distribution series in Table 4.16 and calculate the necessary indicators for each group:

Table 4.16

Grouping of enterprises based on the number of workers

Enterprise groups based on the number of workers	Number of enterprises, units	Number of workers, people	Production volume, mln UAH	Average actual labor productivity of one worker, UAH (col. 4 : col. 3)	The volume of fixed assets, thousand UAH	The average size of fixed assets of one enterprise (col. 6 : col. 2)	Average value of products produced by one enterprise, mln UAH (col. 4 : col. 2)
43 – 78.5	5	303	850	2.805	725	145	170
78.5 – 114	12	1170	4160	3.556	3301	375	346.667
114 – 149.5	5	646	2500	3.870	2128	426	500
149.5 – 185	4	703	2830	4.026	2520	630	707.500
185 – 220.5	2	407	1610	3.956	1990	995	805
220.5 – 256	2	506	1910	3.775	2050	1025	955

As can be seen from Table 4.16, the second group of enterprises is the largest one. It includes 12 enterprises. The fifth and sixth groups (two companies each) are the most numerous ones. These are the largest enterprises (in terms of the number of employees).

As the second group is the largest one, the volume of output per year at the enterprises of this group and the volume of fixed assets are significantly higher than others. However, the average actual productivity of one wage earner at the enterprises of this group is not the highest. Enterprises of the fourth group are leading here. This group also has a large amount of fixed assets.

In conclusion, the average size of fixed assets and the average value of products produced by one enterprise are directly proportional to the size of the enterprise (as to the number of workers).

Task 4. According to the data (Table 4.17) conduct: a) combinational grouping of enterprises according to the given features, forming three groups at regular intervals; present the results graphically; provide an economic interpretation of the data obtained; b) analytical grouping that would show the dependence of labor productivity on stock-taking.

Table 4.17

The input data

Enterprise number	The volume of goods sold, thousand UAH	The amount of money invested, thousand UAH	Enterprise number	The volume of goods sold, thousand UAH	The amount of money invested, thousand UAH
1	9.1	46.6	13	6.4	11.4
2	8.8	26.0	14	5.8	10.8
3	4.5	14.2	15	6.0	19.7
4	5.8	15.2	16	5.8	21.4
5	12.4	42.2	17	8.9	34.5
6	16	54.0	18	12.6	42.1
7	10.2	30.6	19	4.7	18.6
8	6.4	44.2	20	5.7	37.2
9	10.2	38.2	21	12.2	49.5
10	6.7	23.4	22	8.9	37.6
11	10.4	22.5	23	10.6	23.8
12	5.4	20.2	24	11.6	21.5

The solution.

Combinational grouping can be accomplished on two grounds: the volume of goods sold and the amount of invested funds. Applying regular intervals, we define their width and form intervals.

For the volume of goods sold:

$$h = \frac{x_{max} - x_{min}}{m} = \frac{16 - 4.0}{3} = 4 \text{ thousand UAH;}$$

intervals: 4 – 8; 8 – 12; 12 – 16.

For the amount of money invested:

$$h = \frac{54 - 10.8}{3} = 14.4 \text{ thousand UAH;}$$

intervals: 10.8 – 25.2; 25.2 – 39.6; 39.6 – 54.

The combinational grouping of enterprises is given in Table 4.18.

Table 4.18

The combinational grouping of enterprises

The volume of goods sold, thousand UAH	The amount of invested funds, thousand UAH			All together
	10.8 – 25.2	25.2 – 39.6	39.6 – 54	
4 – 8	9	2	–	11
8 – 12	3	4	2	9
12 – 16	–	–	4	4
Total	12	6	6	24

The data in Table 4.18 indicate that there is a direct relationship between the volume of goods sold and the amount of funds invested, which is also confirmed by the appearance of the scatterplot of the available values of the variables (Fig. 4.3).

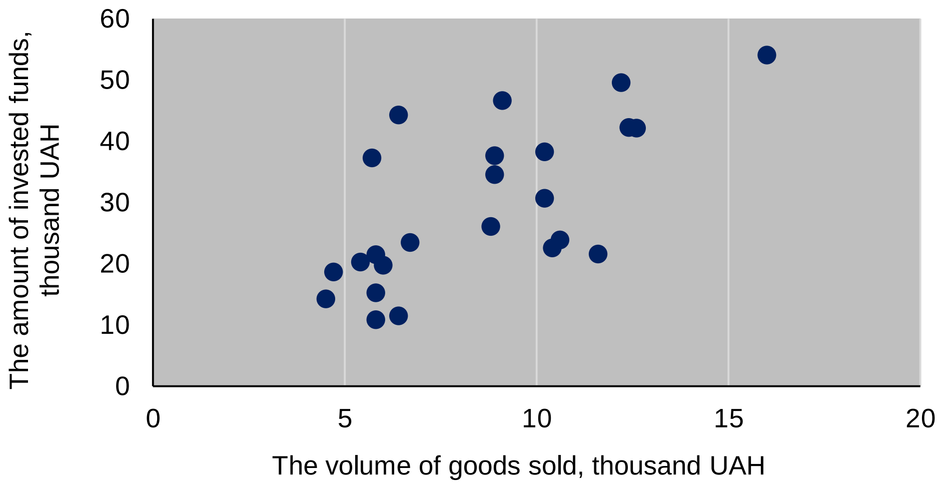


Fig. 4.3. **The dependence of the volume of goods sold on the amount of invested funds**

To confirm the relationship between the indicators, let's conduct analytical grouping (Table 4.19).

Table 4.19

The analytical grouping of enterprises

The volume of goods sold, thousand UAH	The number of enterprises	The amount of invested funds, thousand UAH	
		All together	Per enterprise
4 – 8	11	236.3	21.48
8 – 12	9	282.3	31.37
12 – 16	4	187.8	46.95
Total	24	706.4	–
On average	–	235.5	29.43

The comparison of two indicators with the help of analytical grouping confirms the hypothesis that there is a correlation between the volume of goods sold and the amount of money invested. This relationship is directly proportional, that is as the volume of investments increases, so does the volume of goods sold.

Task 5. We have some data on the quality of work of bank employees servicing the clients of STD Bank last month. Data are presented on a twelve-point system with grouping on a four-point system (Table 4.20).

Table 4.20

The input data

Rating	The number of bank employees
10 – 12 points	25
7 – 9 points	56
4 – 6 points	42
1 – 3 points	14

It is necessary to make a secondary grouping by forming the following groups on the following system: excellent (11 – 12 points); very good (10 points); good (7 – 9 points); satisfactory (5 – 6 points); unsatisfactory (4 points); very unsatisfactory (3 points); extremely unsatisfactory (1 – 2 points).

The results and the order of calculations are given in Table 4.21.

Table 4.21

The calculation data

Group	The order of calculations	The number of bank employees
Excellent	(25/3) × 2 = 17	17
Very good	(25/3) × 1 = 8	8
Good	56	56
Satisfactory	(42/3) × 2 = 28	28
Unsatisfactory	(42/3) × 1 = 14	14
Very unsatisfactory	(14/3) × 1 = 5	5
Extremely unsatisfactory	(14/2) × 2 = 9	9

The number of bank employees in groups is rounded during the calculations because there may not be a noninteger number of bank employees. Thus, due to regrouping, we obtained a grouping of the quality of bank employees' work as to servicing the STD Bank clients according to the proposed scale.

Reference to laboratory work

The guidelines for performing laboratory work on the topic "Summary and grouping of statistics" are presented in [100]. The laboratory work is aimed at mastering the skills in grouping data in MS Excel. The task is to group statistics using the MS Excel add-in "Data Analysis".

Questions for self-assessment

1. What is the task of statistical summary?
2. What are the techniques for grouping statistics?
3. What are the main statistical classifications?
4. List the tasks of grouping and their significance for a statistical survey.
5. Outline the grouping features and the principles of selection of features.
6. Define the number of groups and the size of the grouping intervals.
7. Describe the types of groups.
8. Describe the grouping functions.
9. Describe the principle of the secondary grouping method.
10. What are the principles of making the grouping intervals?

Questions for critical rethinking (essays)

1. The essence of statistics summary in the process of creating the information and analytical research base.

2. The distinction between classification and grouping: a practical aspect.

3. What are the functions performed by means of grouping in the economic and statistical analysis?

4. Explain the distinctions between multivariate and combinational grouping.

5. What is the essence of the practical component in building a typological, structural and analytical grouping?

Section 2
Statistical indicators and distribution series

5. Statistical indicators

Basic questions:

5.1. Absolute indicators.
5.2. Relative indicators.
5.3. Average indicators.

5.1. Absolute indicators

Information on the size, proportions, changes in time, other patterns of socio-economic phenomena is created, transmitted and stored in the form of statistics. From a philosophical point of view, a statistical indicator is a measure, that is, a unity of qualitative and quantitative presentation of a particular property of a socio-economic phenomenon or process.

The qualitative content of the indicator is determined by the essence of the phenomenon, manifested in the name: fertility, yield, profitability, etc. The quantitative aspect is represented by the number and its indicator. Because statisticians study social phenomena in specific conditions of space and time, the value of any indicator is determined in relation to these attributes. In this sense it is said, for example, that the capital of a certain firm at the beginning of 2020 amounted to 4400 million UAH.

The link between the qualitative content and numerical expression is the rule of construction – an indicator model that reveals its statistical structure, establishing what, where, when, and how it should be measured. The model justifies the units taken for measurement, data acquisition technology, computational operations [6; 27; 34]. The indicator model is extremely important to ensure the reliability of statistical information.

The indicators are classified according to the method of calculation, the time and the analytical functions (Fig. 5.1).

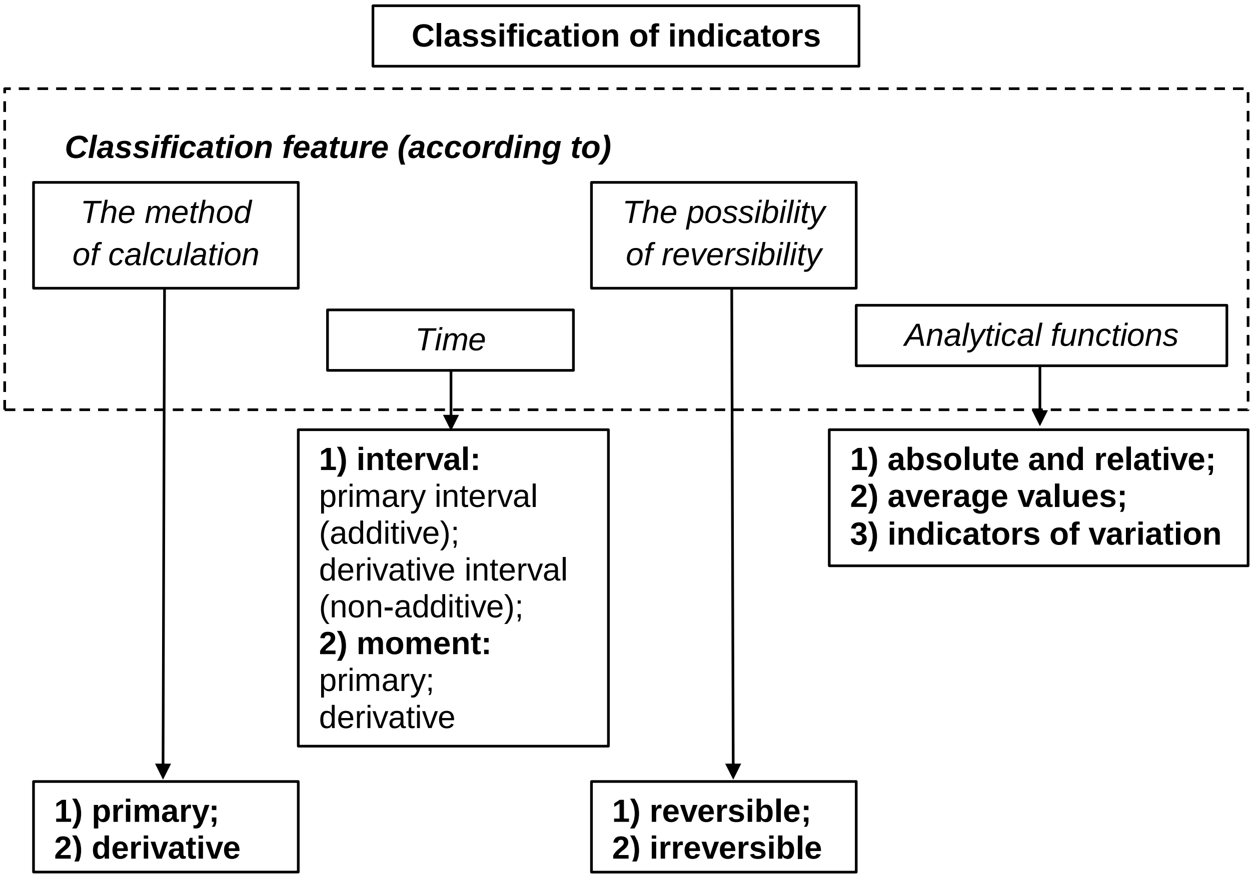

Classification of indicators
Classification feature (according to)
The method of calculation
The possibility of reversibility
Time
Analytical functions
1) interval: primary interval (additive); derivative interval (non-additive); 2) moment: primary; derivative
1) absolute and relative; 2) average values; 3) indicators of variation
1) primary; 2) derivative
1) reversible; 2) irreversible

Fig. 5.1. **Classification of indicators**

The method of calculation considers primary and derivative indicators. *Primary data* are determined by the summary of statistical observation data and presented in the form of absolute values (number and amount of savings, bank deposits). *Derivatives* are calculated on the basis of primary or derivative indicators. They are in the form of average or relative values (average wage, average wage index).

On a time basis, indicators are divided into *interval* and *moment*. Interval indicators characterize the phenomenon over a period of time (day, ten-day period, month, year). An example would be the volume of products produced, housing put into commission, fresh water consumption, etc. *Moment indicators* include indicators that give a quantitative description of the phenomena at a certain point in time: the area of vineyards and citrus plantations, the length of oil pipelines at the end of the year, etc.

Interval and moment indicators can be either *primary* or *derivative*. For example, the area of irrigated land is the primary moment indicator, and the proportion of such land in the total area is a derivative moment indicator;

electricity consumed in the industry is a primary interval indicator, while based on a unit of working time it is a derivative indicator.

Interval indicators depend on the length of time over which they are calculated. The peculiarity of primary interval indicators is *additivity*, that is, the possibility of summation. Derivatives are largely *non-additive*.

Among statistical indicators there are pairs of mutually *reversible indicators*, which in parallel characterize the same phenomenon and are divided on the basis of reversibility. The direct index x increases with the increase of the phenomenon, the inverse one 1/x, on the contrary, decreases. The following indicators can serve as an example:

the purchasing power of a monetary unit is a direct indicator, while the unit price of a product is a reverse one;

labor productivity per unit of time is a direct indicator, labor intensity of a unit of production is a reverse one, etc.

Depending on the *way of performing their functions* (analytical functions), indicators can show the volume of a phenomenon, its average level, the intensity of its manifestation, its structure, changes in time or space. In this regard, there are the following groups of indicators [10; 23; 34]:

absolute and relative;

average values;

indicators of variation.

Absolute indicators characterize the size of a population or the volume of a phenomenon under study within the specific limits of space and time, that is, show the level of development of a phenomenon, its size.

An absolute indicator can be obtained in one of two ways:

by counting of the units of a population having a specific value of a feature (for example: the number of transport enterprises in Kharkiv at a specific date, the number of industrial production personnel of an enterprise, etc.);

by summing the value of a feature over the entire statistical population.

Absolute indicators are always named numbers. Depending on the socio-economic nature of the phenomena studied, they are expressed in natural, value and labor units of measurement.

Natural measures are used when the units of measurement correspond to the consumer properties of the phenomena being studied (for example: the production of cars is measured in pieces, the production of steel is in tons, the yield is in quintals).

Natural units of measurement can be complex. Such units are used when one unit of measurement is not sufficient to characterize the phenomenon under study; then the product of two units is used. For example, electricity production is measured in kilowatt-hours, cargo turnover is measured in tonne-kilometers.

Natural units also include *conventionally natural units* of measurement. They are used when a product has several varieties, and the total volume can be only obtained on the basis of a consumer property common for all varieties.

To summarize, one of the varieties of a product is taken as a unit of measure, and the rest are converted to it by appropriate conversion factors.

Cost units of measurement allow you to give a monetary evaluation of the phenomena and processes under research. These meters are used to summarize data from the enterprise level to the national economy level and to estimate heterogeneous statistical populations. Units of value are used to measure the volume of output of an enterprise, the population, etc. The scheme for obtaining the total volume of a statistical feature in terms of value is as follows:

$$Q = \sum_{i=1}^{m} p_1 \times q_1 , \qquad (5.1)$$

where p_1 is the price (unit value of the feature);

q_1 is the volume of the feature in kind;

m is the number of features.

The values expressed in units of cost can be summed up and summarized, but changes in prices over time must be taken into account when using them. In order to eliminate this shortage of cost measures, one should use the "unchanged" or "comparable" prices of one period [6; 34; 94].

Labor units of measurement are used to estimate the total labor costs and the complexity of individual operations of a technological process. These include man-hours, man-days (estimation of working time), norm-minutes (estimation of complexity).

A classification of absolute indicators is given in Fig. 5.2.

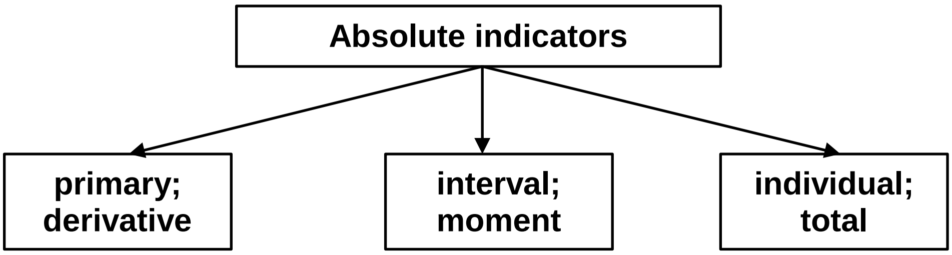


Fig. 5.2. **Classification of absolute indicators**

The entire set of absolute values includes both individual values (characterizing individual units of a population) and total ones (characterizing the totality of several units of the population or a significant feature from one or another part of the population).

Absolute indicators should also be divided into moment and interval. *Moment absolute indicators* characterize the fact of the existence of a phenomenon or process, its size (volume) at a particular date/time. *Interval absolute indicators* characterize the total volume of a phenomenon over a given period of time (e.g. annual or quarterly output), allowing further summation [23; 27; 34].

As such absolute indicators do not give a complete picture of the phenomenon under study, do not show its structure, development over time, the relationship between its parts: on their basis it is difficult to compare it with other similar phenomena. These analytical functions are performed by relative indicators.

5.2. Relative indicators

Although absolute indicators play an important role in the practical and cognitive activities of man, the analysis of facts necessarily leads to the need for various kinds of comparisons. And then the absolute indicators that characterize a particular phenomena under study are considered not only separately, but also in comparison with other indicators taken as a scale for evaluation or a comparison base.

A relative statistical indicator is a summative characteristic, expressed as a numerical measure of the ratio of two absolute values that are compared. Such an indicator is obtained by relating one absolute indicator to another. The logical scheme of calculation of a relative indicator is as follows:

$$\text{Relative value} = \frac{\text{Comparable value}}{\text{Comparison base}}. \qquad (5.2)$$

Depending on the values in the numerator or in the denominator, the relative values can be expressed in the following forms: coefficients (parts), percent (%), per mil ($^{o}/_{oo}$), per decimil ($^{o}/_{ooo}$), when 1, 100, 1 000, 10 000 units respectively are taken as a base. Per mil is widely used in demographic statistics to characterize birth rates, mortality rates and other demographic processes. Per decimil is used to evaluate the provision of hospital beds, places in institutions of higher education.

It is worth noting that a specific unit of measure can be attributed to dimensionless relative indicators. For example, indicators of the natural movement of the population – fertility rate, mortality rate are calculated in ppm, but show the number of births or deaths per year per 1,000 people.

The diversity of relationships in real life requires the use of relative indicators in terms of content and statistical nature. Depending on their functions, in the process of analysis, relative values are classified into types (Fig. 5.3).

Consider the content of each of the selected types of relative indicators.

1. *The relative indicator of dynamics*. Dynamics in statistics means changes in a socio-economic phenomenon or process over time. The relative indicator of dynamics (RID) characterizes the direction and intensity of change in the indicator over time; it is determined by the ratio of its value either over two periods or at times.

The base of comparison may be the previous variable level (the calculation is done by a chain method); constant, time removed level (calculation is done in the basic way). RID is called *growth rate* [18; 34].

Example 5.1. In Ukraine, in 2015, the turnover (in million UAH) was 544 153; in 2016 it was 720 731; in 2017 it amounted to 949 864. Let's calculate the growth rate of this indicator:

1) the chain way:

$$\text{RID} = \frac{\text{the indicator of the current year}}{\text{the indicator of the previous year}}. \qquad (5.3)$$

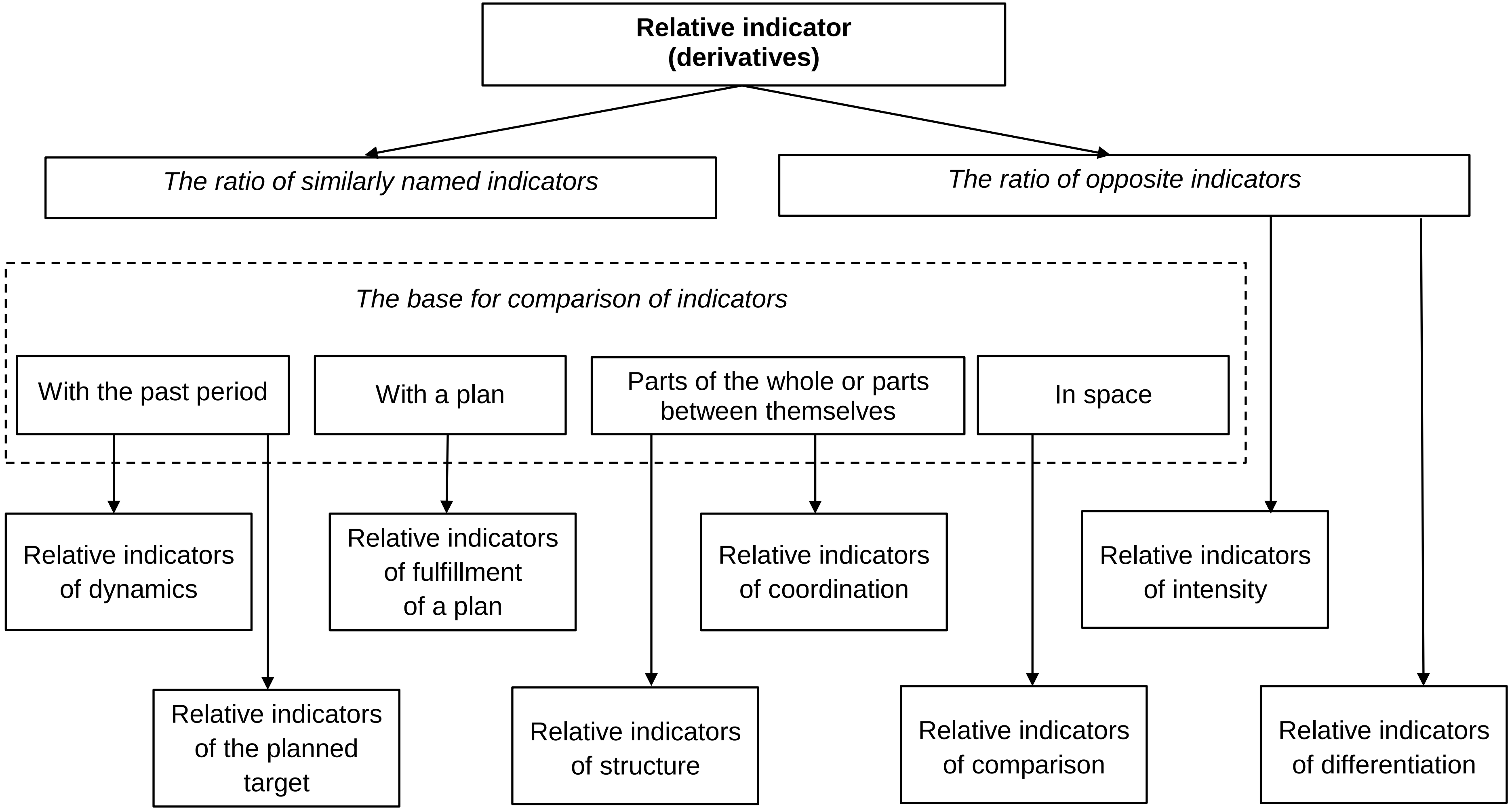

Relative indicator (derivatives)
The ratio of similarly named indicators
The ratio of opposite indicators
The base for comparison of indicators
With the past period
With a plan
Parts of the whole or parts between themselves
In space
Relative indicators of dynamics
Relative indicators of fulfillment of a plan
Relative indicators of coordination
Relative indicators of intensity
Relative indicators of the planned target
Relative indicators of structure
Relative indicators of comparison
Relative indicators of differentiation

Fig. 5.3. **Classification of relative values**

In 2016 the growth rate of the country's turnover was: $\frac{720\ 731}{544\ 153} = 1.33$ (or 133 %); in 2017 it was $\frac{949\ 864}{720\ 731} = 1.32$ (or 132 %). In 2018 the growth rate of the country's trade turnover increased by 33 % compared to 2015, and in 2017 the growth rate of the country's trade turnover increased by 32 % compared to 2016. There is a decrease in the growth rate of the analyzed indicator;

2) the base way:

$$\text{RID} = \frac{\text{the current year indicator}}{\text{the base year indicator}}. \tag{5.4}$$

We take 2015 as the base for comparison. Then the growth rate of the country's trade turnover in 2016 will coincide with the growth rate calculated by the chain method and it will make 32 %; the growth rate of the country's trade turnover in 2017 will be $\frac{949\ 864}{544\ 153} = 1.75$. That is, there is a 75 % increase in the country's turnover growth in 2017 compared to 2015.

2. *Relative indicators of structure* (RIS) characterize the shares of parts of a population in its total volume. They show the structure of a population. The calculation of relative indicators of structure consists in the calculation of shares of individual parts in the totality:

$$\text{RIS} = \frac{\text{the indicator of a part of the total}}{\text{the indicator of the total}}. \tag{5.5}$$

The RIS is usually expressed in the form of coefficients or percentages, the sum of the coefficients should be 1 and the sum of the percentages should be 100 because the shares are reduced to a common base. The relative indicators of structure are used to study the complex phenomena that break up into parts, for example, to study the composition of the population on different grounds (age, education, nationality, etc.). An example of such a set is given in Table 5.1.

Table 5.1

The structure of retail turnover of a city

Indicators	Amount, billion UAH	%
Retail turnover, total:	83.1	100
including: organization of retail trade	42.4	51
non-trading organizations	20.8	25
individuals in the markets	19.9	24

3. *Relative indicators of coordination* (RIC) characterize the ratio of parts of a statistical population to one of them taken as a benchmark. RIC show the number of times one part of the population is larger than another, or the number of units of one part of the population per 1, 10, 100, etc. of units of the other part.

The base for comparison is the part that has the highest share or the priority in the population. Thus, taking into account the turnover of retail organizations in the previous example, it is possible to calculate the RIC for non-trade organizations: RIC = 25/51 = 0.49, which means there are 49 kopiykas of non-trade organizations' trade turnover per every hryvnia of the retail trade turnover.

Relative indicators of coordination play an important part in economic analysis, because with their help the existing relationships are formed more clearly and vividly.

4. *Relative indicators of the planned target* (RIPT) and r*elative indicators of fulfillment of the plan (RIFP)* are used by all financial and economic entities carrying out current and strategic planning. These indicators are calculated as follows:

$$\text{RIPT} = \frac{\text{the current year indicator}}{\text{the previous year actual indicator}}; \qquad (5.6)$$

$$\text{RIFP} = \frac{\text{the current period actual indicator}}{\text{the current period planned indicator}}. \qquad (5.7)$$

The relative indicator of a planned target characterizes the intensity of the planned task, while the *relative indicator of fulfillment of the plan* – the degree of implementation.

Example 5.2. The calculation of RIPT and RIFP: the actual turnover of a firm in 2017 amounted to 22 million UAH; market analysis showed that in 2018 it was realistic to bring turnover to 2.6 million UAH; the actual turnover amounted to 2.5 million UAH.

$$\text{RIPT} = \frac{2.6}{2.2} = 1.18\,;$$

$$\text{RIFP} = \frac{2.5}{2.6} = 0.96\,.$$

The calculations show that the planned target for 2018 is 1.3 times higher than the actual level of 2017, but the plan for 2018 was fulfilled only by 96 %.

The relative indicator of dynamics (RID), calculated in a chain way, the relative indicator of the planned target (RIPT) and the relative indicator of fulfillment of the plan (RIFP) are related to each other as:

$$\begin{aligned} &\text{RID} = \text{RIPT} \times \text{RIFP} = \\ &= \frac{\text{the current year indicator}}{\text{the actual indicator of the previous year}} \times \\ &\times \frac{\text{the actual indicator of the current period}}{\text{the planned indicator of the current period}} = \\ &= \frac{\text{the actual indicator of the current period}}{\text{the actual figure of the previous year}}. \end{aligned} \tag{5.8}$$

5. *Relative indicators of comparison* (RICom) characterize the comparative dimensions of absolute indicators that belong to different objects or territories, but for the same period of time. They are calculated as fractions of the distribution of the same absolute indicators characterizing different objects related to the same period or time.

$$\text{RICom} = \frac{\text{the indicator characterizing object A}}{\text{the indicator characterizing object B}}. \tag{5.9}$$

Using benchmarking, you can compare labor productivity in different countries and determine where and how many times it is higher; compare prices for different products, economic performance of different businesses, etc.

Relative indicators of comparison are classified as follows: relative spatial comparison indicators; relative indicators of comparison with a standard.

Relative spatial comparison indicators (RSCI) show the ratio of the dimensions or levels of spatial indicators of different territories or objects. Most often, they are used to compare economic development or living standards of countries and regions. In this case, any object can be the base of comparison, but the method of calculation of indicators should be uniform [10; 34; 94].

Relative indicators of comparison with a standard compare the actual values of indicators to a *benchmark – a standard*, an optimal level, an industry standard, etc. For example, they can be used to determine the deviation of the current liquidity ratio of an enterprise from the normative values [10; 34; 94].

6. *Relative indicators of intensity* and level of economic development (RII) characterize the degree of distribution or the level of development of the studied phenomena and processes in a particular environment; they are formed as a result of comparing different, but in some way related quantities. These indicators are calculated as follows:

$$\text{RII} = \frac{\text{Indicator characterizing the phenomenon}}{\text{Indicator characterizing the environment of the phenomenon distribution}}. \quad (5.10)$$

RIIs are calculated per 100, 1,000, 10,000, etc. units of the studied population, and are used when it is impossible to determine the magnitude of the spread of the phenomenon by the absolute value. Thus, for the study of demographic processes, the rate of birth, mortality, natural increase (decrease) of the population is calculated as the ratio of the number of births (deaths) or the magnitude of annual natural increase to the average annual population in a certain territory per 1 000 or 10 000 people; medical assistance is calculated per 10 000 people; morbidity or crime rate – per 100,000 people.

Example 5.3. As of January 1, 2019, 1 596 babies were born in Kharkiv region, and 574 in Sumy region. Comparison of absolute indicators doesn't make it possible to evaluate the birth rate and determine where it is higher.

This can be done through RII – birth rates in Kharkiv and Sumy. In January 2019 the population of Kharkiv region was 2 674.6 thousand people, in Sumy region it was 1 080.7.

$$\text{RII}_{\text{Kharkiv}} = \frac{1596}{2674.6} = 0.59 \text{ people/thousand people};$$

$$\text{RII}_{\text{Sumy}} = \frac{574}{1080.7} = 0.53 \text{ people/thousand people}.$$

Conclusion: birth rates in Kharkiv are higher than in Sumy.

7. R*elative indicators of differentiation* (RIdif) are calculated by comparing two structural series, one of which characterizes the ratio of a population in terms of the number of units, and the other in terms of the magnitude of some feature. For example, comparing the share of farms as to the number and share of gross production, fixed assets, employees, etc. [27; 34].

Relative indicators are important in practice, but they cannot be viewed in isolation from the absolute indicators by means of which they are calculated. Otherwise, you can come to wrong conclusions. Thus, only the combined use of absolute and relative indicators makes it possible to perform a qualitative analysis of various phenomena of socio-economic life.

5.3. Average indicators

One of the main tasks of statistical research is to identify patterns of mass phenomena. Patterns can be detected only through generalization of homogeneous phenomena and generalized characteristics of the units of a phenomenon. Therefore, in order to measure a feature that is characteristic of the entire population of units being studied, an average value is calculated.

Significant contribution to the substantiation and development of the theory of averages was made by A. Quetelet (1796 – 1874), a member of the Belgian Academy of Sciences, corresponding member of St. Petersburg Academy of Sciences. When an average value for many units is calculated, the influence of random causes is neutralized. An average value, abstracting from individual characteristics of individual units of a population, expresses the common properties inherent in all units. A. Quetelet emphasized that statistical averages are not just a method of mathematical measurement, but a category of objective reality.

A fundamental essence of statistical cognition lies in the elimination of random phenomena resulting from individual causes, and in the identification of patterns conditioned by common causes [34; 94].

The possibility of switching from particulars to generals, from the random to the regular explains the importance and wide application of the method of averages in statistical surveys.

The conditions of scientific use of mean values are as follows:

- the population should consist of qualitatively homogeneous units;
- the method of averages should be combined with the grouping method;
- the population should be fairly large;
- there should be a possibility of the use of system averages [34].

Thus, the mean value is a generic characteristic of a feature under study, which shows its typical level per unit of population under specific conditions of place and time.

Mean values are used to evaluate the achieved level of the studied indicator; to analyze and plan the production and economic activity of enterprises (associations), firms, banks and other economic units; to identify the relationship of phenomena; to forecast and calculate standards.

The correctness of the use of mean values in analytical practice is first of all conditioned by the availability of a qualitatively homogeneous population, based on which the mean is calculated.

Before calculating the mean, it is necessary to group the units of the population under study by selecting qualitatively homogeneous groups.

The mean calculated for the population as a whole is the *overall mean value*; and the mean value calculated for each group is the *group mean*. The overall mean represents the general features of the phenomenon under study; the group mean characterizes the magnitude of the phenomenon that occurs in the specific conditions of the group.

In addition, the *mean* is *always named*, i.e. it has the same dimension (unit of measure) as the feature in the individual units of the population.

In socio-economic analysis, two classes of mean values are used: a *power mean and a structural mean.*

Power mean values include several types of values, built on one general principle:

$$\overline{x} = \sqrt[k]{\frac{\sum_{i}^{n} x_i^k}{n}}, \qquad (5.11)$$

where x_i is an option;

n is the size of the statistical population;

k is the indicator of the power on which the type of the mean depends.

The combination of individual values of a feature with defining properties of a population determines the type of the power mean.

There are several values of the k exponent used in practice: the harmonic mean for k = −1; the geometric mean for k = 0; the arithmetic mean for k = 1; the mean square for k = 2.

Depending on the form of presentation of the input data, *power mean values* can be *simple* and *weighted*.

If the input data are given as a simple list of the feature values in statistical units, the *simple power mean* formula is used:

$$\overline{x} = \sqrt[k]{\frac{\sum_{i}^{n} x_i^k}{n}}.$$

If the data are pre-grouped (given as a distribution series), the *power weighted mean* formula is used:

$$\overline{x} = \sqrt[k]{\frac{\sum_{i=1}^{m} x_i^k f_i}{\sum_{i=1}^{m} f_i}},$$

where f_i is the repetition rate of individual feature values;

m is the number of homogeneous groups.

The types of mean values should be highlighted in detail.

1. *The arithmetic mean* is the most common type of power mean values. It is used when the volume of the averaged feature is an additive value (formed as a sum of the individual feature values). If the individual values

of a feature in statistical units are replaced by the arithmetic mean, the total volume of the feature in the population will remain unchanged.

The *arithmetic mean* (*simple)*, is used to work with grouped data and is calculated by the formula:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}. \qquad (5.12)$$

If the values of the averaged feature are repeated in the original data, the mean will be calculated based on the grouped data or variance series. In this case, the number of units in the variants of the feature is not the same; it is expressed in the form of weight. In this case, it is necessary to use the arithmetic weighted mean (the mean of the grouped values) for the calculation.

$$\overline{x} = \frac{\sum_{i=1}^{m} x_i f_i}{\sum_{i-1}^{m} f_i}, \text{ or } x = \sum_{i=1}^{m} x_i d_i, \qquad (5.13)$$

where $d = \frac{f_i}{\sum_{i=1}^{m} f_i}$ is the frequency, that is, the proportion of statistical units that have a certain value of the feature in the total volume of the population.

The mentioned formulas for calculating the arithmetic mean are used for both discrete and interval series. Moreover, the interval series must be converted into a discrete one by finding the value of the middle of each interval (center). If there are intervals with indistinctly marked boundaries (lower or upper intervals are open), in order to determine the mean value, the intervals must be closed: for the first interval, the step size of the second interval is taken, and for the last one, the step size is the same as in the previous interval.

The arithmetic mean has a number of useful *properties*, the most important of which include the following:

1) the arithmetic mean of the constant value is equal to this value:

$$A = A \text{ for } A = \text{const};$$

2) the algebraic sum of the deviations from their arithmetic mean is zero:

$$\Sigma(x_i - \bar{x}) = \Sigma x - n\bar{x} = 0 \text{ for a simple mean;}$$

$$\Sigma(x - \bar{x})f = \Sigma xf - \bar{x}\,\Sigma f = 0 \text{ for a weighted mean;}$$

3) the sum of squares of deviations from the mean will always be less than the sum of squares of deviations from any other value:

$$\Sigma(x - \bar{x})^2 f < (x - A)^2 f\,;$$

4) if all variants are reduced (increased) by a constant A number, their arithmetic mean will decrease (increase) by the same number:

$$\frac{\Sigma(x_i \pm A) \times f_i}{\Sigma f_i} = \bar{x} \pm A\,;$$

5) if all variants are equally increased (decreased) by the same number of times, the arithmetic mean will increase (decrease) by the same number of times:

$$\frac{\Sigma(x_i \times A) \times f_i}{\Sigma f_i} = \bar{x} \times A\,;$$

6) if all the weights of the mean are equally increased (decreased) several times, the arithmetic mean will not change:

$$\frac{\Sigma x_i(\frac{f_i}{A})}{\Sigma\frac{f_i}{A}} = \bar{x}\,;$$

7) the multiplication of mean values by the sum of frequencies is equal to the sum of the multiplications of variants by frequencies:

$$\bar{x}\,\Sigma f = \Sigma xf\,;$$

8) the total mean equals the mean of the partial mean values, weighted by the number of the respective groups of the population:

$$\overline{X}_{total} = \frac{\overline{x}f_1 + \overline{x}_2 f_2 + ... + \overline{x}_n f_n}{f_1 + f_2 + ... + f_n} = \frac{\Sigma \overline{x}_i f_i}{\Sigma f_i}.$$

$$\text{If } d_f = \frac{f}{\sum f}, \quad \overline{X}_{total} = \frac{\Sigma \overline{x}_i d_f}{\Sigma d_f}.$$

2. *The harmonic mean* has a more complex construction than the arithmetic mean. It is used in cases where the statistical information does not contain frequencies with respect to individual values of the feature but it is given by multiplying the value of the feature by frequency [10; 34].

$$\overline{x} = \sqrt[-1]{\frac{\sum_{i=1}^{n} \left| x_i^{-1} \right|}{n}}. \tag{5.14}$$

Depending on the form of presentation of the input data, the harmonic mean can be calculated as *simple* and *weighted.*

If the input data are not grouped, the *simple harmonic mean is used:*

$$\overline{x} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}. \tag{5.15}$$

With grouped data the harmonic weighted mean is used:

$$\overline{x} = \frac{\sum_{i=1}^{m} M_i}{\sum_{i=1}^{m} \frac{M_i}{x_i}}, \tag{5.16}$$

where M_i is the statistical weight; $M_i = x_i \times f_i$.

The geometric mean is used in cases where the total volume of the averaged feature is a multiplicative value (when the defining property is formed as the product of the individual feature values) [6; 10].

$$\overline{x}_{geom} = \sqrt[n]{x_1 \times x_2 \times ... \times x_n} = \sqrt[n]{\Pi_{i=1}^{n} x_i}. \tag{5.17}$$

The *geometric weighted mean* is not used in practical calculations. In socio-economic studies, the geometric mean is used in the time series analysis to determine the average growth rate when a sequence of relative values of dynamics is given.

The geometric mean can be used to determine the value equidistant from the maximum and minimum values of a feature.

The quadratic mean is used in the case of replacement of individual values of a feature by the mean value, it is necessary to keep the sum of the squares of the original values unchanged. It is mainly used to measure the degree of fluctuation of individual values of a feature relative to the arithmetic mean, that is standard deviation. In addition, the quadratic mean is used when it is necessary to calculate the mean of a feature expressed in square or cubic units of measurement (in the calculation of the average size of square sections, average diameters of pipes, trunks, etc.).

The quadratic mean can be:

$$\text{simple: } \bar{x}_{sq} = \sqrt{\frac{\sum\limits_{i=1}^{n} x_i^2}{n}}; \tag{5.18}$$

$$\text{weighed: } \bar{x}_{sq} = \sqrt{\frac{\sum\limits_{i=1}^{m} x_i^2 \times f_i}{\sum\limits_{i=1}^{m} f_i}}. \tag{5.19}$$

All power mean values differ in degree values: the higher is the degree, the greater is the quantitative value of the mean:

$$\bar{x}_{harm} \le \bar{x}_{geom} \le \bar{x}_{arith} \le \bar{x}_{sq}. \tag{5.20}$$

This dependence of power averages is called the property of the mean majorities.

In order to *summarize these series of data,* the average level of the time series is calculated, using the *chronological mean* for the moment series with equal intervals of time between the dates:

$$\bar{x} = \frac{\frac{1}{2}x_1 + x_2 + ... + \frac{1}{2}x_n}{n-1}. \tag{5.21}$$

In the case where the average feature is the relationship between logically related quantities (e.g. relative intensity, structure, etc.), the question arises as to the choice of the mean.

The *basic methodological principle* of choosing the *type of mean* is to ensure the *logical and meaningful nature* of the indicator being studied.

Formally, this principle can be written as shown in Table 5.2.

Table 5.2

The basic methodological principles for the choice of the mean type

Indicators	Direct	Reverse
Primary	Simple arithmetic	Simple harmonic
Derivative	Weighed	Weighed harmonic

The mean values can also be calculated when the individual variant values are not recorded but only the results are known. Thus, data on fuel consumption per each kilowatt-hour of electricity, yield on each hectare of acreage of a particular crop, etc. are not calculated.

Important concepts

Absolute indicators are indicators that characterize the size of the population, or the volume of the phenomenon under study within specific limits of space and time, that is, show the level of development of a phenomenon, its size.

The relative indicator of dynamics (RID) is an indicator that characterizes the direction and intensity of change over time; it is determined by the ratio of its value either over two periods or at times.

The relative indicator of a planned target (RIPT) characterizes the intensity of the planned task, and the *relative indicator of the plan fulfillment* (RIPF) shows the degree of implementation of the plan.

The relative statistical indicator is a summative characteristic, expressed as a numerical measure of the ratio of two absolute values that are compared.

Relative indicators of differentiation (RIdif) are indicators that are calculated by comparing two structural series, one of which characterizes the ratio of the population as to the number of units, and the other in terms of the magnitude of some feature.

Relative indicators of coordination (RIC) are indicators that characterize the ratio of parts of a statistical population data to one taken as the comparison base; they show the number of times one part of the population is larger than the other, or the number of units of one part of the population per 1, 10, 100, etc. units of the other part.

Relative indicators of comparison (RICm) characterize the comparative dimensions of homogeneous absolute indicators that belong to different objects or territories, but for the same period of time.

Relative indicators of structure (RIS) characterize the shares of parts of the population in its total volume.

The mean value is a generalizable characteristic of the feature under study in the population under study which shows its typical level per unit of the population in specific conditions of place and time.

A statistical indicator is a measure, that is the unity of qualitative and quantitative representation of a particular property of a socio-economic phenomenon or process.

Typical tasks

Task 1. It is necessary to calculate the absolute indicators according to the following data: in January 2017, the following types of canned fish were produced in Ukraine: 100 thousand cans of 240 cm^3 capacity; 250 thousand cans of 160 cm^3 capacity; 75 thousand cans of 480 cm^3 capacity. Transfer the products to conventional units if the capacity of a conventional can is 353.4 cm^3.

The solution.

1. Let's determine the coefficients of conversion:

$$\frac{240}{353.4}=0.679,\ \frac{160}{353.4}=0.453,\ \frac{480}{353.4}=1.358.$$

2. Let's calculate the total production of canned fish:

$$100\times0.679+250\times0.453+75\times1.358=283 \text{ thousand conventional cans.}$$

Thus, in January 2017, Ukraine produced 283,000 conventional cans of canned fish.

Task 2. Determine the relative values of structure and coordination based on the data given in Table 5.3.

Table 5.3

The input data

Indicators	mln UAH
Foreign trade, total	765.5
Export	356.2
Import	409.3

Justify the results.

The solution.

The relative size of the structure is the ratio of the part to the whole, or the proportion of the part of the units in the total volume of the population (Table 5.4).

Table 5.4

The estimated data

Indicators	%
Foreign trade, total	100.0
Export	46.5
Import	53.5

The relative value of coordination characterizes the ratio of different structural units of the same population.

Thus, based on the above data, we can calculate that each million hryvnias of exports accounted for 1.15 million UAH of imports (409.3 / 356.2).

Task 3. We have the data on the diameter of three pipes: 20, 40, 60 cm. It is necessary to determine the average size of the diameter of the pipes.

The solution.

Since the output is given as square functions, the average diameter of the pipe is determined by the formula of the mean square.

$$\bar{x} = \sqrt{\frac{x^2}{n}} = \sqrt{\frac{20^2 + 40^2 + 60^2}{3}} = \sqrt{\frac{400 + 1600 + 3600}{3}} = \sqrt{\frac{5600}{3}} = \sqrt{1866.7} = 43.21 \text{ cm.}$$

Therefore, the average diameter of the three pipes is 43.21 cm.

Task 4. We have some data from the staffing report of a city industrial enterprise (Table 5.5).

Table 5.5

Reporting data

Quarters	Average number of workers, people
1st	1 000
2nd	1 050
3rd	1 295
4th	1 820

It is necessary to determine the average rate of increase of the number of workers per quarter.

The solution.

To determine the average growth rate of the number of workers, we calculate the quarterly growth rates as the ratio of each successive level to the previous one. We get the chain growth factors:

The 2nd quarter is compared to the 1st quarter: $K_1 = \frac{1050}{1000} = 1.05;$

The 3rd quarter is compared to the 2nd quarter: $K_2 = \frac{1295}{1050} = 1.23;$

The 4th quarter is compared to the 3rd quarter: $K_3 = \frac{1820}{1295} = 1.41.$

In this case, the total volume of the phenomenon is defined as a product of growth factors. Thus, the number of workers in the whole period (year) increased by 1.82 times (1.05×1.23×1.41).

The average quarterly rate of growth of the number of workers is calculated by the formula of the geometric mean, which is used in cases where the total volume of a phenomenon is not the sum but the product of the values of the features:

$$\overline{K} = \sqrt[3]{K_1 \times K_2 \times K_3} = \sqrt[3]{1.05 \times 1.23 \times 1.41} = \sqrt[3]{1.821} = 1.221.$$

Thus, on average, the number of workers in each quarter increased by 1.221 times or by 22.1 %.

This result can be obtained by another formula for calculating the average coefficients:

$$\overline{K} = \sqrt[n-1]{\frac{x_n}{x_1}},$$

where n is the number of periods;

x_1 and x_n are the initial and final levels of the time series.

$$\overline{K} = \sqrt[n\ 1]{\frac{x_n}{x_1}} = \sqrt[3]{\frac{1820}{1000}} = 1.221.$$

We get the same result.

Reference to laboratory work

The guidelines for carrying out laboratory work on the topic "Acquisition of skills in the calculation of relative and average indicators in MS Excel" are presented in [100]. The laboratory work aims to familiarize students with the types of relative and average values and to help them get practical skills in the calculation of these values by means of MS Excel.

Questions for self-assessment

1. What is the essence of statistical indicators, what are their functions and types?

2. Describe the conditions for using absolute and relative values.

3. What is the practical application of relative and mean values?

4. Describe the problem of harmonization and comparability of indicators in statistics.

5. Expand on the principle of using mathematical properties of mean values to simplify calculation of these values.

6. Specify the features of presentation of absolute statistical values.

7. Name the types of relative values according to their analytical function.

8. Outline the conditions for the scientific application of mean values.

9. Provide comparison of measurement units of absolute and relative values.

10. What analytical functions do relative values perform? Is it possible to compare different indicators?

Questions for critical rethinking (essays)

1. Explain the nature of statistical indicators and their role in the statistical analysis.

2. What are the principles of summary statistics that match the use of relevant indicators and methods for determining them with respect to real socio-economic phenomena and processes.

3. What is the importance of calculating the relative values of coordination, comparison, intensity for conducting a statistical survey of socio-economic processes?

4. What is the relationship between the relative values of the planned task, the execution of the plan, the dynamics? Give examples.

5. What types of averages are most commonly used in statistical analysis? What is the criterion for choosing the type of mean?

6. Analysis of distribution series

Basic questions:

6.1. The elements of distribution series.

6.2. The characteristics of the center of distribution.

6.3. The measures of variation.

6.1. The elements of distribution series

The results of statistical summaries and groupings can be presented in the form of *statistical series* – ordered sets of values of the analyzed indicator, that is, a statistical feature. According to the content, statistical series are divided into time series and series of distribution.

A *time series* is a systematized set of numerical data that characterize changes in the studied phenomena over time [34].

A *series of distribution* is a systematic sequence of statistical units grouped based on a specific feature. It characterizes the composition of the phenomenon under study, allows you to evaluate the homogeneity of the totality, patterns of distribution of statistical units. Usually, the distribution series is the result of *structural grouping* [34].

A distribution series is considered to be built when it is known how the feature values change together and how often individual feature values occur.

Series of distributions of different types are constructed for different statistical feature:

• *attributive* series are built on descriptive grounds in the order of increase or decrease of the observed values of a feature. Examples of attributive series are: population – according to nationality, profession, gender; enterprises – according to ownership;

• *variation* series are based on quantitative features (for example the distribution of workers depending on the skill level, wages, students – on a success base.

Variation series are divided into discrete and interval series.

In *discrete series* the feature takes only integer values (for example, family size, tariff category).

Interval series are based on continuous features that have any, including fractional, values. Depending on what structural grouping is the basis of the interval series, the intervals are divided into equal intervals and unequal intervals. In evenly spaced series, the width of the interval is constant, in unequal interval series it is different for different groups [34].

The main *elements of a distribution series* are:

1) the value of a feature (options):

- x_i, a discrete interval in discrete series;
- $x_i^L - x_i^U$, an interval for interval series, where i = 1 ÷ n_i is frequency;

2) frequency f_i which is the number of units of a population that have a certain value of a feature. Frequency shows the number of times this feature value is encountered in the population. The sum of all frequencies is always equal to the size of the statistical population, i.e. $\Sigma f_i = N$.

The study of distribution series is carried out in two stages [34; 80]:

• empirical research aiming to get generalized characteristics of the population under study;

• a theoretical study which aims to identify the pattern of distribution under consideration and its theoretical description.

An empirical study begins with determining the frequency characteristics of a series of distributions.

The frequency characteristics of the distribution series should be considered more carefully.

The output frequency response of any distribution is the frequency n_i. Based on it, the following characteristics can be calculated:

• *frequence,* the proportion (share) of the population units that have a certain value of a feature, that is, the frequency expressed as a relative value (unit or percentage share):

$$d_i = \frac{f_i}{N}, \quad \sum_{i=1}^{m} d = 1, \tag{6.1}$$

where $i = \overline{1, m}$.

This characteristic is important for the study of series of distributions, because it helps to link the indices of series of distributions with corresponding indicators and the terms and concepts of probability theory. In the theory of probability, q_i is the probability that this value of a feature will occur in the population. Frequence is used to compare the distribution series containing equal number of statistical units;

• *accumulated frequency* is the number of units of a population in which the feature value does not exceed a given x*, that is, the frequency with increasing totals:

$$\sum_{i=1}^{m^*} f_i = S_{x^*}; \quad S_{x^*} = N_{x^*}, \tag{6.2}$$

where x* is the feature value in the i-th group for which the cumulative frequency is calculated.

Based on cumulative frequencies, you can build a *cumulative distribution series* – a set of values of a number of units of a population with the feature values smaller and equal to the upper limit of the corresponding interval;

• *accumulated frequence* is the proportion of units in which the feature value does not exceed the given x*, that is, the frequence with increasing totals:

$$Q_{x^*} = \sum_{i=1}^{m} f_i = 1; \tag{6.3}$$

• *distribution density* is a universal frequency characteristic that allows you to proceed from empirical to theoretical distribution. For series with irregular intervals, only this characteristic gives a correct idea of the nature of the distribution. It is calculated in two ways:

as an the absolute density of distribution φ_i, showing the number of units of the population per unit width of the feature value interval:

$$\varphi_i = \frac{f_i}{h_i}; \tag{6.4}$$

as a relative density of distribution φ_i', showing the specific share of units of the population per unit width of the interval:

$$\varphi_i' = \frac{d_i}{h_i}. \tag{6.5}$$

The density of the distribution ensures the comparability of different types of distribution. Different distribution series are characterized by different sets of frequency characteristics: minimum – for attributive series (frequency f_i, and frequence d_i); for discrete series four characteristics (frequency n_i, frequence q_i, accumulated frequency S_i, accumulated frequence Q_i) are used; for interval series all five characteristics (frequency f_i, frequence q_i, accumulated frequency S_i, accumulated frequency Q_i, absolute φ_i and relative φ_i' density of the distribution) are used.

Graphs are a visual form of displaying series of distribution. Line graphs and plane diagrams constructed in a rectangular coordinate system are used to represent series.

For graphical representation of *attributive series* of distribution, different diagrams are used: column, linear, circular, area, sector and so on.

For *discrete variation series*, the graph is a distribution polygon.

A *distribution polygon* is a polyline connecting the points with the coordinates (x_i; f_i) or (x_i; d_i), where x_i is the discrete value of the feature, f_i is frequency, d_i is frequence [80].

The chart is built on an accepted scale. The layout of the distribution polygon is shown in the Fig. 6.1.

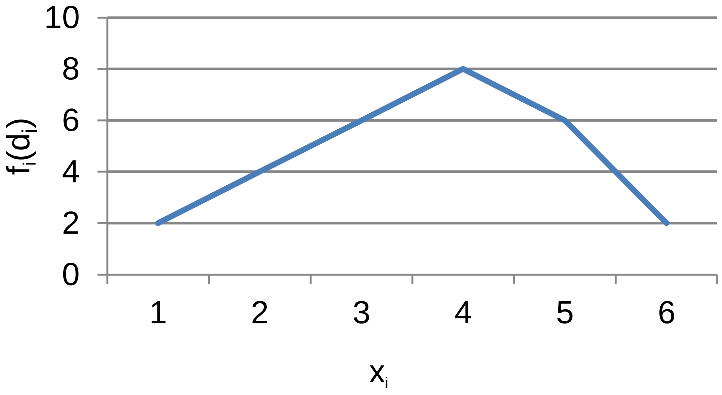


Fig. 6.1. **The polygon of distribution**

Histograms are stepped figures consisting of rectangles used to display interval variation series. The bases of these rectangles are equal to the width of the interval h_i, while the height equals the frequency f_i (frequence d_i) of the equal-interval series or the density of distribution of nonequal-interval series φ_i, φ_i'. The construction of a diagram is similar to the construction of a bar chart. The general view of the histogram is shown in Fig. 6.2.

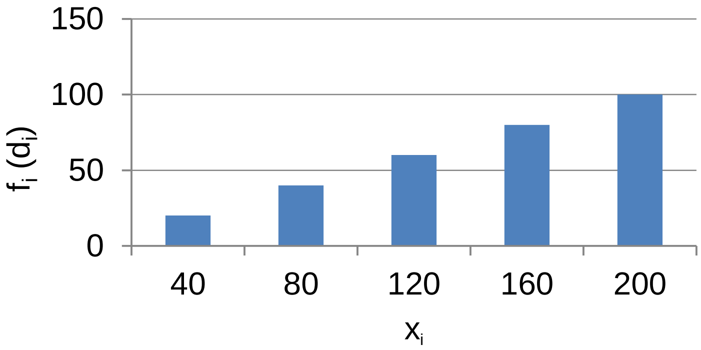


Fig. 6.2. **The distribution histogram**

You can construct a distribution polygon in the form of a histogram if you take the middle of the selected intervals as corresponding points.

For graphical representation of variational series it is possible to use the *cumulative* curve, composed of the accumulated frequencies (frequences). Accumulated frequencies are applied in the form of ordinates, combining the vertices of individual ordinates with segments of straight lines. A broken line is created in such a way that it has a non-descending appearance.

The coordinates of the points on the graph are {x_i; N_i} for a discrete series; {x_i^U; N_i} for an interval series. The starting point of the graph has the coordinates (x_1^L; 0); the highest point is (x_m^U; N). The general view of a cumulative curve is given in Fig. 6.3. The use of cumulative curves is particularly convenient for comparing variations [80; 11].

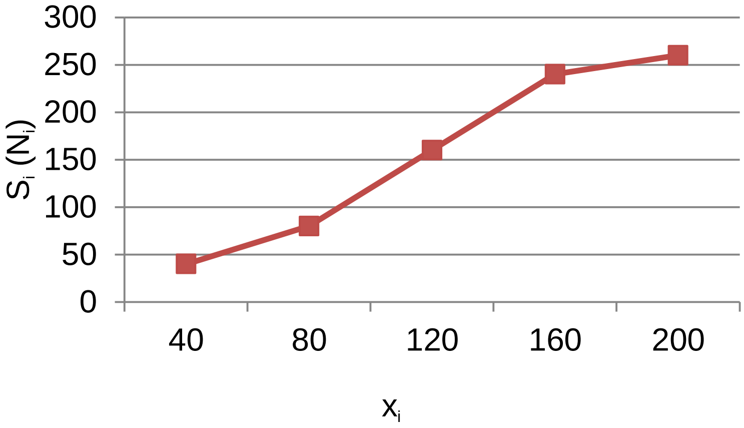


Fig. 6.3. **The graph of a cumulative curve**

For plotting a series of distributions, the ratio of scales on the abscissa axis and the ordinate axis is of great importance. In this case, it is necessary to follow the rule of the golden section, according to which the height of the graph should be approximately twice less than its base.

Analysis of patterns of distribution involves assessing the degree of uniformity of the population, asymmetry and excess of distribution.

The *uniformity* of the population is a prerequisite for using other statistical methods (mean values, regression analysis, etc.). *Homogeneous* populations are those, whose elements have common properties and belong to one type, class. At the same time, homogeneity does not mean complete identity of the elements' properties but only having something in common in what is essential. The criterion for the homogeneity of a population is the quadratic coefficient of variation, which due to the properties in the symmetric distribution is V_σ = 33 %.

Establishing asymmetry and kurtosis helps to establish the symmetry of the distribution of a random variable.

Finding out the general nature of the distribution involves not only evaluating the degree of its homogeneity, but also investigating the shape of the distribution, that is, evaluating symmetry and excess.

6.2. The characteristics of the center of distribution

An empirical study of a series of distributions implies calculation and analysis of the following **groups of indicators**:

• indicators of the position of the distribution center;

• indicators of the degree of homogeneity of the distribution series;

• indicators of the form of distribution.

The indicators of the position of the distribution center include the statistical mean in the form of the arithmetic and structural mean – the mode and the median.

The *arithmetic mean* for a discrete series of distribution is calculated by the formula [34; 80]:

$$\bar{x} = \sqrt[k]{\frac{\sum_{i=1}^{m} x_i^k f_i}{\sum_{i=1}^{m} f_i}}, \qquad (6.6)$$

where x_i is the variants of the feature values;

f_i is the frequency of the feature values;

m is the number of homogeneous groups.

For an interval variation series, the arithmetic mean is determined by the formula:

$$\bar{x} = \frac{\sum_{i=1}^{m} x_i f_i}{\sum_{i=1}^{m} f_i}, \qquad (6.7)$$

where x_i is the middle of the corresponding interval.

Unlike the arithmetic mean, calculated on the basis of all variants, the mode and the median characterize the value of a feature in the statistical unit occupying a certain position in the variational series.

A median (Me) is the value of a feature in a statistical unit standing in the middle of the rank row, dividing the population into two parts equal in size [34].

A mode (Mo) is the value of the most common feature. A mode is widely used in statistical practice for the study of consumer demand, price registration, and so on [34].

For discrete variations, Mo and Me are chosen according to the definitions: the mode – as the value of a feature with the highest frequency f_i: the median position for the odd size of the population is determined by its number $n_{Me} = \frac{N+1}{2}$, where N is the size of the statistical population. In case of even size and number, the median is the average of two variants in the middle of the series.

The median is used as the most reliable indicator of a typical value of an inhomogeneous population, since it is insensitive to extreme values of a feature which may differ significantly from the main array of values of the indicator. In addition, the median finds practical application due to the special mathematical property: $\sum|x_i - Me| \to \min$.

In the interval series, Mo and Me values are calculated in a more complex way. The mode is defined as follows:

• the maximum frequency determines the interval in which the mode value is located. It is called modal;

• within the modal interval, the mode value is calculated by the formula:

$$Mo = x^L_{Mo} + h_{Mo} \times \frac{f_{Mo} - f_{Mo-1}}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} + f_{Mo+1})}, \qquad (6.8)$$

where x^L_{Mo} is the lower limit of the modal interval;

h_{Mo} is the width of the modal interval;

f_{Mo}, f_{Mo-1}, f_{Mo+1} are the frequencies of the modal, pre-modal and post-modal intervals respectively.

The following approaches are used to calculate the median in interval series:

• based on the cumulative frequencies the median interval is found. The median interval is the interval containing the central unit of a population;

• in the median interval, the value of Me is determined by the formula:

$$Me = x^L_{Me} + h_{Me} \times \frac{\frac{N}{2} - S_{Me-1}}{f_{Me}}, \qquad (6.9)$$

where x^L_{Me} is the lower boundary of the median interval;

h_{Me} is the width of the median interval;

N is the size of the statistical population;

S_{Me-1} is the accumulated frequency of the median interval;

f_{Me} is the frequency of the median interval.

For calculating Mo in unequal-interval series a different frequency characteristic is used – *the absolute density of distribution*:

$$Mo = x^{L}_{Mo} + h_{Mo} \times \frac{\varphi_{Mo} - \varphi_{M0-1}}{(\varphi_{Mo} - \varphi_{Mo-1}) + (\varphi_{Mo} - \varphi_{Mo+1})}, \qquad (6.10)$$

where φ_{Mo} is the absolute density of the modal interval distribution;

φ_{Mo-1} is the absolute density of distribution of the premodal interval;

φ_{Mo+1} is the absolute density of distribution of the postmodal interval.

The mode can be defined graphically: by a polygon of distribution in discrete series, by the histogram of distribution in interval series; the median can be defined by the cumulative curve.

To find the mode in the interval series, the right vertex of the modal rectangle must be connected to the right upper corner of the previous rectangle, and the left vertex is to be connected to the left upper corner of the subsequent rectangle. The abscissa of the intersection point of these straight lines is the mode of distribution.

Corresponding to the total population, for the median to be determined, the height of the largest ordinate of the cumulative curve is halved. A straight, parallel axis of the abscissa is drawn through the obtained point to its intersection with the cumulative curve. The abscissa of the intersection point is the median.

Thus, to characterize the position of the distribution series center three indicators can be used: the mean value of the feature, the mode and the median. To choose the type and form of a specific indicator of the distribution center, we must proceed from the following recommendations:

- for sustainable socio-economic processes, the arithmetic mean is used as the index of the center. Such processes are characterized by symmetric distributions in which $\bar{x}$ = Mo = Me;
- for unstable processes, the position of the distribution center is characterized by Mo or Me. For asymmetric processes, the median is the preferred characteristic of the distribution center, since it occupies a position between the arithmetic mean and the mode.

6.3. The measures of variation

After setting the average value ($\bar{x}$, Mo, Me), a question arises to what extent the individual values of a feature differ from each other and from their mean value.

A variation of a feature is the difference in the numerical values of the units of a population and their fluctuations near the mean that characterizes this population. The smaller is the variation, the more homogeneous is the population and the more reliable is the mean [34; 80].

Absolute and relative indicators are used to measure variation in statistics.

The absolute indicators of variation include:

- the magnitude of variation R;
- the mean linear deviation $\bar{l}$;
- the mean squared deviation (dispersion) σ^2;
- standard deviation σ.

The range of variation (R) is the simplest measure of variation, calculated by the formula:

$$R = x_{max} - x_{min}. \qquad (6.11)$$

This indicator makes the difference between the maximum and minimum values of the features and characterizes the scattering of the population elements. The scattering covers only the extreme values of the feature in the population, disregarding the repetitiveness of its intermediate values, and does not show the deviations of all variants of the feature values.

Scattering is often used in practice (for example, the difference between max and min pensions, wages in different industries).

Mean linear deviation ($\bar{l}$) takes into account the differences of all units of the studied population. The mean linear deviation is the arithmetic mean of the absolute values of the individual variation deviations from their arithmetic mean [34]. This indicator is calculated using the formulas of the simple and weighted arithmetic mean:

for non-grouped data: $\bar{l} = \frac{\sum_{i=1}^{N} |x_i - \bar{x}|}{N}$; (6.12) for grouped data: $\bar{l} = \frac{\sum_{i=1}^{m} |x_i - \bar{x}| \times f_i}{\sum_{i=1}^{m} f_i}$, (6.13)

where x_i is the middle of the corresponding interval;

$\bar{x}$ is the average of the totality.

In practical calculations, the average linear deviation is used to estimate the rhythm of production and the uniformity of supply.

In practice, other indicators of the average deviation from the mean are often used − dispersion and standard deviation, because modules have poor mathematical properties.

Dispersion of a feature σ^2 is the mean square of deviation of variations from their mean value; it is a conventional measure of variation. Depending on the original data, the dispersion is calculated using the formulas of the simple and weighted arithmetic mean:

for non-grouped data: $\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}$; (6.14)

for grouped data: $\sigma^2 = \frac{\sum_{i=1}^{m} (x_i - \bar{x})^2 \times f_i}{\sum_{i=1}^{m} f_i}$. (6.15)

When using the weighted average for the calculation of dispersion in interval distribution series, the mean values of chi (mid-intervals) that are not the average in the group are selected as variants of the feature values. The result is an approximate dispersion value.

There are simpler approaches to the calculation of dispersion. The most commonly used one is the shortened dispersion calculation method (moment method), according to which the dispersion σ^2 is the difference between the mean of the squares of the feature values x^2 and the square of their mean ($\bar{x}^2$):

$$\sigma^2 = \bar{x}^2 - (\bar{x})^2, \quad (6.16)$$

where $\bar{x}^2 = \frac{\sum x_i^2}{N}$ for non-grouped data;

$$\bar{x}^2 = \frac{\sum x_i^2 \times f_i}{\sum f_i}$$ for grouped data.

This method makes it possible to calculate the dispersion based on the original data without preliminary calculation of the deviations.

The dispersion dimension corresponds to the square of the dimension of the studied feature, so this indicator has no economic interpretation. In order to preserve economic sense, another measure of variation is calculated, the root mean square deviation.

The root mean square deviation σ is the root mean square of deviations of the feature individual values from their arithmetic mean [34; 80]:

a) for non-grouped data: $$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(x-\bar{x})^2}{n}}; \qquad (6.17)$$

b) for grouped data: $$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{m}(x-\bar{x})^2 f_i}{\sum f_i}}. \qquad (6.18)$$

The mean square deviation is a named value, it has the dimension of the mean feature, and is economically well interpreted. It is also used to estimate the reliability of the mean: the smaller is the mean square deviation σ, the more reliable is the mean value of x, the better the mean characterizes the population under study.

For distributions close to the normal distribution, the following relationship exists between the root mean square deviation and the mean linear deviation: $\sigma \approx 1.25 \times \bar{l}$.

The root mean square deviation plays an important role in the analysis of distribution series. Under normal distribution, the following relationship exists between the magnitude of the mean square deviation and the number of observations: within $x \pm 1\sigma$ the number of observations makes 0.683 (or 68.3 %); within the limits of $x \pm 2\sigma$, observations make 0.954 (or 95.4 %); within $x \pm 3\sigma$ observations make 0.997 (or 99.7 %). In practice there are no deviations in excess of 0.9773, this is called the rule of 3σ (three sigma).

If the median is used as the index of the distribution center, the so-called quadratic deviation can be used to characterize the variation [34]:

$$Q = \frac{Q_3 - Q_1}{2Me}, \qquad (6.19)$$

where Q_1, Q_3 are the first and third quartiles of the distribution respectively.

This indicator can be used in lieu of the variation scattering to eliminate the disadvantages associated with the use of the extreme feature values.

There is the following relationship between the mean square deviation, the mean linear deviation, the quartile deviation, the scattering of variation in a normally distributed population:

$$6\sigma \approx 7.5\bar{l} \approx 9Q \approx R \quad \text{or} \quad \sigma \approx 1.25\bar{l} \approx 1.5Q \approx 1/6R. \qquad (6.20)$$

Relative variation indicators are intended to evaluate and compare variations of populations that have different units of measure or different values of averages. These figures are calculated as the ratio of absolute indicators of variation to the arithmetic mean (median).

The most common relative variation indicator is the coefficient of variation V_σ. It is the ratio of the root mean square deviation to the arithmetic mean, expressed as a percentage:

$$V_\sigma = \frac{\sigma}{\bar{x}} \times 100\ \%. \qquad (6.21)$$

The coefficient of variation is used to characterize the homogeneity of the population under study. The statistical population is considered to be quantitatively homogeneous if the coefficient of variation does not exceed 33 %.

In addition to the variation coefficient, relative indicators of the series variability are:

the coefficient of oscillation: $V_R = \frac{R}{\bar{x}} \times 100\ \%$; (6.22)

the relative linear deviation: $V_d = \frac{\bar{l}}{\bar{x}} \times 100\ \%$; (6.23) the relative quartile deviation:

$$V_Q = \frac{Q}{\overline{x}} \times 100\ \% \ \text{ or } \ V_Q = \frac{Q}{Me} \times 100\ \% \ \text{ or } \ V_Q = \frac{Q_3 - Q_1}{2Me}.$$

Analysis of dispersion (from the Latin *dispersio* – scattering) is a statistical method that allows a researcher to analyze the influence of various factors on the studied variable. The method developed by R. Fisher, a biologist, in 1925 was originally used to evaluate plant raising experiments. Later, the dispersion analysis turned to be of general scientific significance for experiments in psychology, pedagogy, medicine, etc. [34].

The purpose of dispersion analysis is to check the significance of the difference between the mean values by comparing the dispersions. The dispersion of a measured feature is decomposed into independent components, each of which characterizes the influence of a particular factor or their interaction. Further comparison of such components helps to evaluate the significance of each factor investigated, as well as their combinations.

If the null hypothesis (about the equality of the mean values of several observation groups selected from the general population) is significant, the estimate of dispersion associated with intragroup variability should be close to the estimate of the intergroup dispersion.

Dispersion and its properties are basic concepts of the analysis of dispersion, so let us focus on their definitions.

Dispersion or second-order central moment is a measure of deviation of the values of a random variable from the mean. Higher dispersion values indicate larger deviations of the random value from the mean. If the dispersion is 0, all realizations of a random variable are at one point.

Dispersion as a basic measure of variation has a number of mathematical properties that help to understand its nature and simplify the calculation.

The properties of dispersion include the following [34; 80]:

1) the constant value dispersion is zero:

$$\sigma^2_{const} = 0. \tag{6.24}$$

This property is based on the fact that dispersion shows the variants' scattering close to the arithmetic mean, and the arithmetic mean of the constant value is zero;

2) the dispersion does not change if all variants are increased or decreased by the same number A:

$$\sigma_{(x_i-A)} = \sigma^2. \tag{6.25}$$

This means that dispersion can be calculated based on the deviation from any constant number rather than on the basis of the feature values;

3) if all variants are divided (multiplied) by the number A, the dispersion will decrease (increase) by A^2 times, the root mean square deviation will change by A times:

$$\sigma^2_{x/A} = \sigma^2/A. \tag{6.26}$$

This means that all the values of a feature can be divided by a constant number (for example by the value of the interval). It is necessary to calculate the root mean square deviation, and then multiply it by a constant number;

4) if we calculate the mean square of any value deviations of A in varying degrees from the arithmetic mean ($\bar{x}$), it will always be greater than the mean square of the deviations calculated from the arithmetic mean:

$$\sigma^2_A > \sigma^2. \tag{6.27}$$

Moreover, it is increased by a certain value – the square of the difference between the average value and this conditional value, that is

$$(\bar{x}-A)^2,\ \sigma^2_A = \sigma^2 + (\bar{x}-A)^2 \tag{6.28}$$

or

$$\sigma^2_A = \sigma^2 - (\bar{x}-A)^2 = \frac{\sum_{i=1}^{m}(\bar{x}-A)^2 f_i}{\sum_{i=1}^{m} f_i} - (\bar{x}-A)^2.$$

This means that the dispersion from the mean value is always less than the dispersion calculated from any arbitrary value, that is, it has the property of minimality;

5) if a series of observations consists of two groups of observations, the dispersion of the whole series is equal to the sum of the arithmetic mean

of the group dispersions and the arithmetic mean of the deviations of the group averages from the mean of the whole series, and the size of the groups is calculated as a weighted arithmetic mean.

There are *general, intergroup and intragroup dispersions*.

General dispersion (σ^2) measures the variation of a feature in the entire statistical population under the influence of all the factors that cause this variation [34]. It is calculated by the formula:

$$\sigma^2 = \frac{\sum_{i=1}^{m} (x_i - \bar{x})^2 f_i}{\sum_{i=1}^{m} f_i}. \qquad (6.29)$$

Intergroup dispersion (δ^2) characterizes the change in the feature due to the factors underlying the grouping. Thus, intergroup dispersion is the dispersion of local mean values. The calculation is made according to the formula [34]:

$$\delta^2 = \frac{\sum_{i=1}^{m} (\tilde{x}_i - \bar{x})^2}{m}, \qquad (6.30)$$

where $\tilde{x}_i$ is the local mean in each group;

m is the number of groups in the population.

Intragroup dispersion characterizes random variation, that is, fluctuations in a feature that occur under the influence of unaccounted factors and the feature independent of the variation – a factor underlying the grouping [34]. Intragroup dispersion is calculated for each homogeneous group:

$$\sigma_i^2 = \frac{\sum_{j}^{m_i} (x_i - \tilde{x}_i)}{f_i}. \qquad (6.31)$$

On the basis of intragroup dispersion, the mean of intragroup dispersions (residual) $\overline{\sigma_i^2}$ is calculated:

$$\overline{\sigma_i^2} = \frac{\sum_{i=1}^{m_i} \sigma_i^2}{m}. \qquad (6.32)$$

These types of dispersions are related to each other in the following way:

$$\sigma^2 = \delta^2 + \overline{\sigma_i^2}. \tag{6.33}$$

This ratio is called *the rule of adding of dispersions*. Obviously, the greater is the value of intergroup dispersion, the better is the grouping, the stronger the factorial feature's influence on the overall variation. In addition, using the mentioned rule, it is possible to calculate an unknown third dispersion based on two known dispersions.

The estimation of the tightness of the relationship is based on the rule of *making dispersions*. In the model of analytical grouping, the measure of tight relationship is the ratio of intergroup dispersion to total dispersion, which is called *the correlation relation* [34; 80]:

$$\eta^2 = \frac{\delta^2}{\sigma^2}. \tag{6.34}$$

where η^2 is the correlation relationship;

δ^2 is the intergroup dispersion;

σ^2 is the total dispersion.

There is also *the alternative feature dispersion*. Among the varying features there are those whose variation manifests itself in the fact that in some units of a population they occur, while in others don't. Features inherent in some objects but absent in others are called alternative features. The absence of an alternative feature is assumed to be 0, and its presence equal to 1. The proportion of objects possessing the feature of the totality is denoted by P, and the proportion of units not possessing the feature is Q, with p + q = 1.

The dispersion of the alternative feature is calculated by the formula [34; 80]:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_i)f_i}{\sum_{i=1}^{n} f} = \frac{(1-p)^2 p + (0-q)^2 q}{p+q} = \frac{qp(q+p)}{p+q} = pq, \tag{6.35}$$

where $\bar{x} = \frac{\sum xf}{\sum f} = \frac{1 \times p + 0 \times q}{p+q}$.

Analysis of dispersion is one of the main methods of statistical evaluation of the experiment results. It is widely used in the analysis of economic information. The merit of this method is that it gives fairly reliable conclusions about samples of a rather small size.

When investigating the variation of an effective feature under the influence of one or more factors, the researcher using analysis of dispersion can obtain (in addition to general estimates of the significance of dependencies) the differences in the magnitude of mean values, formed at different levels of factors and the significance of the factors' interaction. Analysis of dispersion is used to study the dependencies of both quantitative and qualitative features as well as their combination.

The essence of this method lies in the statistical study of the likelihood of the influence of one or more factors as well as their interaction on the effective feature [34]. Accordingly, analysis of dispersion solves three main problems:

1) a general assessment of the significance of differences between group mean values;

2) estimation of probability of interaction of factors;

3) assessment of the significance of differences between the pairs of mean values.

Most often, these problems are solved in the process of examining the impact of several factors on an effective feature.

When conducting the analysis of dispersion, it is assumed that the following conditions are met: the results of observations of the system input parameter Y are independent random variables following the normal law of distribution; random errors of observation obey the normal law of distribution; the initial factors x_i under study only affect the change in the mean values, while the dispersion of the observations remains constant; experiments are conducted with equal accuracy.

It is mandatory to check these conditions before conducting analysis of dispersions.

The basic scheme of analysis of dispersion includes: establishment of the main sources of variation of the effective feature and determination of the value of variation (sums of squares of deviations) according to the sources of its formation; determining the number of degrees of freedom corresponding to the components of the general variation; calculation of the dispersion as the ratio of the respective volumes of variation to their number of degrees of freedom; analysis of relationships between dispersions; estimation of the probability of difference between the mean values; formulation of conclusions. This scheme holds both for simple models of analysis of dispersion, when the data is grouped based on one feature, and for complex models, when the data are grouped based on two or more features. However, with the increase in the number of grouping features, the process of decomposing the total variation depending on the sources of its formation becomes complicated. According to the schematic diagram, the analysis of dispersion can be presented in the form of five successive stages (Fig. 6.4) [34].

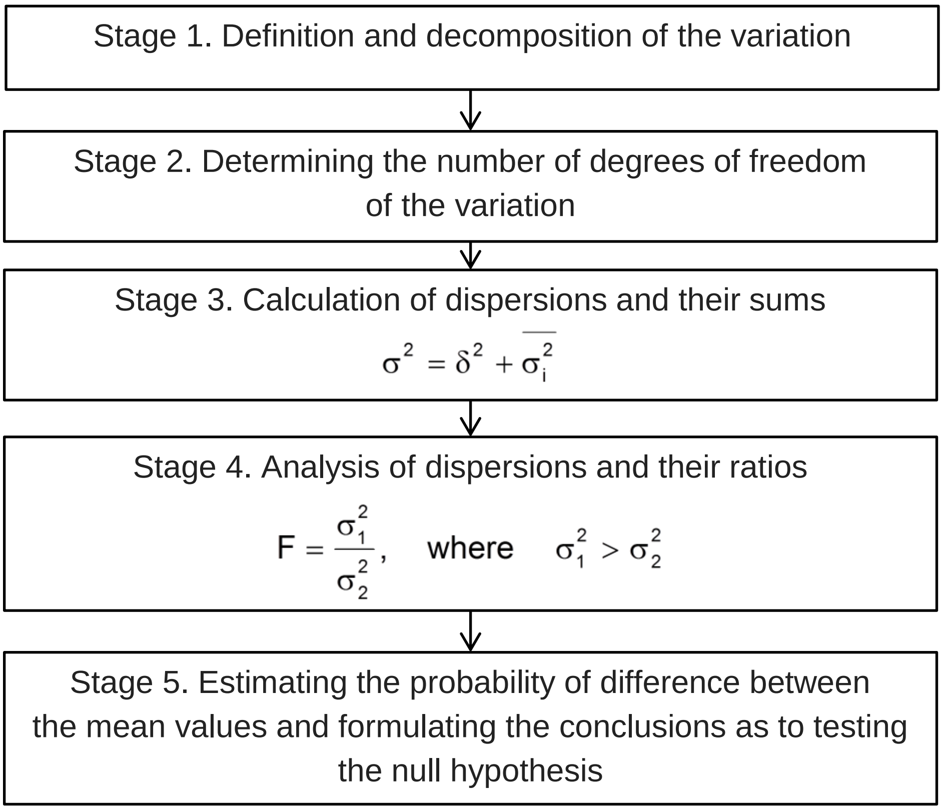

Stage 1. Definition and decomposition of the variation
Stage 2. Determining the number of degrees of freedom of the variation
Stage 3. Calculation of dispersions and their sums
$\sigma^2 = \delta^2 + \overline{\sigma_i^2}$
Stage 4. Analysis of dispersions and their ratios
$F = \frac{\sigma_1^2}{\sigma_2^2}$, where $\sigma_1^2 > \sigma_2^2$
Stage 5. Estimating the probability of difference between the mean values and formulating the conclusions as to testing the null hypothesis

Fig. 6.4. **The stages of the analysis of dispersion**

The solution of the dispersion analysis problems is based on the law of decomposition (addition) of variation. According to it, the total variation (fluctuations) of the effective feature is divided into two ones: the variation caused by the effect of the studied factor(s); the variation resulting from random causes. In order to assess the likelihood of differences between the group mean values, it is necessary to determine the inter- and intragroup variations. If the intergroup (factor) variation significantly exceeds the intragroup (residual) variation, the factor influences the effective feature, significantly changing the values of the group mean values.

To evaluate the significance of differences between the mean values and formulate conclusions from the test of the null hypothesis ($H_0 : \bar{x}_1 = \bar{x}_2 = ... = \bar{x}_n$) in the analysis of dispersion, we use a kind of standard – *F-criterion*, with the law of distribution established by R. Fisher [34; 75; 80]. This criterion is the ratio of two dispersions: the factor generated by the effect of the factor under investigation, and the residual dispersion resulting from random causes:

$$F = \frac{\sigma_1^2}{\sigma_2^2}, \quad \text{where} \quad \sigma_1^2 > \sigma_2^2. \qquad (6.36)$$

Dispersions σ_1^2 and σ_2^2 are estimates of the dispersion of the general population. If the samples with dispersions σ_1^2 and σ_2^2 are taken from the same population, where the variation of the values was random, the difference in the values σ_1^2 and σ_2^2 is also random.

To calculate the actual value of the F-criterion, the greater dispersion is taken in the numerator, so F > 1. Obviously, the larger the F-criterion, the greater the discrepancy between the dispersions. If F = 1, the need to evaluate the significance of the dispersion differences is rejected.

If F_{table} is less than F_{fact}, the differences in sample dispersions are determined not only by random factors; they are significant. The null hypothesis ($H_0 : \bar{x}_1 = \bar{x}_2 = ... = \bar{x}_n$) in this case is rejected; there is reason to argue that the effect of the factor on the effective feature is not significant. F-criterion enables us to establish whether there is or there is not a significant relationship between the group mean values as a whole.

When processing socio-economic data by the method of dispersion analysis, it should be borne in mind that due to the large number of factors and their correlation, it is difficult, even with careful equalization of conditions, to establish the degree of objective influence of each individual factor on the effective feature [34]. Therefore, the level of residual variation is caused not only by random reasons but also by significant factors that were not taken into account when constructing the analysis of dispersion. As a result, residual dispersion, as a base of comparison, is sometimes inadequate for its purpose if it is higher in magnitude and cannot act as a criterion for the essence of the influence of factors. In this regard, the problem of selecting the most important factors and equalizing the conditions for the manifestation of the action of each of them is actualized in the construction of models of dispersion analysis. In addition, the use of dispersion analysis assumes a normal or close to normal distribution of the studied statistical populations. If this condition is not met, the estimates obtained in the analysis of dispersion will be exaggerated.

Important concepts

Attributive distribution series are series constructed on descriptive grounds in the order of increasing or decreasing values of the observed feature.

Variational distribution series are series built on quantitative grounds.

A *variation* of a feature is the difference in the numerical values of the units of a population and their fluctuations near the mean values that characterizes this population.

Intragroup dispersion is a dispersion that characterizes random variation, that is, fluctuations in a feature that occur under the influence of unaccounted factors and a feature independent of variation − the factor underlying the grouping.

Analysis of dispersion is a statistical method that makes it possible to analyze the influence of various factors on the variable under study.

Dispersion or second-order central moment is a measure of the random variable values' deviation from the mean (value). Higher dispersion values indicate larger deviations of the random value from the mean. If the dispersion is 0, all values of the random variable are at one point.

Alternative feature *dispersion* is a dispersion that characterizes the variation of a feature inherent in some objects absent in other ones.

The *dispersion of a feature* is the mean square of dispersion of mean values; it is a common measure of variation.

Total dispersion is a dispersion that measures the variation of a feature in the entire statistical population under the influence of all the factors that cause this variation.

Median (Me) is the value of a feature in the statistical unit that stands in the middle of the rank row and divides the population into two parts equal in size.

Intergroup dispersion is a dispersion that characterizes the change in a feature due to the factors underlying the grouping. Thus, the intergroup dispersion is the dispersion of the local mean values.

Mode (Mo) is the value of the most common feature.

Accumulated frequence is the proportion of units in which the value of the feature does not exceed the given x*, i.e. it is the frequence with a growing result.

Accumulated frequency is the number of units of a population whose feature value does not exceed the given x*, i.e. it is the frequency with increasing totals.

Homogeneous populations are populations whose elements have common properties and belong to one type, class.

The uniformity of a population is a prerequisite for using other statistical methods (mean values, regression analysis, etc.).

A *time series* is a systematic set of numerical data characterizing changes in the studied phenomena over time.

The *order of distribution* is a systematic sequence of statistical units grouped based on a specific feature.

Mean square deviation is the mean square of deviations of the individual values of a feature from their arithmetic mean.

Mean linear deviation is a characteristic of variation of a feature that takes into account the differences of all units of the studied population.

A statistical series is an ordered set of values of the analyzed indicator, that is a statistical feature.

Frequence is the proportion of units of a population that have some value of a feature, i.e. it is frequency expressed as a relative value (a share of a unit or percentage).

Distribution density is a universal frequency characteristic that makes it possible to proceed from empirical distribution to a theoretical one.

Typical tasks

Task 1. We have some data on the distribution of machine tools at the enterprises of a machine-building complex in accordance with the lifetime (Table 6.1).

Table 6.1

The input data

Groups on the lifetime basis	Number of machine tools	Accumulated frequency
1	2	3
1 – 3	43	43

Table 6.1 (the end)

1	2	3
3 – 6	56	99
6 – 9	48	147
9 – 12	32	179
12 – 15	18	197
Total	197	

Identify the indicators of the central trend; determine the variation indicators.

The solution.

To calculate the indicators, it is advisable to draw up an auxiliary table (Table 6.2).

Table 6.2

The auxiliary calculation table

Groups on the lifetime basis	The middle of the interval, x_i	Number of machine tools, f_i	Accumulated frequencies, S	$x_i f_S$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 f_i$
Up to 3	1.5	43	43	64.5	−4.87	23.72	1 019.96
3 – 6	4.5	56	99	252	−1.87	3.50	196
6 – 9	7.5	48	147	360	1.13	1.28	61.44
9 – 12	10.5	32	179	336	4.13	17.06	545.92
12 – 15	13.5	18	197	243	7.13	50.83	914.94
Total		197		1 255.5			2 738.26

$\bar{X} = \frac{1255.5}{197} = 6.37$ years is the average lifetime.

Firstly, the interval containing these indicators is found in order to determine the mode and the median in the interval series, and then the specific values of these indicators are calculated.

The modal interval in this case is the interval of 3 – 6 years, i.e. the largest number of machine tools is in this interval. The mode equals:

$$\mathrm{Mo} = x^{L}_{\mathrm{Mo}} + h_{\mathrm{Mo}} \times \frac{f_{\mathrm{Mo}} - f_{\mathrm{Mo}-1}}{(f_{\mathrm{Mo}} - f_{\mathrm{Mo}-1}) + (f_{\mathrm{Mo}} - f_{\mathrm{Mo}+1})} =$$
$$= 3 + 3 \times \frac{56 - 43}{(56 - 43) + (56 - 48)} = 4.86\,\text{years.}$$

Therefore, in the studied population, the largest number of machine tools has a service life of 4.86 years.

The median interval is also the interval 3 – 6, because the middle of the variant ranked series (98.5) is found based on the accumulated frequencies where their sum is 99. The mode is:

$$\mathrm{Me} = x^{L}_{\mathrm{Me}} + h_{\mathrm{Mo}} \times \frac{\frac{N}{2} - S_{\mathrm{Me}-1}}{f_{\mathrm{Me}}} = 3 + 3\,\frac{\frac{197}{2} - 43}{56} = 5.97 \text{ years.}$$

Therefore, the service life is 5.97 years and it is the option that divides the variational series of distribution into two equal parts (98 machine tools have a service life below 5.97 years and 99 machine tools – more than 5.97 years).

The mean square deviation is:

$$\sigma = \sqrt{\frac{2738.26}{197}} = 3.73 \text{ years.}$$

The average service life is 6.37 years, and the mean deviation, which shows the level of the mean deviation of the individual feature values from the arithmetic mean, is 3.73 years. The coefficient of variation is:

$$V_{\sigma} = \frac{3.73}{6.37} = 0.585 \text{ or } 58.5\ \%.$$

According to the obtained results, the population is not homogeneous, because the obtained value of the coefficient of variation is greater than the threshold value of 33 %.

Task 2. We have some data on the distribution of the enterprise employees according to the length of service (Table 6.3).

Table 6.3

Distribution of workers depending on the length of service

Length of service, years	0 – 4	4 – 8	8 – 12	12 – 16	16 – 20	20 – 24	24 – 28
The number of workers in the same job, people	6	8	11	13	6	4	2

The solution.

Let's consider the calculation of indicators of variation based on the example of the series of distribution of the enterprise employees in accordance with the length of service. To do this, let's compile a subsidiary table (Table 6.4).

Table 6.4

Calculation of variation indicators for the distribution of workers in accordance with the length of service

Group number	Length of service, years			f_i	$f_i \times x_i$	Calculation of the mean linear deviation		Calculation of dispersion	
	x_i^L	x_i^U	x_i			$\lvert x_i - \bar{x} \rvert$	$\lvert x_i - \bar{x} \rvert \times f_i$	x_i^2	$x_i^2 \times f_i$
1	0	4	2	6	12	10	60	4	24
2	4	8	6	8	48	6	48	36	288
3	8	12	10	11	110	2	22	100	1100
4	12	16	14	13	182	2	26	196	2548
5	16	20	18	6	108	6	36	324	1944
6	20	24	22	4	88	10	40	484	1936
7	24	28	26	2	52	14	28	676	1352
Total	–	–	–	50	600	–	260	–	9192

Calculation of the average length of the workers' service:

$$\bar{x} = \frac{\sum x_i \times f_i}{\sum f_i} = \frac{600}{50} = 12 \text{ years.}$$

The spread shows the total range of change in the length of service; it is 28 years.

The mean linear deviation is:

$$\bar{l} = \frac{\sum |x_i - \bar{x}| \times f_i}{\sum f_i} = \frac{260}{50} = 5.2 \text{ years.}$$

The dispersion for this series is:

$$\sigma^2 = \bar{x}^2 - (xx^2 = \frac{\sum x^2 \times f_i}{\sum f_i} - \left(\frac{\sum x_i \times f_i}{\sum f_i} \right)^2 = \frac{9192}{50} - 12^2 = 183.84 - 144 = 39.84.$$

A parameter with this dimension cannot be interpreted, so we calculate the squared mean deviation.

The standard deviation is $\sigma = \sqrt{39.84} = 6.3$ years.

Let's check the relation between the mean linear deviation and the mean square deviation: $\sigma \approx 1.25 \times \bar{l}$ = 1.25 × 5.2 ≈ 6.5. It can be concluded that the distribution of workers depending on the length of service is close to normal.

The coefficient of variation is $V_\sigma = \frac{6.3}{12} \times 100\% = 53\%$ indicating a high level of fluctuation of the feature in the population.

Task 3. Table 6.5 shows interest rates in ten bank branches.

Table 6.5

The input data

Branches of commercial banks	Deposit interest rates in the branches, %					
Central	22	25	24	27	–	–
Affiliates	20	19	21	22	18	24

Identify the different types of deposit interest rate variations, show their relationship, provide interpretations and draw conclusions about the close relationship between the deposit rate and the bank branch.

The solution.

Find the mean value of the deposit interest rate for each branch by the formula for a simple arithmetic mean:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_m}{m} = \frac{\sum\limits_{i=1}^{m} x_i}{m};$$

$$\bar{x}_1 = \frac{22 + 25 + 24 + 27}{4} = 24.5\ \%;$$

$$\bar{x}_2 = \frac{20 + 19 + 21 + 22 + 18 + 24}{6} = 20.7\ \%;$$

$$\bar{x}_0 = \frac{24.5 \times 4 + 20.7 \times 6}{10} = 22.22\ \%.$$

For the central branch of the commercial bank, the intragroup dispersion is:

$$\sigma_1^2 = \frac{\sum\limits_{i=1}^{m}(x_i - \tilde{x}_i)}{f_i} = \frac{(22 - 24.5)^2 + (25 - 24.5)^2 + (24 - 24.5)^2 + (27 - 24.5)^2}{4} = 3.25.$$

It shows the variation of interest rates in the commercial bank's central branch.

Similarly, we make calculations for the affiliate:

$$\sigma_2^2 = \frac{(20 - 20.7)^2 + (19 - 20.7)^2 + (21 - 24.5)^2 + (22 - 20.7)^2 + (18 - 20.7)^2 + (24 - 20.7)^2}{6} = 5.92.$$

The average of group dispersions is:

$$\overline{\sigma_i^2} = \frac{\sum\limits_{i=1}^{m} \sigma_i^2}{m} = \frac{3.25 \times 4 + 5.92 \times 6}{10} = 4.85.$$

It shows the variation in deposit interest rates on average across the population, driven by factors other than differences between commercial bank branches.

The intergroup dispersions is:

$$\delta^2 = \frac{\sum_{i=1}^{m}(x_i - \widetilde{x}_i)^2}{m} = \frac{(24.5 - 22.2)^2 \times 4 + (20.7 - 22.2)^2 \times 6}{10} = 3.47.$$

It characterizes the variation of the group averages caused by differences in commercial bank branches (the deposit interest rate differs by 3.47).

According to the sum rule, the total dispersion is:

$$\sigma^2 = \delta^2 + \overline{\sigma_i^2} = 4.85 + 3.47 = 8.32.$$

It characterizes the effect of all factors on the variation in deposit interest rates.

Task 4. A firm has analyzed the quality of a batch of products. The audit has revealed that 2 % of all products in the batch is defective. The dispersion of the alternative feature is to be determined.

The solution.

Thus, p = 2 % = 0.02.

Since p + q = 1, q = 1 − p.

Accordingly, 98 % of the parts in the batch are qualitative, i.e. q = 98 % = 0.98.

The dispersion of the defective items is $\sigma^2 = pq = 0.02 \div 0.98 = 0.0196$.

This means that the variation of the spoilage from the average number of defective items is 0.14 ($\sigma = \sqrt{0.0196}$).

Reference to laboratory work

The guidelines for carrying out laboratory work on the topic "Analysis of distribution series" are presented in [100]. The laboratory work is aimed at obtaining the skills in the analysis of the distribution series using MS Excel. The task of the laboratory work is to analyze the statistical series of distribution using the add-on "Data Analysis".

Questions for self-assessment

1. What is variation of features? Give examples.

2. What indicators do you use to measure variation? Name them and give formulas.

3. How can the mode and the median be found graphically?

4. What are the disadvantages inherent in the spread of variation and the mean linear deviation?

5. Expand on the dispersion and the root mean square deviation and their place in the system of variation indicators.

6. What are the basic mathematical properties of dispersion?

7. Name the formulas for simplified calculations of dispersion and explain their essence.

8. How is the variation of an alternative feature measured?

9. What kinds of dispersion do you know? What is their essence?

10. Expand on the rule of adding (decomposition) of variations. Where is it used?

Questions for critical rethinking (essays)

1. Explain why the mean value is considered as a typical level of a feature in a population. How is the mean value correlated with other characteristics of the distribution center?

2. Expand on the sequence of calculations of total, intergroup, and intragroup dispersions for alternative features.

3. Describe the mathematical properties of a dispersion that help simplify the calculation of its magnitude.

4. Explain the concept of the moments of statistical distribution and explain the use of them.

5. Explain the need to calculate the empirical correlation relationship and the appropriateness of using it in the assessment of the tightness and strength of the relationship between the factorial and effective features.

7. Sampling and sampling distributions

Basic questions:

7.1. Sampling: types and methods.

7.2. Sampling errors.

7.3. A small sample. Measuring the sample size.

7.1. Sampling: types and methods

In the practice of statistical research, *sampling observation* is widely recognized as the most scientifically sound type of inconsistent observation. The dialectical unity of the particular and the general enables one to think individually and singularly about the common, and in the case of correct identification of the relationship one can partly judge the whole.

Sampling observation yields the most optimal result only if it is organized and aligned with the *scientific principles* of *the sampling method theory* [6; 13; 14; 34; 65; 75; 77]:

ensuring the randomness of sampling (due to the sampling of each of the units of a population under study, they have an equal probability of being selected);

ensuring a sufficient number of units selected (the more units are surveyed, the more accurate is the presentation of the population under study and the smaller is the sampling error).

An important role in the formation of the sampling method belongs to the works of J. Bernoulli. The mathematicians P. Chebyshev, O. Lyapunov, A. Markov, R. Fisher, W. Gosset made a significant contribution to the development of the theoretical foundations of the considered method. The theory of the sampling method was developed in the works of statisticians O. Chuprov, A. Kovalevsky. The classification of the forms of sampling was given by such statisticians as A. Boyarsky, B. Yastremsky.

Sampling is the kind of inconsistent observation, when the nature of a selected part of the units of a population gives an idea of the totality of the whole. That is, not all units of the studied phenomenon (process) are examined during the sample observation, but only some part of them, which is selected in a certain way. However, the observation is organized in such a way that this part of the selected units on a smaller scale represents the whole population.

Sampling observation is used to study various patterns of social phenomena.

Sampling observation has *significant advantages* over other methods of obtaining statistics:

rather high accuracy of survey results due to involvement of more qualified personnel in the process, which leads to reduction of registration errors;

saving time and money as a result of reduced workload; high efficiency in obtaining the results of the survey;

possibility to study very large statistical populations;

minimizing the damage (destruction) of the objects under study;

possibility of exploring completely inaccessible populations.

In a sample survey, a relatively small proportion of a statistical population is studied (5 – 10 %, more rarely 20 – 25 % of its units).

Sampling observation is often used in conjunction with continuous observation for in-depth research, refinement, and control of continuous observation results.

Sampling involves the following *steps* [34; 44; 48; 52; 62]:

justification of the purpose of statistical observation;

drawing up a program of observation and data development;

settling organizational aspects of observation;

determining the percentage and method of selection;

selection;

registration of the features of units under study;

generalization of the observation data and identification of their sample characteristics;

determination of the sampling error;

verification of the reliability of the sample observation results;

generalization of sample characteristics for the whole population.

According to the purpose of the task of the sample research it is necessary to introduce the following *concepts* [33; 34; 52; 56; 77]:

general population (N) – a population under study from which the units to be studied are selected (it may or may not have a finite value);

sampling population (sample) (n) – the part of the units of the general population selected for study.

The quality of the sample results depends on to what extent the sample composition represents the general population, i.e. how representative the sample is.

The *representativeness of a sample* is the correspondence of its properties and structure to the properties and structure of the general population. Sampling representativeness can be ensured only in the case of objectivity of data selection, which is guaranteed by the principles of random selection of units.

The *principle of randomness* implies that the inclusion or exclusion of a statistical unit from the sample cannot be influenced by any other factor except the case. The application of the randomness principle in the formation of a sample makes it possible to use the probability theory terms and concepts in the future.

Most often the following *characteristics of a population are determined* by sampling:

the mean value of a feature *in a population*, that is, the general mean value;

the share of an alternative feature in a population. An alternative feature is a feature that takes two values. If one of them changes as a given one, the proportion of the alternative feature will characterize the share of the statistical units that have the specified value of the alternative feature (for example, the proportion of rejected products in a batch of manufactured products);

the variance of a feature in a population, that is the general variance as an indicator of variation.

The purpose of *sampling observation* is to select *n* units from a general population and to calculate them on the basis of unknown parameters of a generic population.

In general, the *task* of *sample research* is formulated as follows [34]. Suppose there is some general population of a known size (N units) that has certain statistical characteristics:

$D = \frac{P}{N}$, the *general share* (the proportion of statistical units of the general population that have certain values of a feature), where P is the number of units of the general population that have a specific feature;

$\overline{x}$, the *general mean* (the arithmetic mean of the feature in the population);

σ^2, the *general variance* (the variance of the feature being investigated in the population);

σ, the *general mean* deviation (mean square deviation of the feature under study in the population).

To determine them, a sample population of n statistic units (n < N) having similar characteristics is formed:

w, the *sample fraction* (the proportion of statistical units that have this value of the feature in the sample population);

$\tilde{x}$, the *sample mean* (arithmetic mean of the feature in the sample population);

σ_s^2, *the sampling variance* (variance of a feature under study in a sample population);

σ_s, the sampling squared deviation (mean squared deviation of the feature under study).

It is necessary to obtain statistical estimates of the population characteristics on the basis of known sample characteristics.

The selection of units from the general population can be done differently, depending on certain conditions. The system of organization of the selection of units from the general population is called the *selection method.* Depending on whether the selected units are participating in the subsequent sampling, the selection *methods may be* repeated and nonrepeated [25; 52; 64; 77].

In the *case of repeated selection*, each unit participates in the sample as many times as the units are selected. That is, after registration of a feature, the unit returns to the general population, and there is a further likelihood that it may again fall into the sample population.

In the *nonrepeated selection* each selected unit participates in the selection only once and after registration of the features it does not return to the general population. Repeated and nonrepeated selection methods (depending on the nature of the selection unit) are applied in coordination with other types of selection. In practice, there are three *types of selection*:

individual – selection of population units;

group – selection of unit groups;

combinational – a combination of the first and second types.

Different types of selection can be done in different ways. The *method of sampling* is determined by the rule of formation of the sample population. In the practice of sample surveys they use random, mechanical, typical, serial and combined sampling [12; 16; 33; 34; 46; 62; 63].

Depending on how the selection unit changes in the case of successive multiple selections, there is a distinction between *single-stage* and *multi-stage* selection.

Each of these selection methods has its own peculiarities of sampling and methods for calculation of the average sampling error.

Technically, *random selection* is carried out by the method of drawing of lots or by the table of random numbers. It can be expected that there are representatives of different states among the selected units, which are characterized by a certain feature in the population. In this case, the mean value of the feature under study will be represented fairly accurately.

Actually random selection in pure form is rarely used, but it is a source for all other types of selection.

As noted, random selection may be repeated or nonrepeated.

Mechanical selection is the selection of units from the general population in any mechanical order. Most often mechanical selection is used in the case of an objective sequence of the units of selection. The selection is made nonrepeatedly at regular intervals. There is only one unit per sample interval.

In order to perform mechanical selection, it is necessary to set the reference step (the distance between the selected units) and the point of reference (the number of the unit to be surveyed first). The reference step is a set based on the estimated percentage of selection. For example, every tenth unit is selected for a ten percent selection, and every twenty percent for a twenty percent selection.

The peculiarity of mechanical selection is that it may cause systematic errors, which are associated with random coincidence of the selected interval and cyclic regularities in the arrangement of units of the general population. In order to avoid systematic errors, a statistical unit within each interval should be selected. This method is very convenient in cases where it is impossible to precompile a list of units of the general population (the sample is taken from a constantly evolving population).

When studying a complex population that can be divided into several qualitatively homogeneous groups on essential grounds for the purpose of the study, it is advisable to use *stratified sampling*.

Within each group of such sampling, random or mechanical sampling is performed. The resulting groups are usually unequal in the number of units, so the selection of units is carried out in proportion to the size of the group. Thus, the number of observations for each group is determined by the formula [34]:

$$n_{is} = N\frac{n_j}{N}, \tag{7.1}$$

where n_{is} is the number of observations from the i-th group of the general population;

N is the size of the general population;

n_j is the size of the i-th group of the general population.

If the proportions between groups in the sample match the proportions between groups in the population, the selection is called *typical*. According to typical selection the population is predivided into homogeneous groups, types, classes. The essence of typical selection lies in the typical zoning of the statistical population into homogeneous groups from which random or mechanical sampling is carried out.

Serial selection is used when statistical units are grouped into small *groups* or *series (nests).* For example, packages with a certain amount of finished goods can be considered as such series.

Either random or mechanical selection is used for selection of series. All units of the selected series are subject to observation. Serial selection is of great practical importance because a small number of series are inspected and this reduces the cost of observation.

The considered methods of sampling can be used in pure form in different combinations and sequences. The use of several methods of forming a sample in a single sample study is called *combined selection.* This selection is carried out in several stages, and each of them uses its own selection method. For example, in a family income survey, a sample survey is conducted in the following order:

settlements to be surveyed are established. For this purpose, stratified selection is used to make possible the selection of large cities, medium-sized cities, and other settlements;

the places where families live are established in each settlement: streets, houses. For this purpose mechanical selection is used (according to the list of streets and numbering of houses);

at each place of residence of families specific families are selected, to which a random or repeated random or mechanical selection is applied. Lists of apartments or family lists are used for selection.

Sampling methods do not only affect the accuracy of statistical estimates due to sampling errors, but also the size of the sample population.

7.2. Sampling errors

A *statistical estimate* or *a statistical characteristic* (parameter) of a population is the approximate value of the desired characteristic (parameter), which is obtained from the sample data.

There are two *types of sample estimates* used in statistics: point and interval estimates [6; 14; 25; 48; 62].

The *point estimate* characterizes the parameter value and is calculated based on the sample data. The *interval estimate* is the confidence interval of a parameter values with a given probability.

The *quality of statistical estimates* is determined by the following *properties:*

1) consistency. An estimate is considered possible if, with an unlimited increase in the sample size n → ∞ (N), its error tends to 0 [34]:

$$\lim_{n \to x} (\tilde{a} - a) = 0 \text{, because } n \uparrow \lim_{n \to x} \tilde{a} = a, \qquad (7.2)$$

where a is the value of the characteristics of the general population;

$\tilde{a}$ is the value of the sample characteristics;

$\tilde{a} - a$ is the sampling error;

2) unbiasedness. An estimate is considered unbiased if from a given sample size n, the mathematical expectation of error is 0.

For an unbiased estimate, its mathematical expectation will exactly equal the mathematical expectation of the sample characteristic [34]:

$$M[\tilde{a} - a] = 0 \text{ or } M[\tilde{a}] = M[a]. \qquad (7.3)$$

An unbiased estimate does not always give a good approximation of the estimated parameter, since the possible values of the obtained estimate may be strongly scattered around its mean. In this case, the evaluation must meet another requirement – *efficiency*;

3) efficiency. An estimate is considered efficient if its error, called a *sampling error*, is minimal.

In mathematical statistics, a sampling error is defined as [34]:

$$\mu(\tilde{a}) = \sqrt{M^2}\,[\tilde{a} - a] + S^2, \qquad (7.4)$$

where $M^2[\tilde{a} - a]$ is the square of mathematical expectation of the sampling error;

S^2 is the sampling variance.

An estimate is efficient if the condition is $\mu(\tilde{a}) \to \min$.

The following statements are valid for *point estimation*:

the point estimate of the *general share is the sample share*: $p \approx w$;

the point estimate of the *general mean is the sample mean*: $\bar{x} \approx \tilde{x}$.

Thus, it is known in advance that the estimates for these parameters are *consistent and unbiased*.

For *other parameters* of the general population, this statement *is not true*; *so* $\sigma^2 \neq \sigma_s^2$ and $\sigma \neq \sigma_s$.

In mathematical statistics it is proved that the point estimate of the general variance is a sample variance, which is adjusted for the ratio $\frac{n}{n-1}$.

That is $\sigma^2 = \sigma_s^2 \frac{n}{n-1}$, with increasing $n \frac{n}{n-1} \to 1$. Therefore, in samples with a size of more than 30 observation units, this ratio can be neglected.

Similarly, the point estimate of the general standard deviation is the sample root mean square deviation adjusted for $\frac{n}{n-1}$, that is $\sigma = \sigma \frac{n}{n-1}$.

The main *disadvantage* of *point estimates* is that they do not take into account sampling errors, that is, they are not effective. Therefore, it is more appropriate to have interval estimates of the general population parameters that take into account such errors.

Interval estimates meet all three requirements of statistical estimation quality.

Mathematical statistics shows that:

the interval estimate of the general share is its sampling share taking into account the error of the sample share, that is: $p \approx w \pm \mu_w$, where μ_w is the error of the sample share;

the *interval estimate of the general mean* is the sample mean, taking into account the error of the sample mean, that is: $\bar{x} \approx \tilde{x} \pm \mu_x$, where μ_x is the error of the sample mean.

The use of interval estimates means that the characteristics of the general population are within a certain range of values, which is determined by the corresponding sampling error.

If the sample is formed correctly, its error can be calculated in advance. In the general case, *sampling error* refers to the discrepancy that arises objectively between the characteristics of the sample and the general population.

Sampling errors are divided into registration errors and representativeness errors [44; 48; 52; 56; 81].

Registration errors occur due to incorrect or inaccurate information. Their source is the inattention of the registrar, incorrect filling in the forms, descriptions or misunderstanding of the essence of the problem being investigated.

Representativeness errors arise because of the inconsistency of the sample structure with that of the population. The source of their appearance is a different variation of the characteristic of statistical units, which results in the distribution of units in the sample population being different from the distribution of units in the general population. That is, the *error of representativeness* is the discrepancy between a certain characteristic of the population (share, mean, variance, etc.) and its sampling estimate.

According to the *reasons*, representativeness errors are divided into systematic and random.

Systematic errors of representativeness arise from the incorrect formation of a sample, which violates the basic principle of its scientific organization (the principle of randomness). Such errors are unidirectional and will bias the survey results.

Random representativeness errors arise from the randomness of selection and the related differences between sampling structures and the general population.

The mathematical basis for calculating and adjusting the magnitude of random sampling errors is *the sampling method theory.*

For example, the data on the academic success rate of students of a faculty (a general population) are compared based on two ten-percent random samples (Table 7.1).

Table 7.1

The distribution of students based on the level of academic achievement

The score on the national scale	The number of students, people		
	The general population	The first sample	The second sample
2 (unsatisfactory)	100	9	13
3 (satisfactory)	500	50	49
4 (good)	380	30	32
5 (excellent)	120	21	16
Total	1 100	110	110

The total population size *N* was 1 000 students, the size of each sample was 10 % of *N*, that is, *n* = 0.1 × 1 000 = 100 people. The initial data are represented by discrete series of students' distribution according to success rate, so the average success rate is calculated using the weighted arithmetic mean formula.

It is:

for the totality of the population:

$$\bar{x} = \frac{100 \times 2 + 500 \times 3 + 380 \times 4 + 120 \times 5}{1100} = 3.47;$$

for the first sample: $\tilde{x}_1 = \frac{2 \times 9 + 3 \times 50 + 4 \times 30 + 5 \times 21}{110} = 3.57$;

for the second sample: $\tilde{x}_2 = \frac{2 \times 13 + 3 \times 49 + 4 \times 32 + 5 \times 16}{110} = 3.46$.

The proportion of students who received good and excellent grades is:

for the totality of the population: $w \; = \frac{380 + 120}{1100} = 0.45$;

for the first sample: $w_1 = \frac{30 + 21}{110} = 0.46$;

for the second sample: $w_2 = \frac{32 + 16}{110} = 0.44$.

The difference between the sample and general population values is an accidental error of representativeness. Representativeness errors for the mean value are:

$$\tilde{x}_1 - \bar{x} = 3.57 - 3.47 = +0.10;$$
$$\tilde{x}_2 - \bar{x} = 3.46 - 3.47 = -0.01.$$

Accordingly, the representativeness errors for a fraction are equal to:

$$w_1 - w = 0.46 - 0.45 = +0.01;$$
$$w_2 - w = 0.44 - 0.45 = -0.01.$$

From the above calculations, it can be seen that the sample mean and the sample share are random variables that can take different values depending on the units that are in the sample. Sampling errors can also be considered random variables. They can take different values, so the mean (standard) of possible errors is determined.

The magnitude of the sampling error depends on the following *factors* [64; 75; 77; 81]:

the degree of the feature fluctuation in the general population (the more homogeneous the population being investigated, the smaller the average error in the same sample size);

the sample size (increasing or decreasing the sample size *n,* the average error can be adjusted: the more units are included in the sample, the smaller is the error size, because the more accurately the sample will represent the population);

the way of selecting units into a sample population (in practice, different methods of forming a sample population are used, but division of methods into random, repeated and nonrepeated selection is of fundamental importance).

The mean (standard) sampling error (μ) is the difference between the sample mean and the general population mean $(\tilde{x}-\bar{x})$, which does not exceed $\pm\sigma$.

In *a random repeated selection*, the statistical unit that got into a sample returns to the general population after registration of the feature being investigated and can be sampled again. Thus, an equal probability of selection is ensured for all units of a population.

In mathematical statistics it is proved that the average error of selection is determined by the formula [34]:

$$\mu=\sqrt{\frac{\sigma^2}{n}}, \tag{7.5}$$

where σ^2 is the general variance.

General variance, like other parameters of a general population, is an unknown value, but the relationship between the general and sample variance $\sigma^2=\sigma_s^2\frac{n}{n-1}$ is known. Then for a sufficiently large sample size (n > 30) $\frac{n}{n-1}\to 1$ it can be assumed that $\sigma^2\approx\sigma_s^2$. In the case of a small sample with n < 30 it is necessary to take into account the ratio $\frac{n}{n-1}$ and calculate the average error of a small sample by the formula [34]:

$$\mu_{MS} = \sqrt{\frac{\sigma^2}{n-1}}. \quad (7.6)$$

Thus, for an average quantitative feature, the average *sampling error* $\mu_{\tilde{x}}$ will be [34]:

$$\mu_{\tilde{x}} = \sqrt{\frac{\sigma_{\tilde{x}}^2}{n}}, \quad (7.7)$$

where $\sigma_{\tilde{x}}^2 = \frac{\sum (x_i - \tilde{x})^2}{n}$ is the sampling variance of a quantitative feature.

The average sampling error for a fraction is determined by the formula [34]:

$$\mu_w = \sqrt{\frac{\sigma_w^2}{n}}, \quad (7.8)$$

where $\sigma_w^2 = w(1-w)$ is the sampling variance of the fraction of the alternative feature.

The application of simple random sampling is very limited in practice. This is due to the fact that it is impractical and sometimes impossible to reobserve the same units (the unit once surveyed is not recalculated). Therefore, it is advisable to apply nonrepeated selection.

In the *case of random nonrepeated sampling*, the total number of statistical units in a population changes during the sampling process (decreasing each time per unit in the sample), since the sampled units do not return into the general population. Thus, the probability of single units being sampled by nonrepeated random selection also changes (for other units it increases). In general, the probability of any statistical unit being sampled in case of nonrepeated selection can be defined as: $1 - \frac{n}{N}$. The average sampling error for nonrepeated selection should also be adjusted to this value.

Thus, the calculation formulas of the average sampling error for non-repeated selection take the following form [34]:

for an average quantitative feature:

$$\mu_{\tilde{x}} = \sqrt{\frac{\sigma_{\tilde{x}}^2}{n} \times \left(1 - \frac{n}{N}\right)}, \tag{7.9}$$

for a faction of an alternative feature:

$$\mu_w = \sqrt{\frac{w \times (1 - w)}{n} \times \left(1 - \frac{n}{N}\right)}. \tag{7.10}$$

In practice, the sampling method is used to determine the limits of a specific sampling error. The magnitude of the limits of a particular error is determined by the degree of probability.

A sampling error calculated with a given degree of probability is called a marginal sampling error.

The marginal sampling error (Δ) is the maximum of the error given the probability of its occurrence. This means that with a given probability it is guaranteed that any sampling error will not exceed the marginal error. This probability is *trustworthy*.

The marginal sampling error Δ is calculated by the formula [34]:

$$\Delta = t\mu, \tag{7.11}$$

where t is the confidence factor whose values are determined by the confidence probability p_t.

The values of the confidence coefficient t are given in the tables of normal probability distribution. The most commonly used combinations are shown in Table 7.2 [34].

Table 7.2

The values of the confidence coefficient t

t	p_t	t	p_t
1	2	3	4
1	0.683	2.5	0.988
1.5	0.866	3.0	0.997
2.0	0.954	3.5	0.999

Thus, if t = 1, with the probability of 0.683 the discrepancy between the sample characteristics and the population parameters will not exceed one mean error.

Marginal sampling errors Δ for different parameters and different methods of selection of statistical units are calculated by the formulas given in Table 7.3 [34].

Table 7.3

Marginal sampling errors

Selection method	Marginal sampling errors	
	For the mean	For a fraction
Repeated	$\Delta_{\tilde{x}} = t \times \sqrt{\frac{\sigma_{\tilde{x}}^2}{n}}$	$\Delta_p = t \times \sqrt{\frac{w \times (1-w)}{n}}$
Nonrepeated	$\Delta_{\tilde{x}} = t \times \sqrt{\frac{\sigma_{\tilde{x}}^2}{n} \times \left(1 - \frac{n}{N}\right)}$	$\Delta_p = t \times \sqrt{\frac{w \times (1-w)}{n} \times \left(1 - \frac{n}{N}\right)}$

Knowing the magnitude of the marginal sampling error, it is possible to calculate the *intervals* for the *characteristics of a population* [25; 26; 34; 44; 48]:

the confidence interval for the general mean is equal to $\tilde{x} \pm \Delta_{\tilde{x}}$;

for the general fraction it is $w \pm \Delta_p$.

Let's consider the calculation of the mean and marginal sampling error, determining the confidence intervals for the mean, and the fraction based on the following example.

Example 7.1. 2,000 families lived in the region during the reporting period. A five percent one-off family survey was conducted during the assessment of demand for product A. It was found that 90 out of 100 families surveyed consumed the product. On average, each of the surveyed families consumed 5 units of goods ($\tilde{x} = 5$) with a standard deviation of 0.5 units (σ = 0.5 units). With probability ρ = 0.954, the proportion of families consuming product A and the limits of consumption should be found.

To obtain statistical estimates of the general population parameters, the following calculations are to be performed.

1. Determine the characteristics of the sample population:

a) *the sample fraction* w (the share of families consuming product A in the sample): $w = \frac{90}{100} = 0.9$;

b) *the sample mean* $\tilde{x}$ (average consumption of product A per family in the sample): $\tilde{x} = 5$ units.

2. Determine *the marginal* sampling *errors*:

a) for the fraction:

$$\Delta_p = t\sqrt{\frac{w \times (1-w)}{n} \times \left(1 - \frac{n}{N}\right)} = 2 \times \sqrt{\frac{0.9 \times (1-0.9)}{100} \times \left(1 - \frac{100}{2000}\right)} = 0.0585,$$

where $N = \frac{n \times 100}{5} = 2000$ families;

b) for the mean:

$$\Delta_{\tilde{x}} = t \times \sqrt{\frac{\sigma_{\tilde{x}}^2}{n} \times \left(1 - \frac{n}{N}\right)} = 2 \times \sqrt{\frac{0.5^2}{100} \times \left(1 - \frac{100}{2000}\right)} \approx 0.1.$$

3. Calculate *the confidence intervals of the general population characteristics*:

a) for the fraction:

$$w - \Delta_p \le p \le w + \Delta_p,$$

$$0.9 - 0.059 \le p \le 0.9 + 0.059,$$

$$0.841 < p < 0.959;$$

b) for the mean:

$$\tilde{x} - \Delta\tilde{x} \le \bar{x} \le \tilde{x} + \Delta\tilde{x},$$

$$5 - 0.1 \le \bar{x} \le 5 + 0.1,$$

$$4.9 < \bar{x} < 5.1.$$

Thus, with a probability of 0.954 it can be argued that the consumption of product A by the general population will be in the range from 4.9 to 5.1 units.

On the completion of the calculations, it is possible to determine the limit of consumption (demand) of product A in the region:

$$4.9 \times 0.841 \times 2000 < Q < 5.1 \times 0.959 \times 2000,$$

$$8240 < Q < 9780.$$

That is, with a probability of 0.954 it can be argued that the demand for product A will not be lower than 8 240 units, but will not exceed 9 780.

Let's consider the calculation of sampling errors for different selection schemes.

To determine *the mean mechanical sampling error*, the mean error formulas for random sampling are used.

A typical sample provides more accurate results than other methods of selection of units because it eliminates the effect of intergroup variance on the mean sampling error. Therefore, the mean sampling error will only depend on the mean of the intragroup variances. The typical sampling can be repeated or nonrepeated.

The average typical sampling error with repeated selection is determined by the formulas [26; 56]:

for an average quantitative feature:

$$\mu_{\tilde{x}} = \sqrt{\frac{\bar{\sigma}_s^2}{n}}, \tag{7.12}$$

where $\bar{\sigma}_s^2 = \frac{\sum \sigma_i^2 \times n_i}{\sum n_i}$ is the average of the intragroup sample variances;

for a fraction of an alternative feature:

$$\mu_w = \sqrt{\frac{\overline{w_i \times (1 - w_i)}}{n}}, \tag{7.13}$$

where $\overline{w_i \times (1 - w_i)}$ is the mean of the intragroup variance of a fraction of an alternative feature in the sample:

$$\overline{w_i \times (1 - w_i)} = \frac{\sum w_i(1 - w_i)n_i}{\sum n_i}. \tag{7.14}$$

The average typical sampling error with a nonrepeated selection is calculated as follows:

for an average quantitative feature:

$$\mu_{\tilde{x}} = \sqrt{\frac{\bar{\sigma}_s^2}{n} \times \left(1 - \frac{n}{N}\right)}, \tag{7.15}$$

where n is the sample size from a typical group;

N is the size of a typical group;

for a fraction of an alternative feature:

$$\mu_w = \sqrt{\frac{\overline{w_i \times (1 - w_i)}}{n} \times \left(1 - \frac{n}{N}\right)}\,. \tag{7.16}$$

In serial selection, the random error is slightly larger than in other selection methods. Since within series all statistical units are examined without exception, the value of the mean error of serial selection depends only on the intergroup (interserial) variance.

Thus, *the average serial sampling error* for *repeated selection* is determined as follows [26; 56]:

1) for an average quantitative feature:

$$\mu_{\tilde{x}} = \sqrt{\frac{\delta_{\tilde{x}}^2}{r}}\,, \tag{7.17}$$

where $\delta_{\tilde{x}}^2 = \frac{\sum\left(\tilde{x}_i - \tilde{\tilde{x}}\right)^2}{r}$;

$\tilde{x}_i$ is the mean value of the i-th series;

$\tilde{\tilde{x}}$ is the mean value for the entire sample;

r is the number of selected series.

2) for a fraction of the alternative feature:

$$\mu_w = \sqrt{\frac{\delta_w^2}{r}}\,, \tag{7.18}$$

where $\delta_w^2 = \frac{\sum\left(w_i - \bar{w}\right)^2}{r}$ is the intergroup variance of the serial sample fraction;

w_i is the proportion of the feature in the i-th fraction;

$\bar{w}$ is the total proportion of the feature in the entire sample.

The average error of serial sampling in nonrepeated selection is determined in the following way [26; 56]:

1) for an average quantitative feature:

$$\mu_{\tilde{x}} = \sqrt{\frac{\delta^2_{\tilde{x}}}{r} \times \left(1 - \frac{r}{R}\right)}\ , \tag{7.19}$$

where R is the total number of series in the general population;

2) for a fraction of an alternative feature:

$$\mu_w = \sqrt{\frac{\delta^2_w}{r} \times \left(1 - \frac{r}{R}\right)}\ . \tag{7.20}$$

To calculate the marginal sampling error for an average feature and for a fraction using the selection methods considered, the average sampling error must be multiplied by the confidence factor t, whose value depends on the level of the selected probability.

7.3. A small sample. Measuring the sample size

One factor that influences the magnitude of a sampling error is its *size*. The larger is the sample size, the smaller is the error. On the other hand, sample costs are associated with research costs: the larger is the size, the higher is the cost. Thus, the sample should be optimal in size to ensure the reliability of the study results and avoid additional costs.

The sample size can be determined based on the permissible error in sampling and the method of selection of statistical units. In the general case, the marginal sampling error is related to its size by the following ratio [34]:

$$\Delta = t \times \mu \ = t \times \sqrt{\frac{\sigma_s^2}{n}}\text{, where } n = \frac{t^2 \times \sigma_s^2}{\Delta^2}\,. \tag{7.21}$$

That is, with the increase of the estimated error, the required sample size is significantly reduced and vice versa. With different characteristics and different methods of sampling, the required sample size will be determined by the formulas given in Table 7.4 [26; 34; 56].

Table 7.4

The sample size for different selection methods

Selection methods	Sampling formulas	
	For an average feature	For a fraction
Repeated	$n=\frac{t^2\times\sigma_{\tilde{x}}^2}{\Delta_{\tilde{x}}^2}$	$n=\frac{t^2\times w\times(1-w)}{\Delta_p^2}$
Nonrepeated	$n=\frac{t^2\times\sigma_{\tilde{x}}^2\times N}{N\times\Delta_{\tilde{x}}^2+t^2\times\sigma_{\tilde{x}}^2}$	$n=\frac{t^2\times w\times(1-w)\times N}{N\times\Delta_p^2+t^2\times w\times(1-w)}$

Example 7.2. Let's calculate the optimal sample size for the average feature by nonrepeated selection based on the following data: N = 2000 units, $\sigma_{\tilde{x}}=\pm1$ mm, $\Delta\tilde{x}=0{,}5$ mm, t = 2 for p = 0.954.

$$n_{\tilde{x}}=\frac{N\times t^2\times\sigma_{\tilde{x}}^2}{\Delta_{\tilde{x}}^2\times N+t^2\sigma_{\tilde{x}}^2}=\frac{2000\times2^2\times1^2}{0.5^2\times2000+2^2\times1^2}=148\text{ units.}$$

Thus, a sample of 148 units will provide the specified accuracy of the results with nonrepeated selection.

In statistical analysis, there is often a need to compare sampling errors of different features or the same feature in different populations. Such comparisons are made by a relative sampling error (the coefficient of variation) υ_μ. The *coefficient of variation* is the relative standard error, which is determined by dividing the standard (mean) error in absolute terms by the mean value of the estimate; it is expressed as a ratio or as a percentage [34]:

$$\upsilon_\mu=\frac{\mu_{\tilde{x}}}{\overline{x}}\times100, \qquad (7.22)$$

where $\mu_{\overline{x}}$ is the standard (mean) sampling error.

The calculation of the relative standard error is recommended for positive quantitative indicators and not recommended for relative values (especially fractions, dynamics values, because it is very easy to misinterpret its meaning in these cases).

A relative error υ_{μ} indicates the possible deviation level of a sample estimate from the general population parameter.

Sampling observation, where the number of units does not exceed 30, is called a *small sample*.

The Student's criterion [34] is used to determine the possible limits of the error:

$$t = \frac{x - x}{\mu_{sms}}, \tag{7.23}$$

where $\mu_{sms} = \frac{\sigma}{\sqrt{n-1}}$ is the standard small sample error.

The small sample *marginal error* is determined as [34]:

$$\Delta_{sms} = t \times \mu_{sms}, \tag{7.24}$$

where t is the Student's test.

There are special tables compiled on the basis of the established regularity of small sample error distribution. These tables show the values of the Student's t-test and the corresponding probability levels for a different number of sample units. According to the table of probability values for the Student's t-test, the probability that the actual t value for random variables is no more than that in the table of absolute values is determined.

The ultimate goal of any sampling is to generalize its characteristics for the general population.

The methods of generalization of sampling *data* are as follows [6; 26; 34; 75]:

the method of direct recalculation: the average size of the feature, determined by sampling, is multiplied by the total population;

the coefficient method: comparing the sampling data with the solid one, they calculate the correction factor that is used to correct the solid observation material.

The correction factor is calculated as [34]:

$$Y_1 = Y_0 \times \frac{y_1}{y_0}, \tag{7.25}$$

where Y_1 is the size of the population, corrected for defect;

Y_0 is the size of the population without this correction;

y_1 is the size of the population at control points according to the original data;

y_0 is the size of the population at the same points according to the control measures.

The direct recalculation of sampling data for the entire population is used when the purpose of the study is to determine the size of the feature of the population with availability of only the number of its units.

The method of coefficients is used in the case where sampling observation is conducted to clarify the results of the entire observation.

Important concepts

Sample (n) is a part of the general population units selected for study.

Proper random selection is carried out according to the method of drawing lots or a table of random numbers.

General population (N) is the population being investigated from which the units to be studied are selected.

Marginal sampling error is the maximum of the error with the specified probability of its occurrence.

Interval estimate is the confidence interval of parameter values for a given probability.

Combined sampling is the use of multiple sampling methods in a single sampling study.

Small sample is a sample whose number does not exceed 30.

Selection methods (repeated, nonrepeated) are ways determining which units (already selected or not) participate in the subsequent selection.

Methods for generalization of the sampling data are the direct recalculation method and the coefficient method.

Mechanical sampling is the selection of units from the general population which occurs in any mechanical order.

Scientific principles of the sampling method theory is ensuring random selection of units; ensuring a sufficient number of units selected.

Representativeness error is the discrepancy between a certain characteristic of the general population (fraction, mean, variance, etc.) and its sampling estimate.

Serial sampling is applied when statistical units make small groups (series).

Average (standard) sampling error is the difference between the sample average and the population average, which does not exceed $\pm\sigma$.

Selection method is a system of organizing the selection of units from the general population.

Statistical estimate (of a parameter) of the general population is the approximate value of the desired characteristic (parameter), which is obtained from the sample data.

Typical sampling is the kind of sampling where a general population is predivided into homogeneous groups from which random or mechanical sampling is carried out.

Point estimate is an estimate that characterizes a parameter value and is calculated based on the sample data.

The *quality of statistical estimates* is determined by their properties such as consistency, unbiasedness, efficiency.

Typical tasks

Task 1. A five-percent mechanical sampling with 100 accounts was performed to determine the average period of using a bank loan. The survey found that the average loan term is 30 days with a standard deviation of 9 days. With a probability of 0.997, it is necessary to determine the limits of the term of using the loan in the general population.

The solution.

The average term of using a bank loan is within:

$$\tilde{x} - \Delta\tilde{x} \le \bar{x} \le \tilde{x} + \Delta\tilde{x}.$$

Because sampling is mechanical, the marginal sampling error is found as follows:

$$\Delta\tilde{x} = t \times \sqrt{\frac{\sigma_{\tilde{x}}^2}{n} \times \left(1 - \frac{n}{N}\right)} = 3 \times \sqrt{\frac{9^2}{100} \times \left(1 - \frac{100}{2000}\right)} = 2.625 \approx 3 \text{ days.}$$

Thus, it can be stated with a 99.7 % probability, that in general the period of using a bank loan is within $30 - 3 \le \bar{x} \le 30 + 3$, that is from 27 to 33 days.

Task 2. We have some data on the average number of employees at an enterprise – 500 people. According to the materials of a special sample survey of the enterprise employees work experience, it is found that 50 of them have more than 20 years of work experience. Selection was carried out by the method of random sampling.

It is necessary to determine with probability p = 0.954: a) the proportion (share) of workers with more than 20 years' experience in the sample population; b) the average and marginal error of the sample fraction; c) the limits within which the proportion of workers with more than 20 years' work experience in the general population resides. Based on the results of the calculations, draw conclusions.

The solution.

1. Determine the sample proportion (share) of workers with more than 20 years of work:

$$w = \frac{n}{N} = \frac{50}{500} = 0.10, \text{ or } 10.0\ \%.$$

2. Calculate the average error of the sample fraction:

$$\mu_p = \sqrt{\frac{w \times (1-w)}{n} \times \left(1 - \frac{n}{N}\right)} = \sqrt{\frac{0.1 \times (1-0.1)}{50} \times \left(1 - \frac{50}{500}\right)} = \sqrt{\frac{0.1 \times 0.9 \times 0.9}{50}} = 0.04$$

or 4.0 %.

3. To determine the interval estimation and construct the interval for the general share, we define the marginal error of the sample share. According to the table we find that the given probability p = 0.954 corresponds to the normalized deviation t = 2.

The marginal error for the sample share will be:

$$\Delta w = t \times \mu_p = 2.0 \times 0.04 = 0.08, \text{ or } 8.0\ \%.$$

4. Let's construct an interval in which with the given probability there is a general share of the enterprise workers with work experience of more than 20 years:

$$w - \Delta w < p < w + \Delta w;$$

$$0.10 - 0.08 < p < 0.10 + 0.08;$$

$$0.02 < p < 0.18;$$

or 2 % < p < 18 %.

Therefore, with probability p = 0.954 it can be stated that the share of the enterprise workers with more than 20 years of work experience in the general population is in the range from 2.0 to 18.0 %.

Task 3. There are 10 teams of workers in an enterprise shop. In order to study their productivity, a twenty-percent serial sample was made, to which 2 teams were selected. The survey found that the average productivity of workers in the teams was 4.6 and 3 thousand conventional units. With a probability of 0.954, determine the limits within which the average productivity of the workshop workers will be.

The solution.

The average productivity of workshop workers is within:

$$\tilde{x} - \Delta\tilde{x} \leq \bar{x} \leq \tilde{x} + \Delta\tilde{x}.$$

The sample average of the serial sampling is:

$$\tilde{x} = \frac{4.6 + 3}{2} = 3.8 \text{ thou conv. units.}$$

The variance of the serial sample is calculated as follows:

$$\delta_{\tilde{x}}{}^2 = \frac{\sum(\tilde{x}_i - \tilde{x})_2}{r},$$

where $\tilde{x}_i$ is the sample average of the series;

$\tilde{x}$ is the sample average of the serial sampling;

r is the number of series.

Therefore, the value of the variance is:

$$\delta_{\tilde{x}}{}^2 = \frac{(4.6 - 3.8)^2 + (3 - 3.8)^2}{2} = 0.64.$$

The marginal sampling error for the mean is calculated as follows:

$$\Delta_{\tilde{x}} = t\sqrt{\frac{\delta_{\tilde{x}}^2}{r} \times \left(1 - \frac{r}{R}\right)} = 2\sqrt{\frac{0.64}{2} \times (1 - \frac{2}{10})} = 1 \text{ thou conv. units.}$$

Therefore, with a 0.954 probability, it can be argued that the average labor productivity of shop workers is within $3.8 - 1 \leq \bar{x} \leq 3.8 + 1$, that is from 2.8 to 4.8 thou conv. units.

Task 4. The marketing research department is going to carry out a survey of demand for detergent A among city households. Determine the sample size with a 0.954 probability, so that the sampling error does not exceed 2 %, if according to previous surveys it is known that the variance is 0.08.

The solution.

Since the size of a general city household population is not known in the task, the following formula for calculating the required sample size for repeated selection is used:

$$n = \frac{t^2 \times \sigma_{\tilde{x}}^2}{\Delta_{\tilde{x}}^2} = \frac{2^2 \times 0.08}{0.02^2} = 800 \text{ people.}$$

Therefore, to carry out the aforementioned observation in order to satisfy certain requirements, it is necessary to survey 800 households of the city.

Reference to laboratory work

The guidelines for carrying out the laboratory work on the topic "Acquisition of skills in sampling observation using MS Excel" are presented in [100]. The laboratory work is aimed at obtaining skills in conducting a sample survey and processing its results using the MS Excel spreadsheet processor.

Questions for self-assessment

1. What is the essence of sampling, the necessity and expediency of using it?

2. What are the scientific conditions for the use of sampling observations?

3. What methods of selecting units into a sample population do you know?

4 What is the essence of mechanical and typical selection?

5. What is the essence of serial and combined selection?

6. Describe the average sampling error, specify the algorithm for calculation of the error.

7. Describe the marginal sampling error, specify the algorithm for calculation of the error.

8. What is the point and interval estimate of the general average, the confidence interval?

9. How is the required sample size calculated and on what factors does it depend?

10. What is a small sample and what are the peculiarities of finding errors in such samples?

Questions for critical rethinking (essays)

1. Compare repeated and nonrepeated samples, determine their benefits and limitations.

2. A survey was conducted:

a) of one in one hundred families of industrial workers to study the standard of living of the population in the industrial region;

b) of five large families of industrial workers to study their living conditions.

Which do you think of these surveys is sampling and why? Justify your answer.

3. What's your opinion as to whether the sample data agree with the statement that 30 % of business entities do not pay taxes in full, if it is known that out of 450 business entities surveyed, 125 revealed irregularities in the tax return regarding the amount of taxable income? Determine the proportion of business entities that conceal a portion of tax revenue, and the confidence limits of that share in the population. Draw a 99.7 % probability conclusion.

4. A survey of entrepreneurs on the assessment of economic and legal conditions of their activity is being designed. What do you think the sample size should be so that with a 0.954 probability relative sampling error would not exceed 8 %? According to the preliminary survey, it is known that 75 % of entrepreneurs are dissatisfied with the business environment.

5. 350 respondents took part in a sociological survey. The survey revealed that: 15 % avoid risk; 12 % are involved in securities transactions. In your opinion, what is the average sample error size for the proportion of respondents who avoid risk, and what is it for the proportion of respondents working in the securities market? Justify your answer.

8. Analysis of the concentration, differentiation and similarity of distributions

Basic questions:

8.1. Sequential distribution.

8.2. Statistical indicators of concentration and centralization.

8.3. Statistical evaluation of structural changes in time and space.

8.1. Sequential distribution

Quantiles are intended for a deeper study of the structure of series of distribution. A quantile is the value of a feature that occupies a specific place in the population ordered based on that feature.

There are the following types of quantiles [7; 24; 27; 28; 32; 37; 40; 41; 43; 90]:

• **quartiles** ($Q_{1/4}$, $Q_{2/4}$ = Me, $Q_{3/4}$) – the values of a feature that divide an ordered population into four equal parts;

• **deciles** (D_1, D_2, ..., D_9) – the values of a feature that divide a population into ten equal parts;

• **percentiles** – the values of a feature that divide a population into a hundred equal parts.

If the data is grouped, the quantile value is determined based on the accumulated frequencies: the number of the group containing the i-th quantile. It is defined as the number of the first group from the beginning of the series in which the sum of accumulated frequencies equals or exceeds i × S, where i is the quantile index.

If it is an interval series, the value of the quantile is determined by the formula [7; 24; 27; 28; 32; 37; 40; 41; 43; 90]:

$$Q_i = x^n_{Q_i} + h_{Q_i} \times \frac{i \times S - S_{Q_i - 1}}{f_{Q_i}}, \qquad (8.1)$$

where $x^n_{Q_i}$ is the lower boundary of the interval in which the i-th quantile is located;

$S_{Q_i - 1}$ is the sum of the accumulated frequencies of intervals that precede the interval in which the i-th quantile is;

f_{Q_i} is the frequency of the interval in which the i-th quantile is.

The formula for calculating the first and ninth deciles looks as follows [40; 41; 43]:

$$D_1 = x_{D_1} + \frac{i}{f_{D_1}} \times \left(\frac{\sum_{i=1}^{k} f_i}{10} - S_{D_1-1} \right); \tag{8.2}$$

$$D_9 = x_{D_9} + \frac{i}{f_{D_9}} \times \left(\frac{9 \times \sum_{i=1}^{k} f_i}{10} - S_{D_9-1} \right), \tag{8.3}$$

where x_{D_1} is the lower limit of the interval in which the first decile is located;

x_{D_9} is the lower boundary of the interval in which the ninth decile is located;

$\sum_{i=1}^{k} f_i$ is the number of observations;

S_{D_1-1} is the cumulative frequency in the interval preceding the first decile;

S_{D_9-1} is the cumulative frequency in the interval preceding the ninth decile;

f_{D_1} is the frequency in the first decile interval;

f_{D_9} is the frequency in the ninth decile interval;

i is the magnitude of the interval.

If the center of distribution is given by the median, the *quartile coefficient of variation* is taken for the relative degree of variation [32; 37; 40; 41; 43; 101; 104]:

$$V_Q = \frac{Q_3 - Q_1}{2 \times Me}. \tag{8.4}$$

The relative quartile deviation is calculated as [32; 37; 40; 41; 43; 101; 104]:

$$V_Q = \frac{Q}{\overline{x}} \times 100\ \%, \text{ or } V_Q = \frac{Q}{Me} \times 100\ \%, \text{ or } V_Q = \frac{Q_3 - Q_1}{2Me}.$$

The ratio of deciles is also used to estimate the degree of variation.

Thus, the coefficient of decile differentiation shows the multiplicity of the ratio of the ninth (D_9) and first (D_1) deciles [32; 37; 40; 41; 43; 101; 104]:

$$V_D = \frac{D_9}{D_1}. \tag{8.5}$$

The distribution pattern analysis involves assessment of the degree of homogeneity of the totality, asymmetry and excess of distribution. Finding out the general nature of distribution involves not only evaluation of the degree of its homogeneity, but also the study of the distribution form, i.e. the evaluation of symmetry and excess.

It is known from mathematical statistics that with an increase in a statistical population size ($\infty \rightarrow N$) and simultaneous decrease in the grouping interval ($x_i \rightarrow 0$), the polygon or histogram of distribution intensely approaches some smooth curve, which is the boundary for these graphs. This curve is called the empirical distribution curve; it is a graphical representation in the form of a continuous line of the change in frequencies, functionally associated with the variation.

The statistics distinguishes the following types of distribution curves [1; 3; 6; 7; 12; 16; 17; 24; 27; 28; 32; 37]: unimodal curves; multimodal curves.

Homogeneous populations are described by unimodal distributions. The multimodality of distribution indicates the heterogeneity of the population being studied or the poor grouping.

Unimodal distribution curves are divided into symmetric, moderately asymmetric and extremely asymmetric.

The distribution is symmetric if the frequencies of any two variants equally spaced on either side of the distribution center are equal. In such distributions $\bar{x} = Mo = Me$.

Asymmetry coefficients are used to characterize asymmetry.

The Pearson asymmetry coefficient is most commonly used [1; 3; 6; 12; 16; 17]:

$$A_S = \frac{\bar{x} - Mo}{\sigma}. \tag{8.6}$$

In unimodal distributions the value of this parameter changes from −1 to + 1. In symmetric distributions $A_S = 0$.

In the case of $A_S > 0$ a right-sided asymmetry is observed (Fig. 8.1). In distribution with right-sided asymmetry $Mo \le Me \le \bar{x}$.

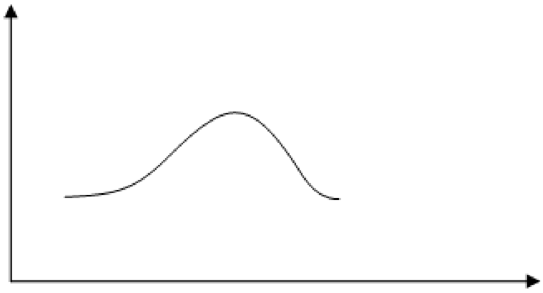


Fig. 8.1. **The right-sided asymmetry** [1; 3; 6; 7; 12; 32; 37; 40; 41; 43]

For $A_S < 0$ the asymmetry is negative left-sided, $Mo > Me > \bar{x}$ (Fig. 8.2).

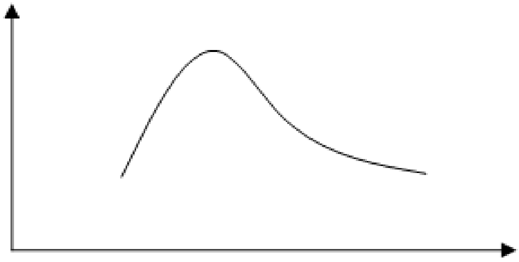


Fig. 8.2. **The left-sided asymmetry** [1; 3; 6; 7; 12; 32; 37; 40; 41; 43]

The closer is the A_S modulus to 1, the more significant is the asymmetry:

- if $| A_S | < 0.25$, the asymmetry is considered to be insignificant;
- if $0.25 < | As | < 0.5$, the asymmetry is considered to be moderate;
- if $| A_S | > 0.5$, the asymmetry is significant.

Pearson's asymmetry coefficient characterizes asymmetry only in the central part of the distribution, so the *asymmetry coefficient* calculated on the basis of the third-order central moment is more common and more accurate [1; 3; 6; 12; 32; 37; 40; 41; 43]:

$$A_S = \mu_3 : \sigma^3, \tag{8.7}$$

where μ_3 is the central moment of the third order;

σ^3 is the mean square deviation of the third degree.

The algebraic central moment of distribution is the arithmetic mean of the k-th degree of deviation of the feature individual values from the mean:

a) the central moment for a nongrouped series is [37; 40; 41; 43]:

$$\mu_k = \frac{\sum_{i=1}^{m}(x_i - \bar{x})^k}{n}; \qquad (8.8)$$

b) the central moment for a grouped series is [1; 3; 6; 7; 12; 32; 37; 40; 41; 43]:

$$\mu_k = \frac{\sum_{i=1}^{m}(x_i - \bar{x})^k \times f_i}{\sum_{i=1}^{m} f_i}. \qquad (8.9)$$

It is obvious that the second-order moment is the variance that characterizes the variation. The moments of the third and fourth orders characterize asymmetry and excess, respectively.

Accordingly the formula for determining the central moment of the third order have the following form:

a) the central moment for a nongrouped series is [1; 3; 6; 12; 32; 37; 40; 41; 43]:

$$\mu_3 = \frac{\sum_{i=1}^{m}(x_i - \bar{x})^3}{n}; \qquad (8.10)$$

b) the central moment for a grouped series is [1; 3; 6; 12; 32; 40; 43]:

$$\mu_3 = \frac{\sum_{i=1}^{m}(x_i - \bar{x})^3 f_i}{\sum_{i=1}^{m} f_i}. \qquad (8.11)$$

In symmetrical distribution μ_3 = 0. The greater is the skewness of the series, the greater is the value μ_3.

Theoretically, the asymmetry coefficient has no boundaries, but in practice its value is not too large and does not exceed unity in moderate distribution.

To estimate the significance of the coefficient of asymmetry calculated by the second method its root mean square error is determined [1; 3; 6; 12; 16]:

$$\sigma_{As} = \sqrt{\frac{6 \times (N - 1)}{(N + 1) \times (N + 3)}}. \qquad (8.12)$$

If $\frac{|A_S|}{\sigma_{A_S}} > 3,$ the asymmetry is essential.

For unimodal distributions another indicator of estimation of its form, *kurtosis*, is used. Kurtosis is an indicator of the peakedness of distribution. It is calculated for symmetric distributions based on the fourth-order central moment [7; 32; 37; 40; 41; 43; 90; 89]:

$$E_x = \frac{\mu^4}{\sigma^4} - 3. \qquad (8.13)$$

In symmetric distribution close to normal $E_X = 0$. Obviously in a peaked distribution $E_X > 0$, in a flat-topped distribution $E_X < 0$.

To estimate the significance of the excess factor, the root mean square error is calculated by the formula [1; 3; 6; 7; 12; 16; 17; 24; 27; 28; 32]:

$$\sigma_{E_X} = \sqrt{\frac{24 \times n \times (n - 2) \times (n - 3)}{(n - 1)^2 \times (n + 3) \times (n + 5)}}. \qquad (8.14)$$

The result of evaluation of the significance of the asymmetry and kurtosis indices enables us to conclude whether this empirical distribution can be attributed to the type of normal distribution curves.

Let's calculate the asymmetry and kurtosis indices for a series of distribution of company employees according to the length of service. We have the following characteristics for this series: $\bar{x} = 12$ years, Mo = 12.9 years, σ = 6.3 years.

Pearson's asymmetry coefficient is [41; 43; 89; 90; 101]:

$A_S = \frac{\bar{x} - Mo}{\sigma} = \frac{12 - 12.9}{6.3} \approx -0.14 < 0,$ indicating that there is a slight left-hand asymmetry in the central part of the distribution.

The asymmetry coefficient is calculated by the central moment of the third order [1; 3; 6; 7; 12; 16; 17; 24; 27; 28; 32; 37; 40; 41; 43; 89; 90]:

$$A_S = \frac{\mu_3}{\sigma^3} = \frac{\sum(x_i - \bar{x}) \times f_i}{\sum f_i} = \frac{61.44}{6.3^3} = \frac{61.44}{250} = 0.24 > 0.$$

This means that in the whole series there is a right-hand asymmetry.

The calculation of the central moment of the third order μ_3 is given in Table 8.1.

Table 8.1

Calculation of the central moment of the third and fourth order

[1; 3; 6; 12; 16; 17; 89; 90; 101; 103; 104]

Number	x_i	f_i	$x_i - \bar{x}$	$(x_i - \bar{x})^3$	$(x_i - \bar{x})^3 \times f_i$	$(x_i - \bar{x})^4$	$(x_i - \bar{x})^4 \times f_i$
1	2	6	–10	–1 000	–6 000	10 000	60 000
2	6	8	–6	–216	–1 728	1 296	10 368
3	10	11	–2	–8	–88	16	176
4	14	13	2	8	104	16	208
5	18	6	6	216	1 296	1 296	7 776
6	22	4	10	1 000	4 000	10 000	40 000
7	26	2	14	2 744	5 488	38 416	76 832
Total	14	50	–	–	3 072		195 360

The rate of excess is [1; 3; 37; 40; 41; 43; 89; 90; 101; 103; 104]:

$$\sum x = \frac{\mu_4}{\sigma^4} - 3 = \frac{195\,360}{50} \div 6.3^4 - 3 = \frac{3\,907.2}{1575.3} - 3 = 2.5 - 3 = -0.5.$$

It indicates that the distribution is flat-topped.

8.2. Statistical indicators of concentration and centralization

In the statistical analysis, along with the characteristics of uneven distribution of a certain feature among individual components of a population, it is important to estimate the concentration of the feature values in its individual parts (the distribution of property or income among individual population groups, distribution of the number of employed according to certain economic activities, distribution of the farmland among individual agro-industrial complexes).

That is, the estimation of differences between two distributions in space and time is based on the comparison of the shares of these distributions [7; 24; 27; 28; 32; 37; 40]:

the share of the distribution of the population elements;

the share of the distribution of the feature values.

Example 8.1. The data given in Table 8.2 show uneven distribution of territorial communities as to the size of received "warm credits" and consumed electricity.

Table 8.2

The estimated data [7; 32; 37; 40; 41; 43; 89; 90; 101; 103; 104; 88]

Size of received "warm credits", thousand UAH	% to total		Share deviation module $\frac{1}{100} \times \left| d_j - D_j \right|$
	Share of territorial communities, d_j	Share of electricity consumed, D_j	
Up to 5	20	4	0.16
5 – 10	38	5	0.33
10 – 20	22	8	0.14
20 –50	13	12	0.01
50 –100	4	25	0.21
100 and more	3	46	0.43
Total	100	100	1.28

The first group is made up of 20 % of territorial communities, and the share of consumed electricity is 4 %. However, the latter group covers only 3 % of the territorial communities, which consume 46 % of electricity. The deviations of the parts of two distributions – according to the number of elements of the population dj and the volume of values of the features Dj – make the base for calculation of *the concentration coefficient*.

If the distribution of the feature values in the *population* is uniform, the shares are the same: $d_j = D_j$. The deviations of the shares indicate a certain concentration.

The upper limit of the sum of deviations $\sum \left| d_j - D_j \right| = 2$, and therefore the *concentration coefficient* is calculated as the half-sum of the deviation modules [1; 3; 6; 12]:

$$K = \frac{1}{2} \sum_{j=1}^{m} \left| d_j - D_j \right|, \qquad (8.15)$$

where d_j is the share of distribution of the population elements;

D_j is the proportion of distribution of the feature values.

Concentration values range from zero (uniform distribution) to one (total concentration). The greater is the degree of concentration, the greater is the value of the K coefficient. In our example K = 1.28 : 2 = 0.64, which indicates a high degree of concentration of electricity consumption by industrial enterprises.

Concentration coefficients are widely used in regional analysis to assess the uniformity of territorial distribution of production capacities, financial resources, etc.

The localization of the feature values in the individual constituents of a population is determined by *the coefficient of localization* characterizing the ratio of shares [12; 16; 17; 24; 27; 28; 32; 37; 40; 41; 43; 89; 90]:

$$L_j = \frac{D_j}{d_j} \times 100. \tag{8.16}$$

According to Table 8.3 the coefficients of localization indicate the unevenness of the receipt of orders for services of the companies.

Table 8.3

The estimated data [7; 28; 32; 37; 40; 41; 43; 89; 90; 101; 103]

Companies	Share of visits to the company website, d_j	Share of service orders received, D_j	Coefficients of localization, L_j, %
A	30	34	113
B	50	42	84
C	20	24	120
Total	100	100	–

The coefficient of similarity of the structures of two populations is calculated by analogy with the coefficient of concentration [101; 103; 104]:

$$P = 1 - \frac{1}{2}\sum_{1}^{m}\left|d_j - d_K\right|, \tag{8.17}$$

where d_j and d_K are the share of distribution of elements of j-th and k-th sets.

If the structures of the populations are the same, P = 1; if they are absolutely opposite, P = 0. The more similar the structures are, the greater is the value of P.

Based on Table 8.4 we calculate the coefficient of similarity of structures for the data on the purchased tours based on the time for their two types. The structures of purchased tours are similar, provided that the value of P obtained will approximate one [1; 3; 6; 7; 12; 16; 17; 24; 27; 28; 32; 37; 40; 41; 43; 89; 90; 101].

Table 8.4

The structure of the purchased tours based on time, %

Type of tour	Share of purchased tours based on time		
	Off-season	Holidays and weekends	Season
Beach	36	24	40
Excursion	25	42	33

$$P = 1 - \frac{|36-25| + |24-42| + |40-33|}{2} = 1 - 0.18 = 0.82.$$

Comparison of structures based on the deviations of shares is appropriate for series with unequal intervals, especially in attributive series. The analysis of variation in the distribution series must be supplemented by the calculation of the differentiation coefficient.

The coefficient of differentiation (K_{dif}) is defined as the ratio of two mean values obtained from 10 % of the largest and smallest values of the feature under study.

Example 8.2. We have some data on the profit margins of 20 commercial banks. Two banks have the lowest profit margins (10 % of the total), which corresponds to 3.7 and 4.3 mln UAH; two banks have the highest profits amounting to 7.9 and 8.1 mln UAH. The average value for the smallest profit is 4.0 mln UAH; the average value in the group of the most profitable banks is 8.0 mln UAH. In this case [1; 16; 17; 24; 27; 28; 32; 37; 89; 90; 101; 103; 104]:

$$K_{dif} = \frac{\bar{x}_{highest}}{\bar{x}_{lowest}} = \frac{8}{4} = 2. \qquad (8.18)$$

That is, the profit margin of 10 % of the highest income banks is twice the profit margin of 10 % of the lowest income banks.

Centralization means the concentration (aggregation) of the feature size in individual units (for example capital in individual commercial banks, products of some kind at individual enterprises, etc.).

The generalized indicator of centralization I_z is calculated by the formula [7; 16; 17; 24; 27; 28; 32; 37; 40; 41; 43; 89; 90; 101; 103; 104]:

$$I_z = \sum_{i=1}^{n} \left(\frac{m_i}{\sum_{i=1}^{n} m_i} \right)^2 , \tag{8.19}$$

where m_i is the feature value of the i-th population unit;

$\sum_{i=1}^{n} m_i$ is the size of the feature in the total population;

n is the the population size (the number of units included in the population).

The maximum value of I_z is only reached if the population consists of only one unit, which owns the entire size of the feature. The minimum value of this indicator approaches zero, but never reaches it.

The estimation of the uneven distribution between the individual constituents of a population is based on the comparison of the shares of two distributions – based on the number of the population elements d_i and the volume of values of the feature D_i. If the distribution of the feature values is uniform, $d_i = D_i$; and the deviations of the shares indicate a certain irregularity, which is measured by the coefficients of localization and concentration [1; 3; 6; 12; 16; 17; 24; 27; 28; 32; 37; 40; 41; 43; 89; 90; 101; 103; 104].

A type of cumulative curve is *the concentration curve*, or the *Lorentz curve*. To plot the concentration curve, on both axes of the rectangular coordinate system, a 0 to 100 percentage scale is plotted. The accumulated frequencies characterizing the distribution of units of the population are plotted on the abscissa axis, and the accumulated values of the share (in percentage) according to the feature volume are plotted on the coordinate axis. According to the graph, the uniform distribution of the feature is the square diagonal. In the case of uneven distribution, the graph is a concave curve, depending on the concentration level of the feature Fig. 8.3.

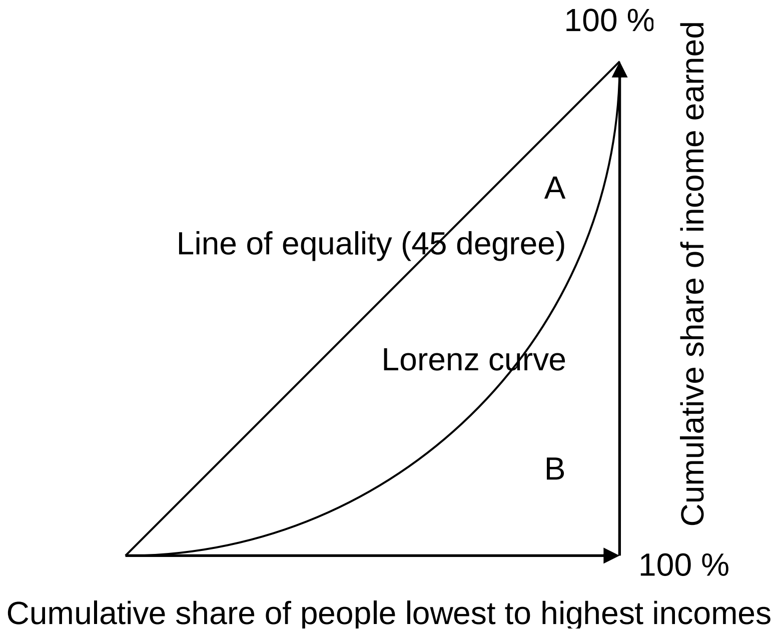

100 %
A
Line of equality (45 degree)
Lorenz curve
B
Cumulative share of income earned
100 %
Cumulative share of people lowest to highest incomes

Fig. 8.3. **The Lorentz curve (the concentration curve)** [24; 27; 32; 40; 41]

The graph of the function (in a rectangular coordinate system) is the Lorentz curve, which is convex downward and extends below the diagonal of a unit square, located in the first coordinate quarter.

Each point on the Lorentz curve corresponds to the statement that 20 % of the poorest population receive 7 % of its total income. In the case of a completely equal distribution, each population group has an income that is proportional to its size. Such a case is described by a line of perfect equality, which is actually a straight line connecting the origin and the point (1; 1), that is, the diagonal of a unit square.

In the case of complete unevenness of distribution (when only one member of the community receives income), the curve (the line of perfect equality) first adheres to the abscissa and then jumps from point (1; 0) to point (1; 1). Any other Lorentz curve is located between the absolute equality curve and the total inequality curve [1; 3; 6; 7; 12; 16; 17; 24; 27; 28; 32; 37; 40; 41; 43; 88; 89; 90].

Lorentz distributions are not only used to model the distribution of social income, but also the property of households, market shares for individual enterprises in the industry, and natural resources of individual countries. Lorentz distribution might also be applied to areas other than economic science.

8.3. Statistical evaluation of structural changes in time and space

The structure of any statistical population is dynamic. The composition and technical level of production funds, the age and professions, the structure of employees, the composition and quality of natural resources involved in production, the range and quality of manufactured products, the structure of the consumer budget, and more are changing. The change in the shares of individual components of a population indicates structural changes.

The absolute and relative indices of change of individual parts of the whole are non-proportional to each other: smaller absolute changes may correspond to larger relative changes and larger absolute changes to smaller relative ones. That is why the analysis of changes in the structure of any population should take into account both absolute and relative indicators of changes in structures to obtain a more accurate data of the structural changes of the compared structures [1; 3; 6; 12; 16; 17; 24; 27; 28; 32; 88; 89; 101; 103; 104].

Turning to the summary indicators, there is one thing to be noted. If the total size of the population under study is increasing, the relative changes in the individual elements of the population may be greater or less than one, that is, they may increase and decrease. Moreover, if the relative index of change of an individual element is greater than the relative change in the whole population, it means that the weight of this element in the population increases. Accordingly, if the relative rate of change of any element or part of the population is less than the same indicator of the whole population, the share of this part is reduced in total. Thus, the change in the structure of the whole is a consequence of uneven intensity of change in the individual parts, that is, the differences in the relative change in the shares.

A generalized characteristic of these changes is often required to analyze changes in structure. For this purpose the following indicators can be used: the sum of absolute changes in the shares; the index of structural changes; standard deviation, the linear coefficient of structural shifts.

The sum of absolute changes in the shares [7; 32; 37; 40; 41; 43; 88; 89; 90; 101; 103; 104] is calculated as follows:

$$A = \sum_{i=1}^{n} \left| d_{i1} - d_{i0} \right|, \tag{8.20}$$

where d_{i1} is the share of the indicator in the structure in the reporting year;

d_{i0} is the share of the indicator in the structure in the base year;

n is the number of elements (groups) in the population.

The sum of absolute changes in the shares is expressed in percentage points. This value characterizes the total deviation of one structure from another.

The structural change index. A different parameter is used to ease the assessment [1; 3; 6; 12; 16; 17; 37; 40; 41; 43; 88; 89; 90; 101; 103; 104]:

$$I_{\text{str.change}} = \frac{1}{2}\sum_{i=1}^{n}\left|d_{i1} - d_{i0}\right|. \tag{8.21}$$

The structural change index, calculated based on proportion, expressed as a percentage can take values from 0 to 100 %. Approaching zero means no changes, approaching the maximum is the evidence of a significant change in the structure.

Example 8.3. According to Table 8.5 the structure of fuel consumed in the region (in terms of conditional units) has changed: the share of gas and fuel oil has decreased, the share of coal and other kinds of fuel has increased.

Table 8.5

The estimated fuel consumption in the region

Kind of fuel	2013, d_0	2018, d_1	Deviation of shares, $d_1 - d_0$	Deviation modules, $\left\|d_1 - d_0\right\|$	Squares of deviations, $(d_1 - d_0)^2$
Coal	29	42	13	13	169
Gas	23	16	− 7	7	49
Oil fuel	45	36	− 9	9	81
Other kinds	3	6	+ 3	3	9
Total	100	100	0	32	308

Let us evaluate the intensity of structural shifts by means of *linear or standard* deviations of shares [1; 3; 6; 7; 12; 16; 17; 24; 27; 28; 32; 37; 40; 41; 43; 88; 89; 90; 101; 103; 104]:

$$\bar{l}_d = \frac{\sum_{j=1}^{m}\left|d_{j1} - d_{j0}\right|}{n}, \tag{8.22}$$

$$\sigma_d = \sqrt{\frac{\sum_{j=1}^{m}\left(d_{j1} - d_{j0}\right)^2}{n}}, \tag{8.23}$$

where d_{j0} and d_{j1} are the shares in the base and current period, respectively;
n is the number of the population components.

The linear coefficient of structural shifts is $\bar{l}_d = \frac{32}{4} = 8$, that is the shares of individual kinds of fuel have changed on average by 8 percentage points. The quadratic coefficient of structural shifts is somewhat larger due to its mathematical properties [1; 3; 6; 7; 12; 16; 17; 24; 27; 28; 32; 37; 40; 41; 43; 89; 90; 101; 103; 104]:

$$\sigma_d = \sqrt{\frac{308}{4}} = 8.8 \text{ percentage points}.$$

When using these indicators, the analysis of changes in structures is made without taking into account the size of the base from which this change occurred. More accurate estimation can be made through the use of relative changes. In particular, it is possible to calculate the relative average linear coefficient of structural shifts as the average of the relative linear deviations (i.e. growth rates), taken modulo [1; 3; 6; 7; 12; 16; 17; 24; 27; 28; 32; 37; 40; 41; 43; 88; 89; 90; 101; 103; 104]:

$$\bar{l}_{rel.} = \frac{\sum_{i=1}^{n} \left| \frac{d_{i1} - d_{i0}}{d_{i0}} \right|}{n}. \tag{8.24}$$

The most sophisticated analytical properties (compared with linear and rms) are characteristic of the Salai index and the integral coefficient of structural differences of Gatev.

The Gatev index is calculated as [1; 3; 6; 7; 12; 16; 17; 24; 27; 28; 32; 37; 40; 41; 43; 88; 89; 90; 101; 103; 104]:

$$K_{Gatev} = \sqrt{\frac{\sum_{i=1}^{n} (d_{i1} - d_{i0})^2}{\sum_{i=1}^{n} d_{i1}^2 + \sum_{i=1}^{n} d_{i0}^2}}, \tag{8.25}$$

where d_{i1} is the share of the indicator in the reporting year structure;
d_{i0} is the share of the indicator in the base year structure.

The index assumes a higher value if for example batches of goods under study are approximately of the same size. At the same time, the larger is the number of goods and the smaller is their size, the higher is the index value.

The Salai index looks as follows [1; 3; 6; 7; 12; 16; 17; 24; 27; 28; 32; 37; 40; 41; 43; 88; 89; 90; 101; 103; 104]:

$$I_{Salai} = \sqrt{\frac{\sum_{i=1}^{n} \left(\frac{d_{i1} - d_{i0}}{d_{i1} + d_{i0}} \right)^2}{n}} \; . \tag{8.26}$$

The Salai index is very sensitive to the incorrect representation of small parts, that differs markedly from all others. The Salai coefficient takes values close to unity when the sum of units is large.

The Ryabtsev index. The values of this indicator do not depend on the number of structure gradations. Estimates are made on the basis of the maximum possible differences between the components of the structure. Anyway, there is a correlation of the actual differences of the structure individual components with the maximum possible values [1; 3; 6; 7; 12; 16; 17; 24; 27; 28; 32; 37; 40; 41; 43; 88; 89; 90; 101; 103; 104]:

$$K_{Ryabtsev} = \sqrt{\frac{\sum (d_1 - d_0)^2}{\sum (d_1^2 + d_0^2)^2}}. \tag{8.27}$$

Thus, the considered indicators make a generalized characteristic of structural changes, but do not give an idea of the magnitude of these changes.

Important concepts

Algebraic central moment of distribution is the arithmetic mean of the k-th degree of deviation of individual values of a feature from the mean.

Decile (D_1, D_2, ..., D_9) is the value of a feature that divides a population into 10 equal parts.

Consent criteria are special statistics used to obtain an objective assessment of the discrepancy between the empirical and theoretical distribution curves (based on the use of different measures of distances between the empirical and theoretical distribution).

Quantile is the value of a feature that occupies a specific place in the population ordered according to that feature.

The decile differentiation ratio shows the multiplicity of the ratio of the ninth (D_9) and first (D_1) deciles.

The coefficient of differentiation (K_{dif}) is defined as the ratio of two averages obtained from 10 % of the largest and smallest values of the feature under study.

Estimation of differences between two distributions in space and time is the equation of the proportion of the elements' distribution of a population and the proportion of distribution of the feature values.

*Percentil*e is the value of a feature that divides a population into 100 equal parts.

Distribution uniformity is the case when the shares of the population elements' distribution and the shares of the feature values' distribution are the same.

Theoretical distributions are dependencies between the density of distribution and the values of a feature that represent the patterns of distribution.

Centralization is the concentration (aggregation) of the size of a feature of some units (for example capital in individual commercial banks, products of some kind at individual enterprises, etc.).

Typical tasks

Task 1. We have some data on the distribution of adolescents according to the time they spend on computer games (Table 8.6). Determine the first and third quartiles as well as the first and ninth deciles.

Table 8.6

The input data

Groups according to time spent on computer games, hours	Number of teenagers, people	Accumulated frequencies
1 – 3	43	43
3 – 6	56	99
6 – 9	48	147
9 – 12	32	179
12 – 15	18	197
Total	197	

The solution.

Let us calculate the first quartile:

$$Q_1 = 3 + 3 \times \frac{0.25 \times 197 - 43}{56} = 3.33.$$

Let us calculate the third quartile:

$$Q_3 = 9 + 3 \times \frac{0.75 \times 197 - 147}{32} = 9.07.$$

Let us calculate the first decile:

$$D_1 = 1 + 3 \times \frac{0.1 \times 197 - 0}{43} = 2.37.$$

Let us calculate the ninth decile:

$$D_9 = 12 + 3 \times \frac{0.9 \times 197 - 179}{18} = 11.72.$$

So, the first quarter (49 teens) spend up to 3.33 hours playing computer games. The last quarter (49 teens) spend over 9.07 hours. 10 % of teens spend up to 2.37 hours on computer games, 10 % spend over 11.72 hours.

Task 2. The level of profitability of enterprises of the light and food industry is characterized by the data given in Table 8.7.

Table 8.7

The input data

Profitability level, %	% to total	
	Light industry	Food Industry
1	2	3
Up to 5	3	8
5 – 10	8	15
10 – 15	16	21
15 – 20	22	26

Table 8.7 (the end)

1	2	3
20 – 25	24	17
25 – 30	18	9
30 and more	9	4
Total	100	100

Determine the quartile of the level of profitability for each industry, explain their content. Compare the variation, draw conclusions.

The solution.

Quartile and decile calculations are based on cumulative shares: $S_1 = d_1$, $S_2 = d_1 + d_2$, $S_3 = d_1 + d_2 + d_3$, etc. (Table 8.8).

Table 8.8

The cumulative share of the profitability level, %

Profitability level, %	% to total			
	Light industry	The cumulative share (S_d)	Food industry	The cumulative share (S_d)
Up to 5	3	3	8	8
5 – 10	8	11	15	23
10 – 15	16	27	21	44
15 – 20	22	49	26	70
20 – 25	24	73	17	87
25 – 30	18	91	9	96
30 and more	9	100	4	100
Total	100	–	100	–

The first and third quartiles are determined by the formulas:

the first quartile:

$$Q_1 = x_0 + h \times \frac{0.25 \times \sum_i^m f_i - S_{Q1-1}}{f_{Q_1}};$$

the third quartile:

$$Q_3 = x_0 + h \times \frac{0.75 \times \sum_i^m f_i - S_{Q3-1}}{f_{Q3}};$$

a) for the light industry:

the first quartile interval is 10 – 15, as S_3 = 27 > 100/4:

$$Q_1 = 10 + 5 \times \frac{0.25 \times 100 - 11}{16} = 14.38 \text{ \%};$$

the second quartile interval is 20 – 25, as S_5 = 73 > 100/2:

$$Q_2 = 20 + 5 \times \frac{0.5 \times 100 - 49}{16} = 20.2 \text{ \%};$$

the third quartile interval is 25 – 30, as S_6 = 91 > 3/4·100:

$$Q_3 = 25 + 5 \times \frac{0.75 \times 100 - 73}{18} = 25.56 \text{ \%};$$

b) for the food industry:

the first quartile interval: 10 – 15, as S_3= 44 > 100/4:

$$Q_1 = 10 + 5 \times \frac{0.25 \times 100 - 23}{21} = 10.48 \text{ \%};$$

the second quartile interval is 15 – 20, as S_4 = 70 > 100/2:

$$Q_2 = 25 + 5 \times \frac{0.5 \times 100 - 44}{26} = 16.5 \text{ \%};$$

the third quartile interval: 20 – 25, because S_5 = 87 > 3/4·100:

$$Q_3 = 20 + 5 \times \frac{0.75 \times 100 - 70}{17} = 21.47 \text{ \%}.$$

In the light industry, a quarter of all enterprises have a profitability level of up to 14.38 %; half – up to 20.2 %; 25 % of all enterprises have the highest level of profitability; the lowest level belongs to 25.56 %. In the food industry, a quarter of all enterprises have profitability level up to 10.48 %, half – up to 16.5 %; 75 % – up to 21.47 %.

The distribution center is given by the median, so we use the quartile coefficient of variation:

$$V_Q = \frac{Q_3 - Q_1}{2 \times Me}.$$

For the light industry:

$$V_Q = \frac{25.56 - 14.38}{2 \times 20.2} = 0.29\ \%.$$

For the food industry:

$$V_Q = \frac{21.47 - 10.48}{2 \times 16.5} = 0.48\ \%.$$

The food industry has a greater variation.

Task 3. Analize the concentration and localization of the distributions according to the data (Table 8.9).

Table 8.9

The input data

Type of economic activity	Number of enterprises based on the type of activity	Fixed capital investment, million UAH
Agriculture, hunting, forestry	14 736	5 199.3
Mining industry	12 983	8 580.8
Manufacturing industry	98 230	21 811.5
Production and distribution of electricity, gas and water	2 654	5 158.7
Construction	35 478	2 991.1
Trade, repair of cars, household goods and personal items	115 231	7 737.1
Transport and communication activities	35 621	14 790.5
Education	15 112	665.3
Health care and social assistance	23 427	907.2
Other activities	27 317	19 114.1
Total	380 789	86 955.6

The solution.

It is necessary to build a calculation table (Table 8.10) to calculate the coefficients of localization and concentration.

Table 8.10

The calculation table

Type of economic activity	Number of enterprises based on the type of activity	Fixed capital investment, million UAH	d_j	D_j	K_L
Agriculture, hunting, forestry	14 736	5 199.3	3.87	5.98	1.55
Mining industry	12 983	8 580.8	3.41	9.87	2.89
Manufacturing industry	98 230	21 811.5	25.80	25.08	0.97
Production and distribution of electricity, gas and water	2 654	5 158.7	0.70	5.93	8.51
Construction	35 478	2 991.1	9.32	3.44	0.37
Trade, repair of cars, household goods and personal items	115 231	7 737.1	30.26	8.90	0.29
Transport and communication activities	35 621	14 790.5	9.35	17.01	1.82
Education	15 112	665.3	3.97	0.77	0.19
Health care and social assistance	23 427	907.2	6.15	1.04	0.17
Other activities	27 317	19 114.1	7.17	21.98	3.06
Total	380 789	86 955.6	100	100	

The localization factor is calculated by the formula $K_L = \frac{D_j}{d_j}$. The results of the calculations are given in the last column of the spreadsheet. The maximum localization of investments in fixed capital is observed in the field of production and distribution of electricity, gas and water.

Calculate the concentration factor:

$$K_C = \frac{1}{2}\sum | D_j - d_j | = \frac{1}{2}(| \quad 5.98 - 3.87 | + | 9.87 - 3.41 | + | 25.08 - 25.8 | +$$

$$+ | 5.93 - 0.7 | + | 3.44 - 9.32 | + | 8.90 - 30.26 | + | 17.01 - 9.53 | +$$

$$+ | 0.77 - 3.97 | + | 1.04 - 6.15 | + | 21.98 - 7.17 | \ ) = 36.27.$$

The obtained value of the concentration coefficient indicates a high concentration of investments according to the type of economic activity.

Task 4. Carry out the structure similarity analysis of the first group students' performance compared to the others (Table 8.11).

Table 8.11

The input data

Progress	Group 1, %	Group 2, %	Group 3, %	Group 4, %
Perfect	12	16	8	4
Good	44	54	36	30
Satisfactory	40	20	42	48
Unsatisfactory	4	10	14	18

The solution.

The coefficient of the structure similarity is calculated by the formula:

$$К_{sim} = 100 - \frac{1}{2}\sum\left|d_{jk} - d_{js}\right|.$$

We calculate the structure similarity coefficient of the second group in relation to the first one:

$$К_{\underset{21}{sim}} = 100 - \frac{1}{2}\sum\left(|16-12| + |54-44| + |20-40| + |10-4|\right) = 80\ \%.$$

We calculate the structure similarity coefficient of the third group in relation to the first one:

$$К_{\underset{21}{sim}} = 100 - \frac{1}{2}\sum\left(|8-12| + |36-44| + |42-40| + |14-4|\right) = 88\ \%.$$

We calculate the structure similarity coefficient of the fourth group in relation to the first one:

$$К_{\underset{21}{sim}} = 100 - \frac{1}{2}\sum\left(|4-12| + |30-44| + |48-40| + |18-4|\right) = 78\ \%.$$

The results of the structure similarity analysis show that the success of students of the third group is the most similar to the first group's success.

Task 5. Determine the quadratic coefficient of structural shifts according to the age composition of customs cleared cars and trucks (Table 8.12), perform a comparative analysis.

Table 8.12

The input data

Service life, years	The structure of customs cleared vehicles, %			
	cars		trucks	
	2012	2019	2012	2019
Up to 10	60	45	56	51
10 – 20	26	34	24	30
20 and older	14	21	20	19
Total	100	100	100	100

The solution.

the quadratic coefficient of structural shifts is $I_{Salai} = \sqrt{\dfrac{\sum\limits_{i=1}^{n}\left(\dfrac{d_{i1}-d_{i0}}{d_{i1}+d_{i0}}\right)^2}{n}}$:

1) for cars:

$$\sigma_d = \sqrt{\frac{(45-60)^2+(34-24)^2+(21-14)^2}{3}} = \sqrt{\frac{338}{3}} = 10.61;$$

2) for trucks:

$$\sigma_d = \sqrt{\frac{(51-56)^2+(30-24)^2+(19-20)^2}{3}} = \sqrt{\frac{62}{3}} = 4.55.$$

The shifts in the structure of cars are more than twice the changes in the structure of trucks over the specified period. That is, in the structure of customs cleared cars, the "younger" cars twice prevail the rest.

Reference to laboratory work

The guidelines for performing laboratory work on the topic "Analysis of concentration, differentiation and similarity of distributions" are presented in [100]. The purpose of the laboratory work is to master students' skills in the analysis of concentration, differentiation and similarity of statistical distributions by means of the MS Excel package.

Questions for self-assessment

1. What are the ordinal characteristics of a distribution?

2. What indicators characterize the degree of homogeneity of a population, asymmetry and excess of distribution?

3. In what ways can the asymmetry of distribution be estimated?

4. How is the asymmetry coefficient calculated and characterized?

5. In what ways can the excess of distribution be estimated?

6. What is the measure of concentration and how is it calculated?

7. How is the similarity of distributions determined?

8. What indicators allow you to estimate the intensity of structural shifts?

9. How are theoretical frequencies determined?

10. What harmonization criteria do you know?

Questions for critical rethinking (essays)

1. Describe the advantages and disadvantages of using the Gini coefficient in the study of differentiation of a population.

2. What economic content is included in the calculation of indicators of similarity of distributions?

3. Prove or refute the statement that the structural shift index can describe qualitative structural changes.

4. How will enhancement of left or right asymmetry of distribution affect the development of the phenomenon or process under study?

5. Explain what happens to a phenomenon or process in a peaked distribution.

Section 3
Methods for analysis of interrelations of phenomena and processes

9. Statistical methods for measuring interrelations

Basic questions:

9.1. The concept and types of relationships in statistics.

9.2. The model of analytical grouping.

9.3. The regression equation and determination of its parameters.

9.4. The indicators of the closeness and significance of the correlation relationship.

9.5. The development of multiple correlation and regression models.

9.6. Methods for studying the relationship of social phenomena.

9.7. Nonparametric methods.

9.1. The concept and types of relationships in statistics

Among many forms of relationships, the most important one is causal, that is, defining all other forms. The essence of causality is the origination of one phenomenon from another. However, the cause as such does not yet recognize the consequence, it also depends on the conditions in which the effect of the cause takes place.

Causation is the relationship of phenomena and processes when a change in one of them – the cause – leads to a change in the other – the consequence. The feature that characterizes the consequence is a *dependent variable* while the feature that characterizes the cause is an *independent variable.*

According to statistical nature, the relationships are divided into functional and stochastic. With *the functional relationship*, each value of an independent variable corresponds to one well-defined value of the dependent variable. Functional relationships are usually expressed by formulas. Most often such relationships are observed in mathematics, physics, chemistry. Functional relationships also occur in economic processes, but rarely, since they show the interrelations of only certain aspects of complex phenomena of social life. For example, labor productivity is the ratio of the output and the number of employees.

Functional relationships can be represented as the equation:

$$y_i = f(x_i), \tag{9.1}$$

where y_i is the dependent variable (i = 1, …, n);

$f(x_i)$ is a known function of the dependent and independent variables;

x_i is the independent variable.

Functional dependence with equal force is manifested in all units of a population, regardless of the change in other variables of a phenomenon. But in the mass phenomena of social life due to a variety of factors there is a wide variation of the dependent variable. This indicates that the relationship between the independent variable and the dependent variable is not complete, but manifests itself only in general, average form. Such relationships are called stochastic.

In *a stochastic relationship*, each value of x is matched by a set of values of y that vary and form a series of distribution. An important peculiarity of stochastic relations is that they are not observed in isolated cases but in mass. Mass observations, that is, statistics, are required to study them.

The stochastic relationship model can be represented in general form by the equation:

$$\tilde{y}_i = f(x_i) + \varepsilon_i, \tag{9.2}$$

where $\tilde{y}_i$ is the estimated value of the dependent variable;

$f(x_i)$ is part of the dependent variable that is formed under the influence of the known and taken into account independent variable (one or more) that are in a stochastic relationship with the dependent variable;

ε_i is part of the dependent variable that appeared as a result of unaccounted factors.

A subtype of stochastic relationship is *correlation relationship* when a change in the independent variable x causes changes in the group mean values of the dependent variable y, that is, instead of conditional distributions, the mean values of these distributions are compared.

An example of a stochastic, including correlation relationship is the distribution of production sites according to the production spoilage (y) and the percentage of violations of the technological discipline (x) (Table 9.1).

Table 9.1

The distribution of production sites based on the variables

Percentage of violation of the technological discipline, x	Number of sites with losses from spoilage of products, thousand UAH						Average losses from spoilage, thousand UAH, $\bar{y}_i$
	0.9 – 1.14	1.14 – 1.38	1.38 – 1.62	1.62 – 1.86	1.86 – 2.1	Total, f_i	
1.2 – 1.575	2	3	1	–	–	6	1.22
1.575 – 1.95	–	–	6	–	–	6	1.5
1.95 – 2.325	–	–	2	3	–	5	1.644
2.325 – 2.7	–	–	–	–	5	5	1.98
Total	2	3	9	3	5	22	1.565

The data shown in Table 9.1 make a correlation table that presents the grouping based on two interrelated variables: independent and dependent. Concentration of frequencies around the diagonal of the matrix indicates that there is a correlation relationship between the variables. As can be seen from the data in Table 9.1, the distribution of production sites is observed from the upper left corner of the matrix to the lower right corner. The nature of the concentration of diagonal frequencies indicates that there is a direct correlation relationship between the variables under consideration.

Each group, on the independent variable basis, corresponds to a distribution y, which differs from other groups and from the unconditional total distribution. Therefore, there is a stochastic relationship between the variables.

Conditional distributions can be replaced by the mean values of the dependent variable which are calculated as the weighted arithmetic mean:

$$\bar{y}_i = \frac{\sum_{i=1}^{m} y_i f_i}{\sum_{i=1}^{m} f_i}. \qquad (9.3)$$

The gradual change of the mean values of $\bar{y}_i$ from one group to another indicates that there is a correlation relationship between the variables.

The *correlation relationship* is the concept that is narrower than the stochastic relationship. The latter can be shown not only in the change in the mean, but also in the variation of one variable depending on the other, that is, any characteristic of the variation. Thus, the correlation relationship is a particular case of the stochastic relationship.

A correlation between the variables may result from:

the causal dependence of the dependent variable (regressor) on the variation of the independent variable;

a conjunction that arises in the presence of a common cause. For example, a well-known statistician O. Chuprov [71] found that if the number of fire brigades in the city is taken as a variable x, and the amount of fire damage per year as a variable y, there is a direct correlation between these variables: on average the more firefighters in the city, the greater the damage caused by fires. But this fact is not natural. Such a correlation cannot be interpreted as a cause and effect relationship. In this case, both variables are the consequence of a common cause – the size of the city. It is clear that in large cities there are more fire brigades, and more fires and, consequently, yearly losses from them are bigger than in small cities;

the interrelation of the variables, each of which is both the cause and the effect. For example, the correlation between the levels of labor productivity and the level of remuneration per 1 hour of work: on the one hand, the higher the productivity, the higher the remuneration, and on the other, remuneration plays a stimulating role for increasing labor productivity. Therefore, in such a system of variables, each acts as an independent variable x and a dependent variable y.

Depending on the direction of action, functional and stochastic relationships may be direct and inverse. In *a direct relationship*, the direction of change in the dependent variable coincides with the direction of change in the independent variable, that is, with the increase of the independent variable the dependent variable also increases and vice versa. Otherwise, there are *inverse relationships* between the values under consideration. For example, the longer the service, the higher the salary – a direct relationship. And the higher the cost of production, the lower the profit – an inverse relationship.

In terms of analytical expression relationships may be straight and curvilinear. If the statistical relationship between phenomena can be expressed by the equation of a straight line, it is called *a linear relation*. If, however, it is expressed by the equation of any curved line (parabola, hyperbola, power, exponential, etc.), this relation is *nonlinear* or *curvilinear*.

It should be noted that only functional relationships are accurately expressed in the analytical equation, while correlation – only approximately, provided we abstract from the influence of all other factors. Therefore, a graph of correlation relationship shows a spread of dots (y and x) around the theoretical line.

Depending on the number of variables under study, pair (simple) and multiple correlations are distinguished. In *a pair correlation*, the relationship between two variables (dependent and independent) is studied, in *a multiple correlation*, the relationship between three and more variables (a dependent variable and two and more independent variables) is investigated. In the case of multifactor relationship, all the factors work together, that is, simultaneously.

9.2. The model of analytical grouping

When building an analytic grouping, it is necessary to create such a number of groups, in which the variation of the group averages will maximize the influence of the grouping variable. As a general rule, the greater the number of groups formed, the greater the intergroup variation. However, it is not advisable to have a large number of groups, especially with a small number of units in the population: in this case the groups will be small, the average of them will be random, the intergroup variation will show not only the influence of the considered factors, but other factors as well. This means that it is necessary to choose the optimal number of groups for each case, when the group averages will be of no accidental nature and will most fully reveal the grouping characteristic.

A characteristic of correlation is *the regression line*, which is considered in two models: analytical grouping and regression analysis.

In the analytic grouping model, it is an empirical regression line formed from the group mean values of the dependent variable $\bar{y}_i$ for each value (interval) x_i. In other words, if on the correlation field you connect the point by segments of a straight line, you will get a broken line with some trend called the *empirical regression line*.

The method of analytical grouping implies that all elements of a set are grouped according to the independent variable x; and in each group, the mean values of the dependent variable y are calculated, that is, the regression line is evaluated only at individual points that correspond to a certain value x.

Analytical grouping allows you to establish quantitative relationships between the characteristics being investigated. The effects of x on y are defined as the ratio of the increments of the mean group values $\Delta y : \Delta x$, where $\Delta y = \bar{y}_i - \bar{y}_{i-1}$, $\Delta x = x'_i - x'_{i-1}$. According to Table 9.1, increments of Δx are the same in all groups making 0.375 %, and the average losses from product spoilage increase in groups as follows: $\Delta y_2 = 1.5 - 1.22 = 0.28$ thousand UAH; $\Delta y_3 = 0.144$ thousand UAH; $\Delta y_4 = 0.336$ thousand UAH. So with the increasing percentage of the technological discipline violations by 1 %, losses from product spoilage increase on average by $\Delta y_2 : \Delta x_2 = 0.28 : 0.375 = 0.74$ thousand UAH and by 0.384 and 0.896 thousand UAH correspondingly.

The next step in analytic grouping is to *measure relationship tightness* which is based on the dispersion sum rule. In the analytic grouping model, the measure of the relationship tightness is the ratio of intergroup variance to the total, that is, the empirical coefficient of determination:

$$\eta^2 = \frac{\sigma_M^2}{\sigma^2}, \tag{9.4}$$

where σ^2 is the total variance that characterizes the variation in the numerical values of the dependent variable associated with the variation of all factors affecting it;

σ_M^2 is the intergroup variance that characterizes the variation in the numerical values of the dependent variable that is related to the variation of the independent variable.

The coefficient of determination varies from 0 to 1. If $\eta^2 = 0$, the intergroup variance is zero. This is possible only if all group averages are the same and there is no correlation between the variables.

If $\eta^2 = 1$, the intergroup variance is equal to the total one, the average of the group variance is zero. In this case, each value of the independent variable corresponds to a single value of the variable, that is, the relationship between the variables is functional.

It should be noted that the value $\eta^2 > 0$ is not always the evidence of a correlation between the variables. A non-zero coefficient of determination may also occur if the population is randomly divided into groups.

According to Table 9.2, the total variance of the loss from the product spoilage is:

$$\sigma^2 = \overline{y}^2 - (\overline{y})^2 = \frac{\Sigma(y'_i)^2 f_i}{\Sigma f_i} - \left(\frac{\Sigma y'_i f_i}{\Sigma f_i}\right)^2 =$$

$$= (1.02^2 \times 2 + 1.26^2 \times 3 + 1.5^2 \times 9 + 1.74^2 \times 3 + 1.98^2 \times 5)/22 - 1.565^2 = 0.086.$$

Table 9.2 provides an analytical grouping of production sites, which describes the dependence of the spoilage of products on these sites on the percentage of violations of the technological discipline. The calculation of intergroup variance is also presented there.

Table 9.2

Analytical grouping

Violations of the technological discipline, %	The number of production sites, f_i	Average losses from the spoilage of products, thousand UAH, $\overline{y}_i$	$\overline{y}_i - \overline{y}$	$(\overline{y}_i - \overline{y})^2 f_i$
1.2 – 1.575	6	1.220	−0.345	0.714
1.575 – 1.95	6	1.500	−0.065	0.025
1.95 – 2.325	5	1.644	0.079	0.031
2.325 – 2.7	5	1.98	0.415	0.861
Total	22	1.565	–	1.631

The intergroup variance is equal to:

$$\sigma_M^2 = \frac{\sum(\overline{y}_i - \overline{y})^2 f_i}{\sum f_i} = \frac{1.631}{22} = 0.074.$$

The coefficient of determination is:

$$\eta^2 = \frac{0.074}{0.086} = 0.86,$$

that is, a 86 % variation of losses from product spoilage is due to a variation in the percentage of the technological discipline violations, and a 14 % variation results from other factors. Therefore, the relationship between the variables is very strong.

A strong relationship between the variables may occur by chance, so it is necessary to check its significance.

The significance of the relationship is verified using the criteria of mathematical statistics. It is based on the comparison of the actual value η^2 with the critical $\eta^2_{1-2}(k_1, k_2)$ for a certain level of significance α and the number of degrees of freedom $k_1 = m - 1$ and $k_2 = n - m$, where m is the number of groups; n is the size of the population.

The critical value η^2 is the maximum value of the coefficient of determination that can occur accidentally in the absence of correlation.

If $\eta^2 > \eta^2_{1-\alpha}(k_1, k_2)$, the relationship between the dependent and independent variables is considered significant. If the actual value η^2 is less than the critical one, there is no correlation relationship between the variables, and the relationship is considered insignificant.

The critical value is chosen so that the probability of obtaining the value η^2 larger than the critical one (provided there is no correlation between the variables) is sufficiently small. This probability is called *the significance level* α. Most often, in economic and statistical studies, the significance levels α = 0.05 and α = 0.01 are used. The critical values of the correlation ratio for α = 0.05 are given in Table (Annex A).

In our example k_1 = 4 − 1 = 3, k_2 = 22 − 4 = 18. The critical value of these degrees of freedom for the significance level α = 0.05 is $\eta^2_{0.95}(3.18) = 0.345$.

Because $\eta^2 = 0.86 > 0.345$, the relationship is considered essential with probability 95 %.

To verify the significance of the relationship they also use a characteristic functionally linked with η^2, F-test (Fisher test) which is calculated by the formula:

$$F = \frac{\eta^2}{1 - \eta^2} \times \frac{k_2}{k_1}. \tag{9.5}$$

The critical values of the F-criterion for different significance levels α are given in Annex B and Annex C.

In our example, the actual value of F is:

$$F = \frac{0.86}{1-0.86} \times \frac{18}{3} = 36.85,$$

that is much more than the critical $F_{0.95}(3.18) = 3.16$. This indicates that the correlation between the analyzed variables is significant.

Combinational analytical groping is used to analyze the relationship of the dependent variable with two or more independent variables. It helps to study the dependence of the dependent variable on each of the factors, provided the fixed values of other variables. Methods for measuring this type of relationship and verification of its significance are called *multivariate dispersion complexes*.

The disadvantage of the method of analytical grouping is that it does not allow us to determine the form (analytical expression) of the influence of independent variables on the dependent ones.

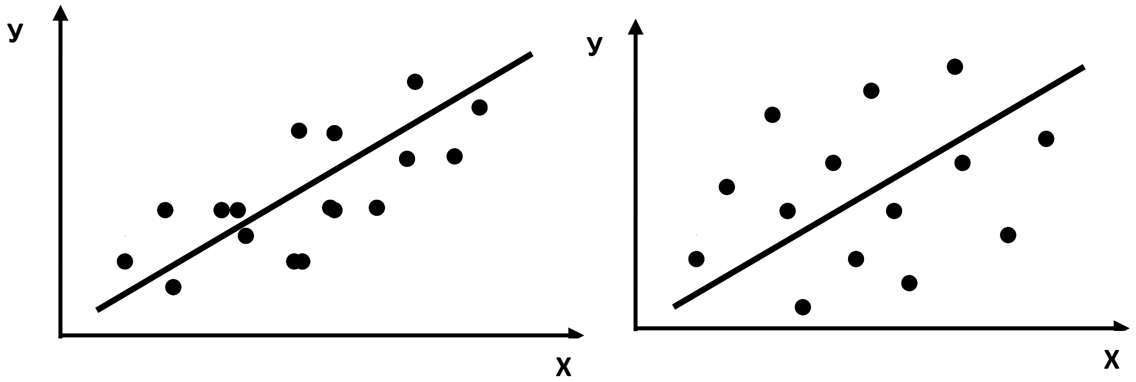
In general, the task of statistics in the study of interrelationships is not only to quantify their presence, direction and strength of relationship, but also to determine the form (analytical expression) of the influence of independent variables on the dependent ones. The method of the correlation and regression analysis is used for this purpose.

The correlation and regression analysis as a general concept includes measurement of tightness, direction of relationship and establishment of analytical expression (form) of the relationship (regression analysis).

The purpose of correlation analysis is to quantify the tightness of the relationship between two variables (paired relationship) and between the dependent and multiple independent variables (multifactor relationship). The relationship tightness is quantified by the magnitude of the correlation coefficients. Such coefficients make it possible to determine the "utility" of the independent variables in building multiple regression equations. The magnitude of the correlation coefficient also serves an estimate of the correspondence of the regression equation to the identified cause and effect relationships.

The purpose of regression analysis is to determine the analytic expression of the relationship, to determine the degree of influence of independent variables on the dependent one and to determine the calculated values of the dependent variable (the regression function).

Let us explain the difference between correlation and regression with the help of Fig. 9.1.



a) close correlation b) weak correlation

Fig. 9.1. **Regression with different intensities of correlation**

As can be seen from Fig. 9.1, the angle of inclination of the regression line with respect to the abscissa axis is the same for parts a and b. But on graph a the points of the correlation field are concentrated around the regression line, while on graph b the points of the correlation field are scattered. This means that the relationship is close, that is, the degree of correlation between x and y would be high in case a and low in case b. Therefore, the regression equation in case a will be statistically significant while in case b it will not be statistically significant. Thus, cases a and b differ in the magnitude of the correlation coefficients ($r_{yxa} \neq r_{yxb}$); at the same time, they have the same regression coefficients ($b_{yxa} = b_{yxb}$).

In statistics, there are the following *dependency options*:

pair correlation (regression), that is the relationship between two variables (dependent and independed or two independent variables);

partial correlation (regression), that is the relationship between the dependent variable and one of the independent variables provided the fixed value of the other independent variables;

multiple correlation (regression), that is dependence of the dependent and two or more independent variables included in the study.

In order for the results of the correlation and regression analysis to be of practical use and to produce scientifically valid results, certain requirements must be fulfilled with respect to the object of study and the quality of the input statistical information. The main *requirements* are as follows:

a sufficient number of observations, since the relationship between the variables is only revealed as a result of the law of large numbers. The number of observation units should be 6 – 10 times greater than the number of factors included in the model;

qualitative homogeneity of the population under study which implies the closeness of the formation of dependent and independent variables. The necessity of fulfilling this condition stems from the content of the relationship equation parameters, which are average values. In a qualitatively homogeneous population they will be typical characteristics, in a qualitatively heterogeneous one they will be distorted, changing the nature of the relationship. The quantitative homogeneity of a population consists in the absence of observation units which in their numerical characteristics are significantly different from the bulk of the data. Such observation units should be excluded from the population and examined separately;

the randomness and independence of individual units of a population which means that the values of the variables of some units of the population must not depend on the values of other units of the same population;

stability and independence of individual factors;

persistence of the dispersion of the dependent variable with the change in independent variables;

normal distribution of variables.

If the size of the population under study is large enough ($n > 50$), the normality of distribution can be confirmed on the basis of calculation and analysis of the criteria of Pearson, Yastremsky, Boyarsky, Kolmogorov and others. If $n < 50$, the law of distribution of the initial data is determined on the basis of construction and visual analysis of the correlation field. At the same time, if there is a linear trend in the location of the points, we can assume that the totality of the input data $(y, x_1, x_2, \dots x_n)$ is subject to normal distribution.

When using the correlation and regression methods, it should be remembered that the correlation relationship equation measures the dependence between variations of dependent and independent variables. The relationship tightness shows the proportion of variation of the dependent variable that is related to the variation of the independent one(s). It is necessary to interpret correlation indices in terms of variation of deviations from the mean. If the purpose of the study is not to measure the relationship between the variation of two variables in the population but between the changes in the variables of the object over time, the method of correlation and regression analysis requires some change. This means that one cannot interpret the correlation of variables as the causal relationship of their levels. In other words, the method of correlation and regression analysis cannot explain the role of the independent variable in creating a dependent variable. This is a limitation of this method.

In the development of a correlation model, the selection of factors is essential, so *the following recommendations* should be followed*:*

1) the regression equation should include only those factors that are essential to the study of the problem. Therefore, mathematical modeling should be preceded by theoretical analysis. The discrepancy between qualitative and quantitative analysis leads to an erroneous correlation of the first kind. If the prediction regarding the necessary properties of the distribution of the source data is broken, the second kind of error correlation occurs. With the identification of false correlation, it is necessary to expand the range of issues under consideration;

2) the factors used in the regression analysis should be measured quantitatively, that is, quantitative rather than qualitative variables should be used. Correlation and regression analysis is a method of exploring the quantitative aspect of relationships, so each variable must have a quantitative (metric) dimension. If attributive factors cannot but be taken into account in a particular problem, the population is first grouped according to attributive variables. This kind of grouping is based on expert judgment or taxonomy. The difference between the selected groups is checked by analysis of variance. If quantitative regularities are to be investigated within each group, regression analysis is performed for each group individually or covariance analysis is applied.

Artificial presentation of attributive variables in the quantitative form by ranking providing different codes and conventional signs does not lead to satisfactory results;

3) each factor in the multiple regression equation must be presented by only one attribute (natural or cost, absolute or relative, but not both). Failure to do so leads to models that have no economic interpretation.

The selection of the variable that best represents the quantitative magnitude of the real factor is based on the consideration of pairwise relationship of the dependent and all possible independent variables. This analysis is performed using the correlation matrix of pairwise relationships.

It is not possible to use aggregate variables and their components at the same time in the regression equation.

If the relationship is sufficiently close and stable, analytical factors can be used to increase the model's specificity. If the relationship is not sufficiently tight and stable, the total (synthetic) index is preferred. With the transition from analytic to synthetic parameters the relationship tightness and stability increase due to the loss of concreteness (this is a manifestation of the law of large numbers).

If the model must include a synthetic indicator and part of it (an analytical indicator), the value of the synthetic indicator is reduced by the value of the analytical one included in the model. But such technique may disrupt the multicollinearity of factors which is acceptable;

4) it is recommended that the regression equation should include only those factors that are in the same cause and effect chain. This is one of the most important requirements of multivariate modeling. If it is not met, the regression coefficients are not economically viable;

5) if the model uses relative values as variables, it is necessary that the denominator is the same indicator or comparable indicators (the same bases). If this requirement is violated, the model parameters are more difficult to interpret economically;

6) the factors that are included in the model must be decorrelated, that is, be in poor correlation with each other. Two factors are considered collinear if the pairwise correlation coefficient is more than 0.8 in the absolute value. This condition applies when the population is medium in size, and the correlation of individual factors and the dependent variable is close. In the general case, the boundary of statistical collinearity does not only depend on the close relationship of the factors, but also on the association of the dependent variable with the independent variables. For practical purposes, factors can be assumed to have no statistical collinear relationship if the pairwise correlation coefficients meet the following requirements: $r_{oi} > r_{ij}$; $r_{oj} > r_{ij}$. Here 0 is the dependent variable number; i, j are independent variable numbers.

If these inequalities (or one of them) are not fulfilled or close to equality, it is recommended that only one of the two factors under consideration (with a higher pairwise coefficient of correlation with the dependent variable) be used for multivariate analysis. A synthetic indicator that presents the complex of both (or more) multicollinear factors is preferable;

7) it is necessary to follow the logical requirements that are considered in the selection of indicators for regression analysis (for example, those that

meet the requirements to analytical grouping based on the independent rather than dependent variable;

8) a minimal but sufficient number of factors that determine the average value of a dependent variable should be used. Increasing the number of factors to obtain a more meaningful and concrete model leads to a decrease in its sustainability and complicates economic interpretation. Therefore, it is not recommended that minor factors be included in the model. This requirement is based on the principle of simplicity. In regression analysis, this principle is implemented through the algorithm of multistep exclusion of variables; it is also taken into account when choosing the form of relationship;

9) the factors included in the model should be accessible and plausible. Factors should not be introduced if data on them cannot be obtained from reporting, but only through specially organized observation, which is time-consuming and costly.

Problems that are solved by the correlation and regression analysis:

1) selection of the most important factors affecting the dependent variable (i.e. the variation of its values in the population). This problem is solved on the basis of strength of tightness of relationship of factors with the dependent variable;

2) evaluation of economic activity based on the efficiency of the use of the factors of production. This problem is solved by calculation (for each unit of the population) of the values of the dependent variable that would be obtained from the average in the population efficiency of the use of factors in comparison with their actual production results;

3) prediction of possible values of the dependent variable based on the given values of independent variables. This problem is solved by substituting the expected values of the independent variables in the equation and calculating the expected values of the dependent variable;

To solve the inverse problem – to calculate the necessary values of the independent variables to provide the necessary value of the dependent variable – the method of solving an optimization problem to find the best possible option is used;

4) preparing the data you need as input for solving optimization problems. Indicators that are involved in solving an optimization problem can be derived from a correlation and regression model.

The use of the mentioned problems will help to form an algorithm for conducting correlation and regression analysis in the study of different phenomena and processes.

9.3. The regression equation and determination of its parameters

In the model of regression analysis, the characteristic of the correlation relationship is *the theoretical regression line*, described by the function y = f (x), called *the regression equation*. In contrast to the empirical line, the theoretical regression line is continuous.

Different phenomena respond differently to changing factors. In order to present the characteristic features of the connection of specific phenomena, statistics uses different functional regression equations. If, with the change of the factor x, the result y changes more or less uniformly, such a relation is described by a linear function $Y = a + bx$*.* With nonuniform correlation of variations in interrelated variables (for example, when increments in the values of y with the change of x are accelerated or decelerated or the direction of the relationship changes) nonlinear regressions are used, in particular:

$$\text{power:}\, Y = ax^b; \tag{9.6}$$

$$\text{hyperbola: } Y = a + \frac{b}{x}; \tag{9.7}$$

$$\text{parabola:}\, Y = a + bx + cx^2. \tag{9.8}$$

The choice and justification of the functional type of regression are based on a theoretical analysis of the nature of the relationship. For example, the relationship between the yield and rainfall is represented as a parabola, and the relationship between the cost y and the output x is the equation of hyperbola, where a is the proportional cost per unit of production; b is the fixed cost of the whole output.

If the regression curvature is small, then within the actual variation of the variables, the relationship between them is fairly accurately described by a linear function. This largely explains the widespread use of linear regression equations: $Y = a + bx$.

Parameter b (*the regression coefficient*) is a named value; it has the dimension of the dependent variable and is considered as the effect of x on y. It shows the average value of change in the dependent variable y with the change in the independent variable x per unit of measurement, that is, the variation of y per unit of variation of x. The sign b indicates the direction of this change.

Example 9.1. We have a regression equation $Y = 4.5 + 0.8x$, characterizing the dependence of labor productivity Y on work experience x. Because the parameter b = 0.8 > 0, with increasing experience, productivity also increases. An increase in work experience by 1 year leads to a growth of labor productivity by 0.6 units on average.

The parameter a is a free term of the regression equation; it is the value of y if $x = 0$. If the limits of variation x do not contain zero, this parameter is only a calculated value. The characteristic of the relative change in y due to x is the coefficient of elasticity:

$$K_{el} = b\frac{\bar{x}}{\bar{y}}, \tag{9.9}$$

which shows how much on average the percentage of the dependent variable changes with the change in the independent variable by 1 %.

It is mathematically proved that the values of the parameters a and b, given a linear dependence, are determined through the system of normal equations:

$$\begin{gathered} \sum y = na + b\Sigma x, \\ \sum xy = a\Sigma x + b\Sigma x^2. \end{gathered} \tag{9.10}$$

When resolving this system, the following parameter values are found:

$$\begin{gathered} b = \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - \Sigma x\Sigma x}; \\ a = \bar{y} - b\bar{x}, \end{gathered} \tag{9.11}$$

where y is the actual series level (the dependent variable);

n is the number of members in the series;

x is the independent variable.

The regression equation does not represent the law of relationship between x and y for individual elements of the population, but for the population as a whole; a law that abstracts the influence of other factors and proceeds from the principle "other conditions being equal".

The influence of factors other than x causes the rejection of empirical values y from theoretical $\tilde{y}$ one way or the other. Deviations $(y-\tilde{y})$ are called *residuals* and denoted by the symbol "ε".

The total variance is calculated by the formula:

$$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y-\bar{y})^2, \tag{9.12}$$

and the residual:

$$\sigma_\varepsilon^2 = \frac{1}{n}\sum_{i=1}^{n}(y-\tilde{y})^2. \tag{9.13}$$

In small populations, the regression coefficient is prone to random fluctuations. Therefore, it is necessary to check its significance. In a linear relationship, the significance of the regression coefficient is checked using the *Student's t-test*:

$$t = \frac{|b|}{\mu_b}, \tag{9.14}$$

where b is the regression coefficient;

μ_b is the standard error.

The standard error of the regression coefficient depends on the variation of the independent variable σ_x^2, residual dispersion σ_ε^2 and the number of degrees of freedom k = n − 2:

$$\mu_b = \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_x^2(n-2)}}, \tag{9.15}$$

where $\sigma_x^2 = \overline{x^2} - (\bar{x})^2$, ($\overline{\overline{x^2}} = \frac{\sum x^2 \times f}{\sum f}$; $(\bar{x})^2 = \left(\frac{\sum x \times f}{\sum f}\right)^2$).

If t > $t_{cr(1-\alpha)}(k)$, the effect of the variable x on the variable y is considered significant.

Confidence limits are also determined for the regression coefficient $b \pm t\mu_b$.

Example 9.2. There are some data on the assets and deposits of ten commercial banks (Table 9.3). A correlation and regression analysis of the relationship between these variables is required. To characterize the given correlation, you have to define: the form and mathematical model of the relationship, the parameters of the regression equation, the closeness of the relationship.

Table 9.3

A spreadsheet for determining the parameters of a regression equation

Bank number	The name of the bank	The amount of deposits, billion units, x	The amount of assets, billion units, y	Estimated data				
				x^2	y^2	x × y	$\tilde{y}$	$(y_i - \tilde{y})^2$
1	A	32	81	1024	6561	2592	80.614	0.1489
2	B	25	75	625	5625	1875	76.043	1.0878
3	C	28	79	784	6241	2212	78.002	0.9960
4	D	18	73	324	5329	1314	71.472	2.3347
5	E	30	80	900	6400	2400	79.308	0.4788
6	F	24	75	576	5625	1800	75.390	0.1521
7	G	29	83	841	6889	2407	78.655	18.879
8	H	17	69	289	4761	1173	70.819	3.3087
9	I	24	73	576	5329	1752	75.390	5.7121
10	J	18	70	324	4900	1260	71.472	2.1667
Total		246	758	6263	57660	18785	758.00	35.2648
On average		24,6	75,8	626.3	5766	1878.5	75.800	3.5264

To determine the form of correlation between the sum of assets (y) and the amount of deposits (x), we plot a correlation field (Fig. 9.2).

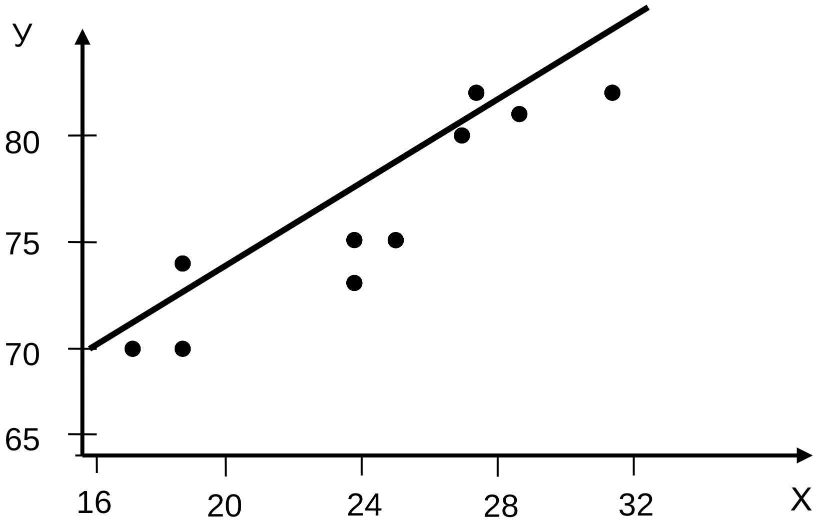


Fig. 9.2. **The correlation field of the dependence of the amount of assets (y) on the amount of deposits (x)**

As can be seen from Fig. 9.2, the relationship between the dependent and independent variables is straight-line and can be expressed by the equation of a straight line:

$$\tilde{y} = a_0 + a_1 x.$$

The parameters (a_0, a_1) of this equation are defined by the least-squares method by solving the system of normal equations:

$$\begin{cases} \Sigma y = na_0 + a_1 \Sigma x \\ \Sigma yx = a_0 \Sigma x + a_1 \Sigma x^2 \end{cases}$$

$$a_0 = \frac{\sum y \sum x^2 - \sum yx \sum y}{n \sum x^2 - \sum x \sum x},$$

$$a_1 = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - \sum x \sum x}.$$

According to Table 9.3 the system of normal equations takes the form:

$$\begin{cases} 758 = 10a_0 + a_1 246, \\ 18\,785 = a_0 246 + a_1 \times 6263, \end{cases}$$

from here:

$$a_0 = \frac{758 \times 6263 - 18\,785 \times 246}{10 \times 6263 - 246 \times 246} = 59.718,$$

$$a_1 = \frac{10 \times 18\,785 - 246 \times 758}{10 \times 6263 - 246 \times 246} = 0.653.$$

Thus, the regression equation that expresses the relationship between the amount of assets and the amount of deposits looks like:

$$\tilde{y} = 59.718 + 0.653x.$$

If the parameters of the regression equation are correctly defined, the equivalence of the sums of the theoretical and empirical values of the amount of assets should be observed (see Table 9.3).

The regression coefficient $a_1 = 0.653$ shows that with the increase in deposits by 1 bln units, the volume of assets increases by 0.653 bln units on average for the population of banks.

To evaluate the effect of the independent variable on the dependent variable one can use the coefficient of elasticity:

$$K_{el} = a_1 \times \frac{\overline{x}}{\overline{y}};$$

$$K_{el} = 0.653 \times \frac{24.6}{75.8} = 0.21.$$

The coefficient of elasticity shows that with an increase in the volume of deposits by 1 %, the volume of assets increases by 0.21 %.

Example 9.3. Let's check the adequacy of the one-factor regression model and the significance of the correlation relationship.

In small populations (in this example, a small sample is taken, n < 30) the regression coefficient (a_1), is prone to random fluctuations. Let us check its significance by the Student's t-test:

$$t = \frac{|a_1|}{\mu_{a_1}},$$

where μ_{a_1} is the average parameter error a_1,

$$\mu_{a_1} = \sqrt{\frac{\sigma^2_{total}}{\sigma^2_x (n-2)}},$$

where σ^2_{total} is the residual variance calculated by formula (9.13) and equal to:

$$\sigma^2_\varepsilon = \frac{32.2648}{10} = 3.52648,$$

where σ^2_x is the variance of the independent variable which according to formula (9.15) is equal to $\frac{6263}{10} - \left(\frac{246}{10}\right)^2 = 11.4.$

The average parameter error a_1 calculated by formula (9.15) is:

$$\mu_{a_1} = \sqrt{\frac{3.52648}{11.14(10-2)}} = 0.1989 \text{ bln units}.$$

The actual value of the Student's t-test is equal to:

$$t = \frac{0.653}{0.1989} = 3.28.$$

According to the table of critical points of the Student's distribution, for $\alpha = 0.05$ and the number of degrees of freedom k = n − m = 10 − 2 = 8, the critical value $t_{0.05}$ = 2.31.

Because t_{fact} > t_{crit} (3.28 > 2.31), the selective regression coefficient a_1 = 0.653 is probable and significant.

Let's determine the interval for the regression coefficient in the population:

$$a_1 - t \times \mu_{a_1} \le a_1 \le a_1 + t \times \mu_{a_1};$$
$$0.653 - 2.31 \times 0.1989 \le a_1 \le 0.653 + 2.31 \times 0.1989;$$
$$0.194 \le a_1 \le 0.112.$$

So, with the level of significance $\alpha = 0.05$ (with the probability of being mistaken in five cases out of a hundred) it can be stated that the value of the regression coefficient characterizing the relationship between the amount of assets and the amount of deposits in the population ranges from 0.194 to 1.112 bln units.

9.4. The indicators of the closeness and significance of the correlation relationship

Checking the adequacy of the regression model can be supplemented by correlation analysis. To do this, determine the correlation between the variables x and y. The tightness of a correlation relationship, like any other, can be measured by *an empirical correlation* η_e, when intergroup variance σ_M^2 characterizes the deviation of the group mean of the dependent variable from the total variance: $\eta_e = \sqrt{\sigma_M^2 / \sigma^2}$.

The theoretical relationship η should be distinguished from the empirical correlation relation. It shows the relative value obtained by comparing the mean square deviation of the values of the dependent variable σ_M with the root mean square deviation of the empirical (actual) values of the dependent variable σ:

$$\eta = \sqrt{\sigma_M^2 / \sigma^2} \ ,$$

where $\sigma_M = \sqrt{\frac{\Sigma(\tilde{y} - \overline{y})^2}{n}} = \sqrt{\sigma_M^2}$;

$$\sigma = \sqrt{\frac{\sum (y - \overline{y})^2}{n}} = \sqrt{\sigma_y^2} \ .$$

Then:
$$\eta = \sqrt{\frac{\Sigma(\tilde{y} - \overline{y})^2}{\Sigma(y - \overline{y})^2}} \ . \qquad (9.16)$$

The change of the value η is explained by the influence of the independent variable.

The basis for the calculation of the correlation ratio is the rule of sum of the variances, that is, $\sigma^2 = \sigma_M^2 + \overline{\sigma}_i^{\,2}$, where $\overline{\sigma}_i^{\,2}$ represents the variation of all factors except x, that is, the residual dispersion:

$$\overline{\sigma}_{total}^{2} = \frac{\Sigma(y - \tilde{y})^2}{n} \ .$$

Then the formula of the theoretical correlation relation will take the form:

$$\eta = \sqrt{\frac{\sigma_M^2}{\sigma^2}} = \sqrt{\frac{\sigma^2 - \sigma_i^2}{\sigma^2}} = \sqrt{1 - \frac{\sigma_\varepsilon^2}{\sigma^2}}, \qquad (9.17)$$

or

$$\eta = \sqrt{1 - \frac{\sum (y - \tilde{y})^2}{\sum (y - \overline{y})^2}} \ . \qquad (9.18)$$

The radicand of the correlation ratio is *the coefficient of determination (causality)*. This coefficient shows the proportion of variation in the dependent variable under the influence of variation of the independent variable.

The theoretical correlation ratio is used to measure the tightness of the relationship in the linear and curvilinear dependencies between the dependent and independent variables. In curvilinear relationships, the theoretical correlation relation calculated by formulas (9.17), (9.18) is called the *correlation index R*.

As can be seen from formulas (9.17) and (9.18), the correlation ratio can be in the range from 0 to 1, that is, $0 \le \eta \le 1$. The closer η to 1, the closer the relationship between the variables.

Let us demonstrate the calculation of the theoretical correlation ratio as a measure of the closeness of the relation based on the example of the data given in Table 9.4, for which the values of daily productivity of one worker depending on the length of service are calculated using the direct regression equation $\tilde{y} = 4 + 0.6x$.

The theoretical correlation relation can be calculated in two ways (Table 9.4):

by formula (9.16), $\eta = \sqrt{\frac{29.7}{32.1}} = \sqrt{0.925} = 0.962$;

by formula (9.18), $\eta = \sqrt{1 \quad \frac{2.4}{29.7}} = \sqrt{0.92} = 0.96$.

Table 9.4

The estimated values required for the calculation of η

$y-\bar{y}$	$(y-\bar{y})^2$	$\tilde{y}-\bar{y}$	$(\tilde{y}-\bar{y})^2$	$y-\tilde{y}$	$(y-\tilde{y})^2$
−3.3	10.89	−2.7	7.29	−0.6	0.36
−2.3	5.29	−2.1	4.41	−0.2	0.04
−1.3	1.69	−1.5	2.25	0.2	0.04
−0.3	0.09	−0.9	0.81	0.6	0.36
−0.3	0.09	−0.3	0.09	0.0	0.00
0.7	0.49	0.3	0.09	0.4	0.16
0.7	0.49	0.9	0.81	-0.2	0.04
1.7	2.89	1.5	2.25	0.2	0.04
2.7	7.29	2.1	4.41	0.6	0.36
1.7	2.89	2.7	7.29	−1.0	1.00
Total	32.10	–	29.70	–	2.40

The obtained value of the theoretical correlation ratio indicates a very close relationship between the variables (according to the Chaddock scale, Table E.1 (Annex E).

The coefficient of determination is 0.925. Hence, 92.5 % of the total variation in labor productivity is due to the variation in the factor of work experience; only 7.5 % of the total variation cannot be explained by the variation in work experience.

In addition, with the linear form of relationship, another indicator of the relationship tightness is used, *the linear correlation coefficient* suggested by the English mathematician K. Pearson:

$$r = \frac{\overline{xy} - \bar{x} \times \bar{y}}{\sigma_x \sigma_y} = \frac{\sum (x - \bar{x}) \times (y - \bar{y})}{n \times \sigma_x \sigma_y}, \tag{9.19}$$

where n is the number of observations.

For practical calculations with a small number of observations, namely for $(n \leq 20 \leq 30)$, it is more convenient to calculate a linear correlation coefficient by the formula:

$$r = \frac{\sum xy - \dfrac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \dfrac{(\sum x)^2}{n}\right) \times \left(\sum y^2 - \dfrac{(\sum y)^2}{n}\right)}}. \tag{9.20}$$

Therefore, the coefficient is in the range from −1 to 1.

The negative value means inverse relationship; the positive one shows direct correlation. If r = 0, there is no linear relationship. The closer the correlation coefficient is to 1, the closer the relationship between the variables. If r = ± 1, the relationship is functional.

Example 9.4. Let's determine the closeness of the correlation between the volume of assets and deposits using the linear correlation coefficient and the coefficient of determination based on the data given in Table 9.3.

The linear correlation coefficient calculated by formula (9.20) will be equal to:

$$r = \frac{18785 - \dfrac{246 \times 758}{10}}{\sqrt{\left(6263 - \dfrac{246^2}{10}\right) \times \left(57660 - \dfrac{758^2}{10}\right)}} = 0.6662.$$

The correlation coefficient shows that there is a noticeable correlation between the volume of assets and the volume of deposits (on the Chaddock scale).

The determination factor r^2 = 0.6662^2 = 0.4438 shows that 44.38 % of the total variation in banks' assets is due to a variation in deposits, and 55.62 % (100 % − 44.38 %) due to other factors that were not taken into account in this example.

The square linear correlation coefficient r^2 is called *a linear coefficient of determination* with the value in the range from 0 to 1. Its interpretation is similar to the coefficient of determination η^2.

The values η and r are the same only if there is a straight relationship. If they are different, this indicates that the relationship between the variables is curvilinear. It is stated that if the difference of squares η^2 and r^2 does not exceed 0.1, the straight-line relationship hypothesis can be considered as confirmed.

Tightness indicators, calculated on the basis of small-size population data, are affected by random variables. This necessitates a check for their significance which gives the grounds to generalize the results of the sample for the general population.

Example 9.5. The probability of the sample correlation coefficient is verified using the critical values of the sample correlation coefficient with different degrees of freedom and significance levels (Annex D). For example (see Table 9.3), the table value of the correlation coefficient for $\alpha = 0.05$ and k = 8 is $r_{0.05}$ = 0.632. Because $r_{fact} > r_{0.05}$ (0.6662 $>$ 0.632), the sample correlation coefficient can be considered probable.

The adequacy of the regression model for n < 30 is estimated using the Fisher test:

$$F = \frac{\sigma^2_{y_x}}{\sigma^2_{\varepsilon}} \times \frac{n-m}{m-1}, \tag{9.21}$$

where m is the number of model parameters;

σ^2_y is the factor variance that characterizes the variation of the dependent variable under the influence of the independent variable included in the model. The factor variance is calculated by the formula:

$$\sigma^2_{\tilde{y}} = \frac{\sum(\tilde{y} - \overline{y})^2}{n} \qquad (9.22)$$

$$\sigma^2_{\tilde{y}} = \sigma^2_{y} - \sigma^2_{\varepsilon} = \frac{\sum y^2}{n} - \left(\frac{\sum y}{n}\right)^2 - \sigma^2_{\varepsilon}. \qquad (9.23)$$

or

$$\sigma^2_{\tilde{y}} = \frac{57660}{10} - \left(\frac{758}{10}\right)^2 - 3.52648 = 16.83;$$

$$F = \frac{16.83}{3.52648} \times \frac{10-2}{2-1} = 38.179.$$

The table value F with significance level $\alpha = 0.05$ and the number of degrees of freedom k_1 = n − m = 10 − 2 = 8 and k_2 = m − 1 = 2 − 1 = 1 is 5.32. Because $F_{calcul} > F_{tabl.}$, the regression equation can be considered adequate.

To test the reliability of the regression model, we calculate the approximation error:

$$\overline{\varepsilon} = \frac{1}{n} \sum \frac{|y - \tilde{y}|}{y} \times 100\,\%, \qquad (9.24)$$

$$\overline{\varepsilon} = \frac{1}{10} \times 0.19855 \times 100\,\% = 1.98\,\%.$$

Because $\overline{\varepsilon}$ does not exceed 12 – 15 %, the regression model can be used in practice.

9.5. The development of multiple correlation and regression models

Pair correlation is used when the factors affecting the dependent variable are dominant.

However, in the practice of economic analysis it is often necessary to study phenomena that are not influenced by one factor but depend on many different factors, each of which may have no decisive influence separately.

The combined influence of factors is sometimes strong enough to draw conclusions about the magnitude of the indicator of the studied phenomenon. Methods for measuring the correlation between two, three or more correlation variables at the same time create the doctrine of *multiple correlation* (the question of multiple correlation was first investigated by the English scientist F. Y. Edgeworth at the end of the 19th century).

In multiple correlation models, the dependent variable y is regarded as a function of several (in general n) independent variables x.

The use of applications makes it possible to solve correlation and regression models of different dependencies and to choose from this set the equation that most accurately describes the degree of approximation of the actual data to the theoretical values and, accordingly, gives the least sum of squares of deviations of the actual data from the data calculated by the equation.

Multivariate correlation and regression analysis can be applied to:

1) the calculation of theoretical (expected) values of the dependent variable;

2) the estimation and comparison of actual and estimated values of the dependent variable;

3) comparative analysis of different populations;

4) identification of production reserves and objective evaluation of the results of the object under research;

5) prediction of social phenomena;

6) development of standards.

Due to the fact that along with the factor under study the dependent variable is also influenced by other factors, pair correlation does not always give a correct idea of the relationship between the dependent and independent variables (it overestimates or underestimates the degree of dependence). The advantage of multivariate correlation and regression analysis over paired correlation is that it allows you to estimate the degree of influence on the dependent variable of each of the factors included in the model, provided the fixed (average) position of the other factors.

Selection of the most significant factors for the correlation model is one of the most important and fundamental tasks of multivariate correlation and regression analysis. Naturally, all the factors that affect the dependent variable being investigated cannot be included in the regression equation. From the whole complex of such factors it is necessary to select the most important, essential ones. Covering a large number of factors with a relatively small population can produce poor results. In addition, an increased number of parameters in the regression equation makes it difficult to interpret the results.

Preconstructed and analyzed factor groups play a key role in the selection of factors. Of particular importance here are combinatorial groupings, which allow the researcher to determine the impact of the variable of interest on the dependent variable (with fixed values of other variables). It can be concluded that statistical groupings form the basis of correlation and variance analysis; the highest efficiency is achieved in combination with the grouping method.

As already mentioned, to ensure the stability of the parameters of the coupling equation, the number of factors included in the model should be 6 – 10 times smaller than the size of the population under study. At the same time, the population from which the factors are selected must be qualitatively homogeneous.

When selecting factors, one should exclude those that are mutually duplicating and are in functional relationship. Functional or close relationship between the variables proper indicates *multicollinearity* (for two variables it means collinearity). The presence of multicollinearity indicates that these factors show the same side of the effect on the dependent variable.

With high correlation of factors (when the close relationship between two factors exceeds г > 0.8) the influence of one of them accumulates the influence of the other. The obtained correlation models become unstable.

In the case of multicollinearity, the inclusion in the correlation model of interrelated factors is possible when the closeness of relationship between them is less than the closeness of relationship of the dependent variable with each factor. The correlation model is required to contain independent, non-overlapping factors. It is undesirable to include partial and general factors in one model. Factors functionally related to the dependent variable should be completely excluded.

The choice of a regression equation for development of a multifactor model is based on the theory of the phenomenon under study or the practical experience of previous studies. The graphical method in this case is ineffective. If there is no preliminary data, combinatorial groupings are built, expert estimates are used, paired relationships between the dependent variable and each factor are studied, functions of different types are sorted out, serial transition from linear equations to more complex types is carried out.

The implementation of all these techniques involves a large number of unnecessary calculations. However, in most cases correlation relationships are represented by linear or power functions that can be reduced to linear form by logarithmizing or replacing variables, so multiple regression equations can be constructed in linear form.

For n variables, the linear multivariate equation looks like:

$$\tilde{y} = a_0 + a_1 x_1 + a_2 x_2 + ... + a_n x_n , \qquad (9.25)$$

where $\tilde{y}$ is the dependent variable;

x_1, x_2, …, x_n are independent variables (factors);

a_0 is the beginning of a reference that does not make economic sense or indicates the presence of factors not taken into account in the model;

a_1, a_2, …, a_n are coefficients of the regression equation.

A multiple regression equation is the equation by which the correlation between several variables is expressed. The parameters of the regression equation, like in the case of pairwise correlation, are found by the least squares method.

Multiple regression coefficients show the degree of average change in a dependent variable with a change in the corresponding independent variable per unit, provided that all other factors included in the regression equation remain fixed at one (average) level.

Multiple regression coefficients which characterize the relationship between the dependent variable and a variable with a fixed value of other factors are called *net regression coefficients*, while the pairwise regression coefficients – *gross regression coefficients*.

Net regression coefficients that have different physical meaning and units of measurement do not give a clear idea of which factors most significantly affect the dependent variable. In addition, the magnitude of the regression coefficients depends on the degree of variation of the variable. To bring net regression coefficients to a comparable form, they are expressed in standardized form as elasticity coefficients (E) and beta coefficients (β).

The coefficients of elasticity show, how many percent the value of the dependent variable changes with the change in the relevant independent variable by one percent, provided the fixed values of other factors.

The elasticity coefficients and the net regression coefficients are correlated as (9.26):

$$E_i = a_i \times \frac{\overline{x_i}}{\overline{y}}, \tag{9.26}$$

where a_i is the net regression coefficient for the i-th factor;

$\bar{x}_i$ and $\bar{y}$ are the average values for the i-th factor and the dependent variable correspondingly.

Beta coefficients show how many standard deviations σ_y the dependent variable will change with the change in the relevant factor by one standard deviation σ_x (provided the constancy of other factors included in the regression equation).

Beta coefficients are calculated by the formula:

$$\beta_i = a_i \frac{\sigma_{x_i}}{\sigma_y}, \tag{9.27}$$

where a_i is the net regression coefficient for the i-th factor;

σ_{x_i} and σ_y are mean square deviations respectively, for the i-th factor and the dependent variable.

From the above formula, it follows that the beta coefficients have the same sign (plus, minus) as the net regression coefficients.

Thus, beta coefficients characterize the factors in the development of which the greatest reserves for improving the dependent variable are hidden.

In a pairwise linear relationship, the correlation coefficient is the beta coefficient:

$$r = b\frac{\sigma_x}{\sigma_y} = \beta. \tag{9.28}$$

A multivariate regression model having being developed, the characteristics of the relationship between the dependent and the independent variables are calculated: pairwise, partial, and multiple correlation coefficients; the multiple coefficient of determination; and then the model is checked for adequacy.

To measure the closeness of the relationship between the two variables being considered (without taking into account their interaction with other variables), *paired correlation coefficients* are used. The method for calculation and interpretation of such coefficients is similar to the method for calculation of a linear correlation coefficient in the case of a one-factor relationship.

Paired correlation coefficients are calculated by the formulas:

$$r_{y_{x_1}} = \frac{\overline{x_1 y} - \bar{x}_1 \bar{y}}{\sigma_{x_1} \sigma_y}; \qquad (9.29)$$

$$r_{y_{x_2}} = \frac{\overline{x_2 y} - \bar{x}_2 \bar{y}}{\sigma_{x_2} \sigma_y}; \qquad (9.30)$$

$$r_{x_1 x_2} = \frac{\overline{x_1 x_2} - \bar{x}_1 \bar{x}_2}{\sigma_{x_1} \sigma_{x_2}}. \qquad (9.31)$$

In real life, all variables are interrelated. The tightness of this relationship is determined by *partial correlation coefficients* which characterize the degree and influence of one of the arguments on the function, provided that other independent variables are fixed at a constant level. Depending on the number of variables whose influence is eliminated, the partial correlation coefficients can be of different order: with the exclusion of the influence of one variable we obtain the partial correlation coefficient of the first order; with the exclusion of the influence of two variables – the second-order coefficient and so on. The paired coefficient of correlation between the function and the argument is usually not equal to the corresponding partial coefficient.

The partial coefficient of first-order correlation between the variables x_1 and y with the exclusion of the influence of the variable x_2 is calculated by the formula:

$$r_{y_{x1(x_2)}} = \frac{r_{y_{x_1}} - r_{y_{x_2}} r_{x_1 x_2}}{\sqrt{(1 - r^2_{y_{x_2}}) \times (1 - r^2_{x_1 x_2})}}; \qquad (9.32)$$

the dependence of y on x_2 with the exclusion of the influence of x_1:

$$r_{y_{x_2(x_1)}} = \frac{r_{y_{x_2}} - r_{y_{x_1}} r_{x_1 x_2}}{\sqrt{(1 - r^2_{y_{x_1}}) \times (1 - r^2_{x_1 x_2})}} . \qquad (9.33)$$

It is possible to calculate the correlation of the independent variable with the exclusion of the effect of the dependent variable:

$$r_{x_1 x_2 (y)} = \frac{r_{x_1 x_2} - r_{y_{x_1}} r_{x_1 x_2}}{\sqrt{(1 - r^2_{y_{x_1}}) \times (1 - r^2_{y_{x_2}})}} , \qquad (9.34)$$

where r is the paired coefficient of correlation between the relevant variables.

Studying the paired and partial correlation coefficients allows us to select the most significant factors.

The indicator of a close relationship between the dependent variable and two or more independent variables is *the aggregate multiple correlation coefficient* ($R_{y x_1, x_2, \ldots, x_n}$). In the case of a linear two-factor relationship, this factor is calculated by the formula:

$$R_{y_{x1x2}} = \sqrt{\frac{r^2_{y_{x1}} + r^2_{y_{x2}} - 2 r_{y_{x1}} r_{y_{x2}} r_{x_1 x_2}}{1 - r^2_{x_1 x_2}}} . \qquad (9.35)$$

The aggregate multiple correlation coefficient measures the simultaneous effect of independent variables on the dependent variable. Its values range from −1 to +1. The smaller the deviation of empirical values of the considered parameter from the multiple regression line, the stronger the correlation relationship is, and therefore the closer R value is to unity.

To measure the proportion of variation in a dependent variable, which is explained by the influence of the factors included in the multiple regression equation, *the aggregate coefficient of multiple determination* (R^2) is used. The values of this coefficient are in the range from 0 to 1. The closer R^2 to one, the more the variation of the analyzed variable is characterized by the influence of the selected factors.

But it should be noted that the coefficient R^2 cannot be used to test the adequacy of the regression equation. The reason is that multivariate regression analysis uses random observations, but not necessarily distributed according to multidimensional normal law (the deviations of the actual values of the function from the calculated ones must be subject to this law). The aggregate multiple determination coefficient shows only the quality of alignment by the regression equation.

The significance of the regression equation is verified on the basis of the Fisher test (9.21). It is believed that the regression equation can be used in practice at least four times.

To estimate the weight of regression coefficients by linear dependence of y on x_1 and x_2, the Student's t-test is used provided n − m − 1 degrees of freedom:

$$t_{a_1} = \frac{a_1 \sigma_{x_1} \sqrt{1 - r^2_{x_1 x_2}} \sqrt{n - m - 1}}{\sigma_y \sqrt{1 - R^2_{y x_1 x_2}}}; \tag{9.36}$$

$$t_{a_2} = \frac{a_2 \sigma_{x_2} \sqrt{1 - r^2_{x_1 x_2}} \sqrt{n - m - 1}}{\sigma_y \sqrt{1 - R^2_{y x_1 x_2}}}. \tag{9.37}$$

The values a_1 and a_2 are taken as positive. The parameters are weighty if $t_{calc} > t_{tabl}$ with the level of significance α and the number of degrees of freedom n − m − 1.

The estimation of regression coefficients' weight by the t-criterion is used to complete the selection of significant factors in the process of multistep regression analysis. Its essence is in the fact that after the estimation of the regression the factor where the coefficient is negligible and has the least criterion value is excluded from the model. Then the regression equation is built without this factor. Again, the adequacy of the equation and the weight of the regression coefficients are evaluated. This process continues until all regression coefficients are significant, indicating that only significant factors are present in the regression model. In some cases, the calculated value t_{calc} is close to t_{tabl}, so this variable can be left to further test its significance in conjunction with another set of factors.

Consistent elimination of irrelevant factors is the basis of multistep regression analysis.

The significance of the aggregate correlation coefficient is also estimated by the Student's t-test using the formula:

$$t_{Ryx_1x_2} = \frac{R^2_{yx_1x_2}\sqrt{n-m-1}}{1-R^2_{yx_1x_2}}. \qquad (9.38)$$

If $t_{calc} > t_{tabl}$ with the significance level of 0.01 or 0.05 and the number of degrees of freedom ($n - m - 1$), the multiple correlation coefficient is considered significant.

9.6. Methods for studying the relationship of social phenomena

The use of correlation and regression models (CRMs) to solve specific socio-economic problems is necessary:

to build a regression equation based on the considered principles and conditions;

to identify those variables whose regression coefficient signs do not correspond to theoretical concepts, based on the analysis of signs of regression coefficients. This may be due to functional relationships between explanatory variables, i.e. the presence of collinearity;

to perform calculations using software-based applications which give a t-value for each of the regression coefficients. Therefore, if the regression coefficient exceeds the standard error of at least 1.95 times (that is $t \geq 1.95$), it is considered statistically significant.

The following options are possible to analyze the adequacy of the regression equation:

the developed model, based on the Fisher test, is generally adequate; all regression coefficients are significant. This model can be used to make decisions and predictions;

according to the Fisher test, the model is adequate, but some of the regression coefficients are not significant. In this case, the model can be used to make some decisions, but not to make predictions;

the Fisher-tested model is adequate, but all regression coefficients are not weighty. In this case, the model is considered completely inadequate; neither decisions no forecasts are made on this basis;

regression can be used to analyze the activity of population units with a coefficient of determination close to unity and not lower than 0.5;

regression is used to make predictions: if the value of the variables is in the middle of the interval of the actual values of the factors, it is a point forecast; if they go beyond these limits, they give an interval estimate of the expected value of the result.

If it is necessary to go beyond the real limits of measuring an independent variable, the following limitation should be observed: it is impossible to substitute in the regression equation the values of x which are significantly different from those on the basis of which this equation was obtained. To select the predictive variable (factor) values, it is recommended that the magnitude of variation be within 1/3 of both the minimum and the maximum value of the variable (factor) that occurred in the source data.

Thus, we can offer the following recommendations for the development of a correlation and regression model:

1) independent variables must be causally linked to the dependent variable (consequence);

2) independent variables should not be part of the dependent variable or its functions;

3) independent variables should not duplicate each other, i.e. be collinear (with a correlation coefficient greater than 0.8);

4) you shouldn't include in the model the factors of different levels of the hierarchy, that is, the closest-order factor and its subfactors. For example, the product cost model should not include both labor productivity and the age of workers as subfactors of productivity itself;

5) it is desirable for the dependent and independent variables that unity of the unit of the population to which they are assigned be fulfilled. For example, if y is the output of an enterprise, the factors x must relate to the enterprise (the level of specialization, the level of spoilage, the number of skilled workers, etc.);

6) the mathematical form of the regression equation must be consistent with the logic of linking the factors to the result in the real environment. Therefore, they choose an additive or a multiplicative form of relationship;

7) preference is given to a model with fewer factors with the same or even slightly less coefficient of determination;

8) if, in addition to quantitative factors, in the multivariate regression analysis, the equations include nonquantitative ones, their presence in the

units of a population is denoted by one, and absence by zero, that is, dummy variables are introduced: $U = \begin{cases} 1 \\ 0 \end{cases}$. The number of dummy variables should be one less than the number of gradations of a qualitative (nonquantitative) factor. This technique measures the influence of qualitative factors, isolating them from the influence of quantitative ones.

The main difference between the method of correlation and regression analysis and the method of analytical groupings is that correlation and regression analysis allows you to divide the influence of a set of independent variables, to analyze different sides of a complex system of relationships. While the method of combined analytical grouping does not allow you to analyze more than three factors, the correlation method with the size of the population of about 100 units makes it possible to analyze the systems with 8 – 10 factors and separate their effects.

The main limitations of the method of correlation and regression analysis is that this method cannot explain the role of independent variables in creating a dependent variable. In addition, regression equation models have weak extrapolation properties because they do not represent trends in socio-economic phenomena and processes and can only be used to make short-term probabilistic forecasts.

The method of analytical grouping and correlation and regression analysis uses the basic parameters of distribution – average values and variances, so they are called *parametric*. In statistics, *nonparametric methods* are also widely used for identification of relationships that are based on quantitative values of variables and do not require calculation of their distribution parameters. While in the correlation and regression analysis all the variables belong to the metric scale, and in the method of analytical grouping this refers to the dependent variable, nonparametric methods are used even when there are variables of ordinal or nominal scale. But it should be emphasized that this advantage of nonparametric methods is achieved due to the smaller depth of interconnection analysis; they only determine the presence of the relationship and measure its tightness.

The use of nonparametric methods for estimation of the closeness of a stochastic dependence is based on the comparison of frequencies or fractions of conditional distributions in tables of mutual conjunction.

A mutual conjunction table is a table that contains a summary of numerical characteristics of a population based on two or more attributes or a combination of quantitative and attributive variables. The layout of the conjunction table with dimension i x j, where i = 1, 2, …, k is the number of variants of values of one variable (A); j = 1, 2, …, n is the number of variants of values of the second variable (B) is given in Table 9.5.

Table 9.5

The general scheme of a conjunction table

B A	B_1	B_2	…	B_j	Total
A_1	f_{11}	f_{12}	…	f_{1j}	f_{10}
A_2	f_{21}	f_{22}	…	f_{2j}	f_{20}
…	…	…	…	…	…
A_i	f_{i1}	f_{i2}	…	f_{ij}	f_{i0}
Total	f_{01}	f_{02}	…	f_{0j}	f_{00}

Two qualitative variables, each consisting of only two groups, are used to determine the closeness of the relationship between the *association and contingency coefficients:*

$$K_a = \frac{f_{11} \times f_{22} - f_{12} \times f_{21}}{f_{11} \times f_{22} + f_{12} \times f_{21}}; \qquad (9.39)$$

$$K_c = \frac{f_{11} \times f_{22} - \ f_{12} \times f_{21}}{\sqrt{(f_{11}} + f_{12}) \times (f_{12} + f_{22}) \times (f_{11} + f_{21}) \times (f_{21} + f_{22})}. \qquad (9.40)$$

The value of the contingency coefficient is always less than the value of the coefficient of association. The link between the variables is considered confirmed, provided $K_a \geq 0.5$ or $K_c \geq 0.3$.

A useful measure in the analysis of 4-cell conjunction tables is *the ratio of cross products, or the odds ratio*:

$$W = \frac{f_{11} f_{22}}{f_{12} f_{21}}. \qquad (9.41)$$

The odds ratio is a measure of relative risk.

Example 9.6. According to Table 9.6, let's evaluate the closeness of the relationship between the ownership form of enterprises and the level of satisfaction with the living conditions of workers employed at these enterprises, as well as the effectiveness of change in the form of ownership of the enterprise.

The contingency coefficient indicates the presence of a stochastic relation:

$$K_c = \frac{13 \times 60 - 24 \times 3}{\sqrt{16 \times 63 \times 37 \times 84}} = 0.40.$$

The odds ratio is: $W = \frac{13 \times 60}{24 \times 3} = 10.8$.

Table 9.6

The table of conjunction between two features

Form of ownership of the enterprise	Number of workers		Total
	Satisfied with the standard of living	Dissatisfied with the standard of living	
Private	13	3	16
State	24	60	84
Total	37	63	100

That is, the chances of being satisfied with living conditions among workers of private enterprises are 10.8 times higher than among workers of state-owned enterprises. So, it makes sense for a company to change the ownership form, for a worker to change their place of work.

When each of the qualitative variables consists of more than two groups, to determine the relationship tightness, *Pearson contingency coefficient, Chuprov contingency coefficient* and a modification of *Chuprov* coefficient – *Cramer coefficient* – are used.

Pearson contingency ratio is:

$$C_P = \sqrt{\frac{\chi^2}{n + \chi^2}}, \qquad (9.42)$$

where χ^2 is *Pearson's chi-square ratio*, which characterizes the discrepancy between the frequencies (fractions) of the conditional and unconditional distribution; it is calculated by the formula:

$$\chi^2 = n\left[\sum_i \sum_j \frac{f_{ij}^2}{f_{i0} \times f_{i1}} - 1\right], \tag{9.43}$$

where n is the number of observations.

The actual Pearson values χ^2 are compared to the critical ones. Critical values χ^2 for α = 0.05 and the number of degrees of freedom $k = (m_x - 1) \times (m_y - 1)$ are given in Table E.2, Annex E.

If the actual value exceeds the critical one, the significance of the association between the features is probably 0.95.

The Chuprov mutual conjunction coefficient is:

$$K_{ch} = \sqrt{\frac{\chi^2}{n\sqrt{(m_x - 1) \times (m_y - 1)}}}, \tag{9.44}$$

where m_x is the number of groups on the basis of the variable x;
m_y is the number of groups based on the variable y.

Because with the independence of the variables $\chi^2 = 0$, $K_{ch} = 0$. In the functional relationship $K_{ch} = 1$ provided that $m_x = m_y$.

When $m_x \neq m_y$, it is more convenient to use the Cramer formula:

$$K_c = \sqrt{\frac{\chi^2}{n(m_{min} - 1)}}, \tag{9.45}$$

where m_{min} is the minimum number of groups (m_x or m_y).

If $m_x = m_y$, the values of the coefficients calculated by the Chuprov and Cramer formulas coincide.

The Chuprov, Cramer, and Pearson coefficients range from 0 to 1. The Chuprov coefficient takes into account the number of selected groups according to each variable and gives the most approximate estimation of the relation. With the value $K_{ch} \geq 0.3$, we can talk about a moderate or close relationship between the variables.

The significance of the relation is verified on the basis of the criterion χ^2 – Pearson's square.

Example 9.7. Using the coefficients of contingency, let's investigate the tightness of the relationship between the type of economic entity and the type of deposit depending on the currency (Table 9.7).

Table 9.7

Deposits to credit institutions of the region

Types of deposits	All deposits, million units	including		
		enterprises and organizations	individuals	banks
Total	702.7	69.3	580.6	52.8
including:				
in UAH	222.1	18.0	200.7	3.4
in foreign currency	480.6	51.3	379.9	49.4

Using the Pearson, Chuprov, and Cramer coefficients, we can determine the relationship between the variables:

$$K_p = \sqrt{\frac{18.84}{702.7 + 18.84}} = 0.16;$$

$$K_{ch} = \sqrt{\frac{18.84}{702.7 \times (2-1) \times (3-1)}} = 0.12;$$

$$K_c = \sqrt{\frac{18.84}{702.7 \times (2-1)}} = 0.16.$$

Since the calculated coefficients of relationship tightness are low (even less than 0.3), we can conclude that there is no relationship between the variables and that they are independent of each other. Let us verify the significance of this finding using Pearson's χ^2 – Pearson's squared test.

The actual value of the quadratic conjunction coefficient χ^2 – Pearson's squared test is:

$$\chi^2 = 702.7\left[\begin{array}{l}\frac{18^2}{222.1\times 69.3}+\frac{200.7^2}{222.1\times 580.6}+\frac{3.4^2}{222.1\times 52.8}+\\ \frac{51.3^2}{480.6\times 69.3}+\frac{379.9^2}{480.6\times 580.6}+\frac{49.4^2}{480.6\times 52.8}-1\end{array}\right]=18.84,$$

which far exceeds the critical value for α = 0.05 and the number of degrees of freedom k = (2 − 1) × (3 − 1) = 2, which is 5.99. Therefore, a 95 % probability of no correlation between the type of deposit depending on the currency and the entity is proved.

The methods of analysis of conjunction tables can also be used for quantitative variables.

The biserial correlation coefficient proposed by K. Pearson is intended for the study of correlations in 2 × n tables, which are dichotomies based on a certain nominal variable and classifications based on a nominal or ordinal variable, which is classified according to q classes and may be ordered or disordered. The output distribution should be two-dimensional normal. In case of classification based on the ordered variable, the biserial coefficient is calculated by the formula:

$$r_b = \frac{(\bar{x}_1 - \bar{x})}{ns_x z_k}, \tag{9.46}$$

where $\bar{x}_1$ is the first-line average;

$\bar{x}$ is the table average;

s_1 is the sample mean square deviation;

n is the total number of samples;

z_k is the ordinate of the density of normal distribution at point k, where k is the solution of the equation:

$$1-F(k)=n_1/n, \tag{9.47}$$

where n_1 is the number of samples in the first line.

The biserial correlation coefficient values can vary from −1 to +1. Its error can be determined by the formula:

$$m_{rb} = \frac{1 - r_b}{\sqrt{n}}. \qquad (9.48)$$

The biserial correlation coefficient has a t-distribution with the number of degrees of freedom (n − 2).

9.7. Nonparametric methods

In the analysis of socio-economic phenomena and processes conditional estimates (for example, ranks) are used. **Ranking** is a prearranged ordering of the objects of study. A **rank** is the ordinal number of the values of a variable, arranged in ascending or descending order.

If the variable values have the same quantitative rating, the rank of all these values is assumed to be equal to the arithmetic mean of the corresponding place numbers that are determined. Such ranks are called *connected*.

To measure the relationship between the variables of the order scale the **Spearman rank correlation coefficient** is used. The calculation of the coefficient is based on the difference of ranks $d = R_x - R_y$, where R_x, R_y is the rank of the elements of a population based on the first and second variables respectively. This coefficient is calculated by the formula:

$$\rho = 1 - \frac{6\,\Sigma d^2}{n(n^2 - 1)}, \qquad (9.49)$$

where n is the number of elements in the population.

The rank correlation coefficient, as well as the linear correlation coefficient, can be from −1 to +1. If two rows of a rank are exactly the same, Σd^2 = 0. Therefore, there is a complete direct relationship and ρ = 1. With a fully inverse relationship (ranks of two rows are in reverse order) ρ = −1.

The critical values of the Spearman rank correlation coefficient are given in Annex F.

As an example, let's use the data in Table 9.8 which lists ten enterprises according to the volume of sales and profit.

Table 9.8

Calculation of the Spearman rank correlation coefficient

Company number	Ranking				d	d^2
	Sales volume, million UAH	R_x	Amount of profit, thousand UAH	R_y		
1	1.2	2	200	1	1	1
2	1.4	3	240	3	0	0
3	5.2	10	384	6	4	16
4	0.9	1	256	4	−3	9
5	1.9	4	215	2	2	4
6	2.7	6	395	7	−1	1
7	3.1	7	310	5	2	4
8	3.9	8	425	8	0	0
9	2.4	5	469	9	−4	16
10	4.0	9	471	10	−1	1

Substituting the necessary data into formula (9.49), we obtain:

$$\rho = 1 - \frac{6 \times 52}{10 \times (100 - 1)} = 0.68.$$

This indicates a direct and noticeable relationship between the amount of profit and the volume of sales.

The significance of the relationship can be checked by comparing the critical value of the rank correlation coefficient for the significance level α = 0.05 and n = 10, which is $\rho_{0.95}(10)$ = 0.5515 (Table E.1, Annex E) with the actual value. Since actual ρ is more than critical, the presence and significance of the association between the variables is proven with 95 % probability.

It should be noted that for the inverse relationship, the absolute value of the actual level of ρ is compared to the critical value of the rank correlation coefficient.

Rank coefficients have both advantages and disadvantages compared to parametric ones. There is no need to adhere to certain mathematical preconditions for the distribution of variables (in particular, the preconditions for the normality of distribution). However, some relationship information is lost because it is not the values that are used, but only the variable ranks.

The method for measuring a relationship and its closeness characteristics should be based on previous theoretical analysis of the nature of the phenomena, the nature of the relationships, the available information.

To study stochastic (correlation) relationships, we use the method of comparing parallel series of two indicators, one of which is independent (variable x) and the other is dependent (variable y). The main task of applying this method is to evaluate the closeness (strength) of the relationship and to determine its direction based on the calculation of special coefficients.

The simplest indicator is the *Fechner coefficient* (K_f), which is calculated by the formula:

$$K_f = \frac{C - H}{C + H}, \tag{9.50}$$

where C is the number of coincidence signs of deviations from the mean;

H is the number of half-matches of deviations from the mean.

If inequality $x \geq \bar{x}$ or $y \geq \bar{y}$ holds, the value is assigned a "+" sign, otherwise a "−" sign. If the signs are the same in both indicators, they coincide, and when they are different, there is no coincidence. The Fechner coefficient ranges from −1 to +1. If $|K_f| \to 0$, the relationship between indicators is weak, and for $|K_f| \to 1$ the relationship is tight. This coefficient has a positive value in the presence of direct relationship and a negative value if the relationship is inverse.

The *coefficients of rank correlation of Spearman or Kendall*, considered to be a more perfect indicator, are used to test the hypothesis of a relationship between two variables after a previous ranking. For the correct calculation of both coefficients (Spearman and Kendall), the measurement results must be presented on a scale of ranks or intervals. There are no fundamental differences between these criteria, but it is accepted that the Kendall coefficient is "more substantial" since it analyzes the relationships between the variables more fully and in detail, focusing on all possible correspondences between pairs of values. The Spearman coefficient takes into account more precisely the quantitative degree of the relationship between the variables.

The *Spearman rank correlation coefficient* is a nonparametric analogue of the classical Pearson correlation coefficient, but its estimates do not include the variables related to the distribution of the compared variables (arithmetic mean and variance), but are based on the ranks.

The limitations for applying the Spearman rank coefficient are:

there must be at least five observations for each variable;

with a large number of identical ranks in one or two variables the coefficient gives an inaccurate value.

Example 9.8. Let (x_i, y_i), i = 1, 2, …, n is a sample of observations of two variables x and y measured on sequential or quantitative scales. Suppose that among the sample elements x_i and y_i (I = 1, 2, …, n) there are no matching items. Let's arrange the elements x_i in ascending order, that is, write a variation series ($x^{(1)}$, $x^{(2)}$, …, $x^{(n)}$), and assign a rank $x_i^{'}$ to every x_i, i.e. the item number x_i in the variation series. Obviously, the smallest element of the sample $x^{(n)}$ is the rank n. Similarly, the ranks $y_i^{'}$ of items $y_i^{'}$, (i = 1, 2, …, n) are defined. Each pair (x_i, y_i) matches the pairs of ranks ($x_i^{'}$, $y_i^{'}$). The Spearman rank correlation coefficient is calculated by the formula:

$$r_s = 1 - \frac{6\sum\left(x_i^{'} - y_i^{'}\right)}{n(n^2 - 1)}. \qquad (9.51)$$

The value obtained, r_s, is called *the sample Spearman rank correlation coefficient*. The coefficient r_s is to be within 0 to 1. Higher values of the sample coefficient r_s show that there is a dependence between the random variables x and y (in this case they say that Spearman rank correlation coefficient is significant).

If $|r_s| < r(\alpha, n)$, where α is the given level of significance, n is the sample size, the hypothesis H0: r_s = 0 is accepted at the level of significance 2α; if the condition is not satisfied, the alternative hypothesis H1: $r_s \neq 0$ is taken.

The Kendall's tau rank correlation coefficient (Kendall's tau-b) is an independent original method that relies on the calculation of the ratio of pairs of values of two samples that have the same or different trends (an increase or a decrease in values). This factor is also called *the coefficient of concordance*. Thus, the basic idea of this method is that the direction of relationship can be evaluated by pairwise comparison of observations: if in a pair of observation a change for x coincides with the direction with a change for y, it indicates a positive relationship; if not, it is a negative relationship.

The Kendall's tau rank correlation coefficient is calculated by the formula:

$$T = 1 - \frac{4k}{n(n-1)}, \qquad (9.52)$$

where k is the number of inversions in the rank of the second variable ($y_i^{'}$) provided that the ranks of the first variable ($x_i^{'}$) are ordered.

In the case of convergent ranks, to calculate the rank correlation coefficients r_s and T the adjusted formulas are used. The sample value of Spearman rank correlation coefficient is calculated by the formula:

$$r_s = \frac{\frac{1}{6}(n^3 - n) - \sum (x_i^{'} - y_i^{'})^2 - T_x - T_y}{\sqrt{\left[\frac{1}{6}(n^3 - n) - 2T_x\right]\left[\frac{1}{6}\left(n^3 - n\right) - 2T_y\right]}}, \qquad (9.53)$$

where $T_x = \frac{1}{12}\sum\limits_{t=1}^{m_x}\left[(n_t)^3 - n_t\right]$, $T_y = \frac{1}{12}\sum\limits_{l=1}^{m_y}\left[(n_l)^3 - n_l\right]$, in which:

m_x is the number of groups of overlapping ranks in the sequence of ranks $x_i^{'}$;

n_t is the number of overlapping ranks in the group number t (t = 1, 2, …, m_x);

m_y is the number of groups of overlapping ranks in the sequence of ranks $y_i^{'}$;

n_l is the number of overlapping ranks in the group with the number l, (l = 1, 2, …, m_y).

The adjusted formula for calculating the Kendall rank correlation coefficient has the form:

$$T' = \frac{T - \frac{2(U_x + U_y)}{n(n-1)}}{\sqrt{(1 - \frac{2U_x}{n(n-1)})(1 - \frac{2U_y}{n(n-1)})}}, \qquad (9.54)$$

where T is the Kendall's rank correlation coefficient calculated without correction.

$$U_x = \frac{1}{2}\sum\limits_{t=1}^{m_x} n_t\left(n_t - 1\right); \qquad (9.55)$$

$$U_x = \frac{1}{2}\sum\limits_{l=1}^{m_x} n_l\left(n_l - 1\right). \qquad (9.56)$$

The use of the Kendall coefficient is preferred if the initial data are anomalous.

The peculiarity of the rank correlation coefficients is that the maximum modulo rank correlations (+1, −1) do not necessarily correspond to strictly directly or inversely proportional relationships between the original variables x and y: only a monotonic functional relationship between them is sufficient. Rank correlations reach their maximum modulo value if a larger value of one variable always corresponds to a larger value of another variable (+1) or a larger value of one variable always corresponds to a smaller value of another variable, and vice versa (−1).

The statistical hypothesis that is tested, the procedure for making a statistical decision and formulating a meaningful conclusion are the same as for the case of Spearman or Pearson rank correlation coefficients.

If no statistically significant relationship is found, but there is reason to believe that the relationship actually exists, you should first proceed from Spearman to Kendall procedure (or vice versa), and then check for possible causes of relationship inaccuracy. There may be such reasons as:

relationship nonlinearity (consider a two-dimensional scatter plot for this). If the relationship is not monotonous, divide the sample into parts in which the relationship is monotonous or contrast groups and compare them as to the level of manifestation of the variable;

sampling heterogeneity (consider a graph of two-dimensional scattering). You should try to divide the sample into parts where the relationship may have different directions.

If the relationship is statistically reliable, before drawing meaningful conclusions, it is necessary to exclude the possibility of false correlation (by analogy with metric correlation coefficients).

Important concepts

Beta coefficients are coefficients indicating how many rms deviations the dependent variable will change with the change in the corresponding factor by one rms deviation (provided the constancy of other factors included in the regression equation).

Empirical correlation is the coefficient by which the correlation relationship is measured.

Residual variance is the quality indicator of a regression equation.

Inverse relationship is a relationship where the direction of change in the dependent variable does not coincide with the direction of change in the independent variable; that is, with the increase of the independent variable the dependent variable decreases, and vice versa.

Pearson mutual conjunction coefficient, Chuprov mutual conjunction coefficient, Cramer coefficient are relationship tightness ratios used when each qualitative variable consists of more than two groups.

Determination factor is the coefficient that shows the proportion of variation of a dependent variable under the influence of variation of the independent variable.

Pearson chi-square ratio is the coefficient characterizing the difference between the frequencies (fractions) of the conditional and unconditional distribution; it is the criterion for verification of the relationship significance.

Spearman rank correlation coefficient is the coefficient used to measure the association between the features of an order scale.

Kendall's tau rank correlation coefficient (Kendal's tau-b) is an independent original method that relies on the calculation of the ratio of pairs of values of two samples that have the same or different trends (increasing or decreasing values). This coefficient is also called the coefficient of concordance.

Regression coefficient is a named value that has the dimension of a dependent variable and is considered as the effect of x on y.

Fechner coefficient is the factor that characterizes the elementary degree of tightness of the relationship that it is appropriate to use to establish the fact of the presence of the relationship if there is a small amount of initial information.

Association and contingency ratios are the coefficients used to determine the relationship between two qualitative variables, each consisting of only two groups.

Full regression coefficients are pairwise regression coefficients.

Net regression coefficients are multiple regression coefficients that characterize the relationship between the dependent and independent variable, provided that other factors have a fixed value.

Correlation analysis is the analysis quantifying the closeness of relationship between two variables (paired relationship) and the dependent variable and multiple independent variables (in a multifactor relationship).

Correlation relationship is a subtype of stochastic relationship when with a change in the independent variable x the group mean values of the dependent variable y change. That is, instead of conditional distributions, the average values of these distributions are compared.

Linear relationship is the relationship between phenomena, which can be expressed by the equation of line.

Linear coefficient of determination is the squared linear correlation coefficient.

Linear correlation coefficient is a measure of the closeness of relationship used in the linear form of relationship (proposed by the English mathematician K. Pearson).

Multiple correlation is measuring the correlation between two, three or more correlation variables.

Nonlinear or curvilinear relationship is the relationship between phenomena which is expressed by the equation of any curve line (parabola, hyperbola, power, exponential, etc.).

Nonparametric methods are methods for determining relationships that are based on the quantitative values of variables that do not require the calculation of the parameters of their distributions.

Parametric methods are methods that use basic distribution parameters – averages and variances.

Paired correlation coefficients are coefficients that measure the closeness of a relationship between two of a set of variables (excluding their interaction with other variables).

Direct relationship is the relationship in which the direction of change in a dependent variable coincides with the direction of change in the independent variable; that is, with the increase of the independent variable the dependent variable increases, and vice versa.

Rank is a sequential number of the variable values arranged in ascending or descending order.

Ranking is the preference-based ordering of the objects of study.

Regression analysis is determining the analytic expression of a relationship, establishing the degree of influence of independent variables on the dependent one, and determining the calculated values of the dependent variable (a regression function).

Stochastic relationship is a relationship in which each value of the variable x is given by a set of values of the variable y that vary and form a conditional series of distribution.

Aggregate multiple determination coefficient is an indicator that measures the proportion of variation in the dependent variable, which is explained by the influence of the factors included in the multiple regression equation.

Aggregate multiple correlation coefficient is an indicator of the closeness of relationship between the dependent and two or more independent variables.

Conjunction table is a table containing a summarized numerical characteristic of the population under consideration with two or more attributive variables or a combination of quantitative and attributive variables.

Theoretical regression line is a continuous line described by the function y = f (x) called the regression equation.

Factorial feature (independent variable) is a variable characterizing the cause.

Functional relationship is a relationship for which each value of an independent variable corresponds to one well-defined value of a dependent variable.

Partial correlation coefficients are coefficients characterizing the degree and influence of one of the arguments on a function, provided that other independent variables are fixed at a constant level.

Typical tasks

Task 1. Ascertain the stochastic relationship by using the combinational distribution of the elements of the statistical population.

The solution.

A set of mines in the region is divided into groups on two grounds: x, the depth of development of coal layers; y, the capital/output ratio of the coal. Each group, depending on the depth of development, is characterized by a special distribution of mines based on the complexity of coal production. These are conditional distributions. Comparison of conditional distributions indicates a tendency to an increase in capital intensity with the increase of depth of development of layers. Of course, for each individual mine, such dependence might not be revealed due to the influence of other factors. There are specific limits to the variation in capital intensity in each group. Thus, in mines where the depth of development of layers is 500 – 700 m, the capital intensity varies from 18 to 26 UAH per ton. However, the average level of the capital/output ratio in this group is higher compared to the previous group (300 – 500 m) and lower than in the next one (700 m and more) (Table 9.9).

Table 9.9

Combinational distribution of mines depending on the depth of coal development and coal capital/output ratio

Depth of the layer development, m	Number of mines according to the capital/output ratio, UAH/t						Average level of the capital/ output ratio, UAH/t
	Up to 20	20 – 22	22 – 24	24 – 26	26 and more	Total	
Less than 300	9	7	1			17	20.0
300 – 500		8	27	5		40	22.9
500 – 700			6	15	4	25	24.8
700 and more				8	10	18	26.1
Total	9	15	34	28	14	100	23.5

The results of the calculations of the average capital/output ratio for each group are as follows:

$$\bar{y}_1 = \frac{19 \times 9 + 21 \times 7 + 23 \times 1}{17} = 20; \quad \bar{y}_2 = \frac{21 \times 8 + 23 \times 27 + 25 \times 5}{40} = 22.9;$$

$$\bar{y}_3 = \frac{23 \times 6 + 51 \times 15 + 27 \times 4}{25} = 24.8; \quad \bar{y}_4 = \frac{25 \times 8 + 27 \times 10}{18} = 26.1.$$

The average levels of the coal capital/output ratio are shown in the last column of Table 9.9. The increase in group averages from group to group indicates that there is a correlation between the depth of development of the bed and the capital/output ratio. Thus, the correlational relationship, like the stochastic one, is a property of the totality, rather than its individual elements.

Thus, it is possible not only to state that there is a correlation between the independent variable x and the dependent variable y, but also to determine how the average y changes with the change of x by one. *Impact effects* of x on y are determined by the ratio of the increments of the average group values D_y:D_x. For example, in the second group, the depth of development of the coal layer is greater by 200 m as compared to the first group, and the capital intensity of coal production is higher by 22.9 − 20.0 = 2.9 UAH/t.

That is:

$$\frac{\Delta y}{\Delta x}=\frac{2.9}{200}=0.0145.$$

So, with the increase in the depth of development of the bed by 100 m, the capital/output ratio increases on average by 1.45/t UAH. Similarly, the calculated effect of the depth of development of the bed on the capital/output ratio of coal is 0.95 UAN/t in the third group, 0.65 UAH in the fourth group.

Task 2. It is necessary to derive an equation of linear regression dependence between the yield of cereals and the amount of fertilizer applied (in quintals without changing the primary nutrient).

The solution.

Let's consider the procedure for calculation of linear regression parameters using the relationship between the yield of cereals and the amount of fertilizer applied (in centers of active nutrient). The values of the interrelated characteristics and the necessary parameters for the calculation of the parameter values are given in Table 9.10.

Table 9.10

Calculation of linear regression parameters, theoretical values and residual values

The farm number	Amount of the fertilizer X, p.n.	Yield of cereals Y, c/ha	xy	x^2	$\hat{y}$	$y-\hat{y}$	$(y-\hat{y})^2$
1	1.1	23	25.3	1.21	24	−1	1
2	1.4	25	35.0	1.96	27	−2	4
3	1.2	26	31.2	1.44	25	1	1
4	2.0	33	66.0	4.00	33	0	0
5	1.5	27	40.5	2.25	28	−1	1
6	1.3	2,8	36.4	1.69	26	2	4
7	1.8	30	54.0	3.24	31	−1	1
8	1.7	32	54.4	2.89	30	2	4
Total	12.0	224	342.8	18.68	224	×	16

The interim results obtained from the calculation table are:

$$\Sigma x = 12; \quad \Sigma y = 224; \quad \Sigma xy = 342.8; \quad \Sigma x^2 = 18.68;$$
$$\overline{x} = 12 : 8 = 1.5; \quad \overline{y} = 224 : 8 = 28.$$

Using these quantities, we determine:

$$b = \frac{8 \times 342.8 - 12 \times 224}{8 \times 18.68 - 12 \times 12} = \frac{54.4}{5.44} = 10.0 \text{ (c/ha)};$$
$$x = 28 \quad 10.0 \times 1.5 = 13.0 .$$

So the regression equation looks like this $Y = 13.0 + 10.0x$, that is, each centner of fertilizer (in terms of the primary nutrient) yields an average crop increment of 10 c/ha. If fertilizers are not applied at all (x = 0), the grain yield will not exceed 13.0 c/ha.

The regression equation does not represent the law of dependence between x and y for the individual elements of the population but for the population as a whole; the law that abstracts the influence of other factors proceeds from the principle of "other things being equal". Under these conditions, the expected yield of cereals with the use of fertilizers in the amount of 1.1 c/p.n. per 1 ha is Y = 13 + 10 × 1.1 = 24 (c/ha). For other values of the independent variable x, the theoretical yield levels are given in Table 9.10. The influence of other factors (except x-factors) causes the deviation of empirical values from the theoretical to one side or the other. The deviation $y - \hat{y}$ is called *residual* and it is denoted by the symbol e. The residuals are usually smaller than the deviation from the average, that is $(y - Y) \leq \left(y - \overline{y}\right)$.

In our example:

$$\sum_1^n \left(Y - \overline{Y}\right)^2 = 84 , \quad \sum_1^n \left(Y - \hat{Y}\right)^2 = 16 .$$

Accordingly, *the total* yield *variance* is:

$$\sigma_y^2 = \frac{1}{n} \sum_1^n (y - \overline{y})^2 = \frac{84}{8} = 10.5 ,$$

the residual variance is:

$$\sigma_e^2 = \frac{1}{n}\sum_1^n (y - Y)^2 = \frac{16}{8} = 2.$$

In small populations, the regression coefficient is prone to random fluctuations. Therefore, it is necessary to check its significance. When the relationship is linear, the significance of the regression coefficient is verified by the t-test (Student) whose statistical characteristic for the hypothesis H_0: $b = 0$ is determined by the ratio of the regression coefficient b to its own standard error μ_b, that is $t = b/\mu_b$.

The standard error of the regression coefficient depends on the variation of the independent variable σ_x^2, residual dispersion σ_e^2 and the number of degrees of freedom $k = n - m$, where m is the number of the regression equation parameters:

$$\mu_b = \sqrt{\frac{\sigma_e^2}{\sigma_x^2(n - m)}}.$$

For the linear function m = 2. According to Table 9.9 we have:

$$\sigma_x^2 = \frac{18.68}{8} - 1.5^2 = 0.085, \quad \sigma_e^2 = 2.$$

From here, $\mu_b = \sqrt{\frac{2}{0.085(8 - 2)}} \approx 2.0$ (c/ha), and $t = \frac{b}{\mu_b} = \frac{10}{2} = 5$, exceeding the critical value of the two-sided t-test $t_{0.95(6)} = 2.45$. The hypothesis about the random nature of the regression coefficient is rejected, and therefore, with a probability of 0.95, the effect of the amount of fertilizer applied on the grain yield is considered significant.

For a regression coefficient as well as for any other random variable confidence limits are defined: $b \pm t\mu_b$. In our example, the confidence limits of the regression coefficient with a probability of 0.95 (t = 2.45) are $10.0 \pm 2.45 \times 2.0$.

An important characteristic of the regression model is the relative effect of the independent variable x on the result Y – *the coefficient of elasticity*:

$$Y = b\frac{\overline{x}}{\overline{y}}.$$

It shows how much on average the result y changes with a factor x change by 1 %. In our example $Y = 10.0\frac{1.5}{28} = 0.8035$, that is, an increase in the amount of fertilizer applied by 1 % causes an increase in the grain yield by an average of 0.8 %.

Task 3. Based on the data given in Task 2, estimate the relationship between the amount of fertilizer applied and the yield of the grain.

The solution.

In practice, various modifications of the correlation coefficient formula are applied. To estimate the tightness of the relationship between the amount of fertilizer applied and the yield of grain, we use one of the modifications of the formula:

$$r = \frac{\sum_{1}^{n} xy - n\overline{x}\,\overline{y}}{\sqrt[n]{\sigma_x^2\,\sigma_y^2}}.$$

According to Table 9.10, $\sum_{1}^{n} XY = 342.8$; $\overline{X} = 1.5$; $\sigma_x^2 = 0.085$; $\overline{y} = 28$; $\sigma_y^2 = 10.5$.

According to these values, the correlation coefficient is 0.900, which indicates a significant influence of the amount of fertilizer applied on the grain yield:

$$r = \frac{342.8 - 8\times1.5\times28}{\sqrt{0.085\times10.5}} = 0.900.$$

The correlation coefficient, when evaluating the tightness of the relationship, also indicates its direction: when the relationship is direct, r is positive and when it is reverse, it is negative. The signs of correlation and regression coefficients are the same, their values are interrelated functionally:

$$r = b\frac{\sigma_x}{\sigma_y}; \qquad b = r\frac{\sigma_y}{\sigma_x}.$$

This makes it possible to calculate one factor knowing the other one. For example:

$$r = 10.0\sqrt{\frac{0.085}{10.50}} = 0.900.$$

The deviation of the individual value of the variable from the mean $Y - \overline{Y})$ can be broken down into two components. In regression analysis, this is the deviation from the regression line $(Y - \hat{Y})$ and the deviation of the regression line from the mean $\left(\hat{Y} - \overline{Y}\right)$.

The deviation $\left(\hat{Y} - \overline{Y}\right)$ is a consequence of the impact of the factor X, the deviation $Y - \hat{Y}$ is due to other factors. The relationship between the factor and residual variations is described by the decomposition rule of variation:

$$\sigma_y^2 = \delta_Y^2 + \sigma_e^2,$$

where $\sigma_y^2 = \frac{1}{n}\sum_1^n\left(Y - \overline{Y}\right)^2$ is the total variance of the variable y;

$\delta_Y^2 = \frac{1}{n}\sum_1^n\left(Y - \overline{Y}\right) = \frac{1}{n}\left(a\,\Sigma x + b\,\Sigma XY\right) - \overline{Y}^2$ is the factor variance;

$\sigma_e^2 = \frac{1}{n}\sum_1^n\left(Y - \hat{Y}\right)^2$ is the residual variance.

Obviously, the value of the factor variance δ_Y^2 increases with the increase of the influence of the factor x on y. The ratio of the factor variance to the total variance is considered as a measure of the correlation tightness; it is called *the coefficient of determination*:

$$R^2 = \frac{\delta_Y^2}{\sigma_y^2}.$$

If, according to Table 9.10, $\sigma_y^2 = 10.5$, $\sigma_e^2 = 2.0$, then $\delta_Y^2 = 10.5 - 2.0 = 8.5$. The calculations give a similar result:

$$\sigma_Y^2 = \frac{1}{8}(13 \times 224 + 10 \times 342.8) - 28^2 = 8.5.$$

The coefficient of determination is $R^2 = \frac{8.5}{10.5} = 0.81$, that is 81 % of the variation in the grain yield depends on the variation in the amount of the fertilizer applied, and 19 % is due to other factors.

The square root of the coefficient of determination is called *the correlation index R.* When the relationship is linear, $R = |r|$, it is confirmed by the calculations: $R = \sqrt{R^2} = \sqrt{0.81} = 0.90$. Therefore, knowing the linear correlation coefficient r, one can determine the contribution of the variable x to the variation of the variable y.

Task 4. It is necessary to determine the coefficient of rank correlation based on expert assessments of the economy efficiency and the degree of political risk for seven countries with transformational economies (Table 9.11).

Table 9.11

Calculation of the rank correlation coefficient

No.	Expert assessments, points		Rank		$d_j = Rx_j - Ry_j$	d_j^2
	Economy efficiency (max = 10)	The degree of political risk (max = 100)	Rx_j	Ry_j		
1	6.6	64.5	1	7	−6	36
2	5.8	57.8	2	6	−4	16
3	2.9	23.6	6	1	5	25
4	3.4	36.2	5	4	1	1
5	4.5	45.3	3	5	−2	4
6	2.7	28.4	7	2	5	25
7	4.2	32.7	4	3	1	1
Total	×	×	×	×	×	108

The solution.

Because peer reviews are given in scores, countries must be ranked. Economy with the highest efficiency is rated 1; that with the lowest efficiency being given the rank n = 7. Political risk is estimated in the opposite way with 1 being the lowest and 7 the highest risk.

The sum of squares of rank deviations $\overset{n}{\underset{1}{\Sigma}} d_j^2 = 108$, and the rank correlation coefficient is:

$$\rho = 1 - \frac{6 \times 108}{7(49-1)} = 1 - \frac{648}{336} = -0.928.$$

The value of the rank correlation coefficient indicates that there is an inverse and sufficiently high correlation between the economy efficiency and the degree of political risk.

The critical value of the rank correlation coefficient (Table 9.12) for the significance level α = 0.05 and n = 7; $\rho_{0.95}(7) = 0.71$. Therefore, the relationship significance is proved with a probability of 0.95.

Table 9.12

The critical values of the Spearman rank correlation coefficient for α = 0.05

Sample size n	5	6	7	8	9	10	11	12
$\rho_{0.95}(n)$	0.90	0.83	0.71	0.64	0.60	0.56	0.53	0.50

If two or more elements of a population have the same variable values, they are given an average rank. For example, suppose the second largest variable value has three elements in the population (No. 2, 3, 4), then all of them are given a rank $\frac{1}{3}(2+3+4)=3$, and the tightness of the relationship can be estimated by the formula of the linear correlation coefficient.

Reference to laboratory work

The guidelines for performing laboratory work on the topic "Mastering the skills in the correlation, regression and variance analysis in MS Excel" are presented in [100]. The laboratory work aims to gain practical skills in conducting the correlation, regression and variance analysis with the help of the package MS Excel.

Questions for self-assessment

1. Explain the concept of functional and correlation relationships. How does the correlation relationship manifest itself? What problems are solved by correlation and regression analysis?
2. Give a definition of multiple correlation.
3. What are the regression coefficients in the multiple regression equation?
4. What is the difference of the coefficient of elasticity from the regression coefficient?
5. What is the essence of the coefficient of elasticity and the β-coefficient?
6. What coefficients are used to quantify the relationship of social phenomena?
7. What is a nonparametric indicator of relationship?
8. What ranking ratios do you know?
9. What is the reason for the need to check the significance of a relationship?
10. How is the null hypothesis formed? Under what conditions is it accepted, and under what conditions is it rejected?

Questions for critical rethinking (essays)

1. The correlation and regression analysis of investment attractiveness of enterprises.
2. Correlation analysis of enterprise finances.
3. Correlation studies in psychology.
4. Correlation analysis in sociology. A two-dimensional distribution of two variables: visual image, correlation coefficients.
5. The correlation and regression analysis in the study of social phenomena.

10. Analysis of the intensity of dynamics

Basic questions:

10.1. The concept of time series and their types.

10.2. Analytical indicators of time series.

10.1. The concept of time series and their types

Social phenomena are constantly changing both in space and time. Studying the gradual development and change of social phenomena is one of the main tasks of statistics. It is gained on the basis of analysis of series of dynamics (sometimes called *dynamic, chronological,* or *time* series).

A time series in statistics is a sequence of values of statistical indicators that characterize the change in the magnitude of a social phenomenon or process.

Each time series consists of two *items* [17; 20; 34]:

time period (t), for which (or as of which) the numerical value is given;

series level (Y) – a specific value of a particular indicator.

The series levels have certain *features*:

the level of future time depends on the level reached in the previous time;

the longer the time interval between the events, the more different their quantitative and qualitative state is.

For example, the number of university graduates in the speciality "Economics" of the educational and professional program "Business Analytics and Management" of the first (bachelor) degree changed in 2017 compared to 2010. But not only the number of graduates changed; the number of educational and professional programs which cover the speciality "Economics", the qualitative level of training of students, the requirements for the practical training of specialists also changed.

Thus, the construction and analysis of time series make it possible to reveal the patterns of development of phenomena of social life and their peculiarities.

To properly analyze time series you need to know the types of them, which are distinguished by grouping the elements of a series based on different grounds.

Classification of time series is carried out based on the following features:

the statistical nature of the indicator [33; 34; 43; 54]:

primary indicators (the population of a city);

derivative indicators (the female population of the city 15 – 49 years old);

the nature of the dynamics presentation:

moment indicators that represent the state of a phenomenon at a certain point in time (the euro exchange rate every day, the inventory at the beginning of the day, population according to the census);

interval indicators that characterize the phenomena over a period of time (store turnover per quarter, household income per month, production of products per year).

The level of the interval time series depends on the length of the time period it characterizes. The moment time series levels do not depend on the time interval between the dates.

Both primary and derivative indicators, calculated on the basis of interval series, as opposed to moment ones, depend on the length of the time interval (average daily or average annual electricity production per capita).

If the levels of an interval series are given in absolute terms, it is possible to proceed from a time series with small time intervals to larger intervals of time, that is, it makes sense to summarize their levels.

By consistent adding of interval levels, you obtain accumulated results over a certain period (the payroll in a company, the number of man-hours worked). Summing up the levels of moment time series as such is not possible, because the obtained quantities are of no economic importance (the remainder of goods in the trading network changes daily, so the sum of them over the working week does not make sense).

The ways the levels of a series are presented can be the following*:*

rows of *absolute values* (the size of the bank's credit resources, the company's profit);

rows of mean values (average selling price, average salary);

rows of relative values (change in household income, change in capital structure).

According to *completeness of time, series are:*

complete, when dates or periods follow each other at regular intervals;

incomplete, when the intervals of time are not equal.

As to *the number of indicators* they are:

one-dimensional, characterizing the change in one indicator (gold production);

multidimensional, characterizing the change in two or more indicators.

Multidimensional time series in turn are divided into:

parallel, showing the dynamics of either the same indicator for different objects (change in the profit of enterprises in a particular industry) or different indicators for the same object (coal, oil and gas production in a region),

series of interrelated indicators.

The relationship between the indicators of a multidimensional time series may be *functional* (population change, including men and women) or *correlative* (the effect of changes in the amount of capital of a commercial bank on its profit) [34; 53].

Analysis of the dynamics of social phenomena, as a rule, is performed on the basis of multidimensional time series. They make it possible to estimate the intensity and describe the nature of the development of all components, to carry out a comparative analysis of the dynamics of two or more phenomena, to estimate the influence of the intensity of the development of one phenomenon on the other, to make scientifically sound forecasts.

Dynamic series of economic indicators, unlike time series in mathematical statistics, are mostly nonstationary, they demonstrate a trend that shows changes in the economy. Along with dynamics, economic processes have such a property as inertia: the mechanism of formation of phenomena and the nature of development (pace, direction, fluctuations) are preserved. The dialectical unity of variability and constancy, dynamism and inertia shape the nature of dynamics, providing a fundamental possibility of statistical forecasting of socio-economic development.

Some basic *rules for constructing time series* are [3; 14; 19; 34; 46]:

1) all time series indicators should be *plausible and scientifically sound*;

2) time series levels should be *comparable* to each other as for:

the territory (changing borders of a country, region or district);

the method for calculation of indicators (the time spent on performing a particular job by one employee and one worker for a certain period of time);

the period or time of observation (phenomena with seasonal nature of levels – the number of snowboards sold in the winter increases, so it is not correct to combine these data into one time series according to the critical moment of registration, because its levels are attributed to different dates of registration);

the object of observation, that is, all levels of a time series must relate to the same object of observation;

the observation units (in the study of a number of industrial enterprises of a city, the observation unit is the enterprise. Therefore, it is necessary to clearly identify which enterprises belong to the production sector and which do not);

the extent of coverage of units of a population (according to the NACE, construction is classified as services, while according to the balance of the national economy, this sector belongs to the sphere of material production);

the structure of the population (birth rates and death rates due to age structure of population over different years);

the units of measurement;

3) the length of the time series is determined by to *the period of study of the previous history* of development (or process)).

Thus, to build the predictive value of a particular feature, you need to know the background of its development. In this case, the entire time series cannot be considered as a single period of prehistory. For example, to forecast GDP in the country in 2019, the previous history of the study of changes in this indicator defines the period 1992 – 2018; however, at that time, the monetary units in Ukraine were ruble, coupon, coupon karbovanets, and hryvnia.

4) ensuring comparison of levels in the same intervals with simultaneous use of several time series in the analysis.

For example, to characterize the efficiency of the use of fixed assets, the enterprise profit over a period of time (the interval series) and the value of fixed assets at the end of the year (moment series) are compared. Comparison of indicators is possible if the absolute levels of the moment time series are averaged (i.e. it is necessary to calculate the average annual cost of fixed assets and to use changes in this indicator in further comparisons).

Comparability of data is a necessary condition for analysis of a time series.

If the main *cause of incomparability* is data, in particular [21; 32; 34]:

the incomparability of units or changes in prices for cost indicators, comparability can be achieved by *direct conversion of data using conversion rates, indexes, exchange rates*;

change in the population structure, the data can be made comparable through the *use of a standardized structure*;

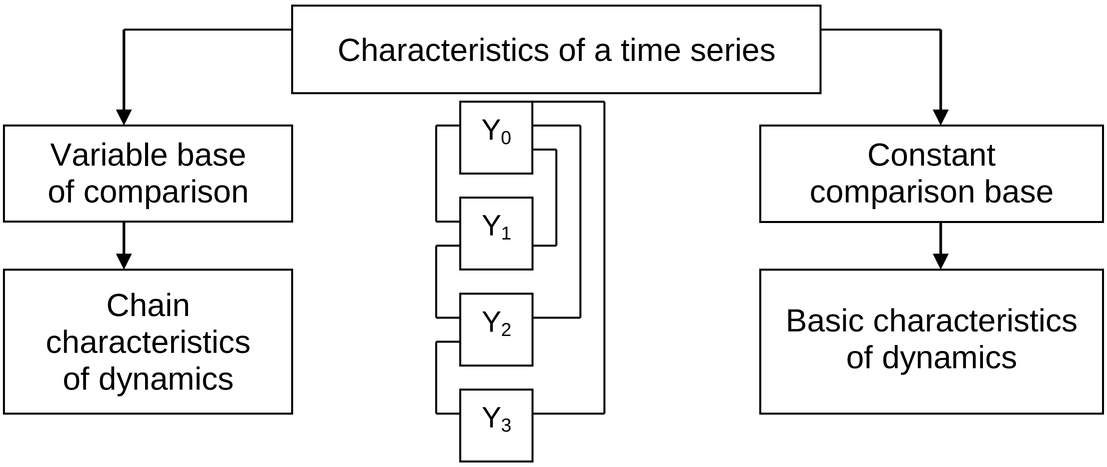
change in the range of the objects of study, territory, accounting and calculation methodology, the *way* to make data comparable is the use of special methods of closing a torn time series – *"statistical keys"* (two series are combined based on the ratio of the transition period levels, or the transition period level is the base of comparison for which a closed series of percentage ratios is formed).

The task of statistics is to determine the intensity of development of socio-economic phenomena through the analysis of a time series and to identify and describe the trends in this development, to evaluate structural shifts, constancy, and series fluctuations, to identify the factors of economic growth.

To study the dynamics of socio-economic phenomena, certain statistical characteristics are used.

10.2. Analytical indicators of time series

To estimate the direction and magnitude of changes in series levels over time, statistics uses interrelated characteristics, in particular: absolute increment, growth rate, increment rate and the absolute value of 1 % increment, acceleration (deceleration) rate. The order of calculation of the basic and chain characteristics of a time series is schematized in Fig. 10.1.


Characteristics of a time series
Variable base of comparison
Chain characteristics of dynamics
Y₀
Y₁
Y₂
Y₃
Constant comparison base
Basic characteristics of dynamics

$\Delta Y = Y_i - Y_{i-1}$	Absolute increment	$\Delta Y = Y_i - Y_0$
$G_r = \frac{Y_i}{Y_{i-1}}$	Growth rate	$G_r = \frac{Y_i}{Y_0}$
$\Delta G_r = G_r - 1(100)$	Increment rate	$\Delta G_r = G_r - 1(100)$
$A = 0.01 \times Y_{i-1}$	Absolute value of 1 % increment	$A = \frac{Y_{i-1}}{100}$
$\Delta^* = \Delta Y_i - \Delta Y_{i-1}$	Acceleration (deceleration) rate	$\Delta\% = \frac{\Delta Y_i}{\Delta Y_{i\ 1}}$
absolute		relative

Fig. 10.1. **Analytical characteristics of a time series**

The calculation of the dynamics characteristics is based on the comparison of the levels of a series. The level of the dynamics being compared is called *current (Y_1)*, and the level with which the comparison is made, is *basic (Y_0)*. The base for comparison can be either the previous level or the start-up. In the first case, the base of comparison is variable, in the second case it is constant. The base of comparison is determined depending on the purpose of study. Characteristics of dynamics calculated by comparison of adjacent levels are called *chain,* and those calculated by comparison with a constant base of comparison are *basic.*

Comparing the levels of a time series with each other is the methodological basis for assessing the intensity of dynamics. Chain characteristics estimate the intensity of changes at separate periods within the period of study, and the basic characteristics indicate the final result of changes over the whole period.

Absolute growth (ΔY) shows the absolute rate of change of series levels over a period of time. It is calculated as the difference of the levels of the row, the signs "+" or "−" indicate the direction of the dynamics.

Chain and basic absolute increments are additively related: the sum of the chain absolute increments is equal to the total basic increment over the entire period. That is:

$$\sum_1^n \Delta Y = \sum_1^n (Y_i - Y_{i-1}) = Y_n - Y_0 \,. \tag{10.1}$$

The absolute increase, depending on the statistical nature of the indicator, may be a relative value. For example, at a particular firm, the share of water pollution in the total amount of environmental protection cost was 82 % in 2016, and 79 % in 2018, i.e. decreased by three percent.

The intensity of change in the levels of a series is estimated by the relative value – *growth rate* (G_r), which is a multiple of the level ratios in the form of a ratio or percentage.

The growth rate indicates how many times the level of the period under study is greater or less than any level taken as the base of comparison.

There is a multiplicative relationship between the chain and basic growth rates: the product of the chain growth is equal to the basic rate of growth:

$$\prod_1^n G_r = \prod_1^n \frac{Y_n}{Y_{n-1}} = \frac{Y_n}{Y_0} \,. \tag{10.2}$$

The ratio of the absolute increment and the basic level is *a measure of the relative rate of growth*. Algebraic transformations of this relation give a deviation of the growth rate from the base of comparison, which is 100 %. The relative rate of growth is called *increment rate* (ΔG_r) which (as opposed to growth rate) is always expressed as percentage.

$$\Delta G_r = \frac{\Delta Y}{Y_{i-1}} - 100 = 100(G_r - 1). \tag{10.3}$$

Chain increment rates do not have such properties as additivity or multiplicity. They correlate with basic increment rates though growth rates.

The *absolute value of a 1 % increment* (A) is the quotient of division of the absolute gain by a corresponding rate of increment. Algebraically, this value is expressed by one hundredth of the previous level (taken as the base of comparison) of a time series:

$$A = \frac{\Delta Y}{\Delta G_r} = 0.01 \times Y_{i-1}. \tag{10.4}$$

For basic increment rates, values (A) are the same. The weight of the percentage increment depends on the basic level.

Comparative analysis of the dynamics intensity is based on the ratio of unidirectional characteristics of the rate of development of processes. If the rate of development of a phenomenon (process) within the analyzed period is not the same, then, by comparison of the same characteristics of rate, the acceleration or deceleration of dynamics is determined.

To evaluate changes in the rate of dynamics at different stages of development of a certain process (phenomenon) the indicator of *absolute and relative acceleration (deceleration)* is applied.

Absolute gains (increments) indicate the rate of change of the levels of a series per unit of time, and their systematic increase shows that the series develops with acceleration. With a systematic reduction, the difference of absolute chain increments is characterized by an *absolute slowdown of the series.* In statistics, *absolute acceleration (deceleration)* is the difference between the next and previous absolute increments. Acceleration (deceleration) shows how much the rate considered is greater (less) than the previous one, that is, absolute acceleration (deceleration) is the rate of the change of rate.

$$\Delta^* = \Delta Y_i - \Delta Y_{i-1}. \tag{10.5}$$

If the time intervals as to the time series indicators are the same, you can compare the basic characteristics of the rate; if not, medium rates should be used. Absolute acceleration (growth) is characterized by a positive value ($\Delta^* > 0$), slowdown is negative ($\Delta^* < 0$) [5; 10; 15; 22; 34].

Comparing the growth (increment) rate, you get *the rate of acceleration (deceleration) of the relative speed of development.* For the sake of clarity and convenience of interpretation, the divisor is the greater value of growth (gain). Both should be in one direction.

If the chain growth rates are systematically increasing (decreasing), the series evolves with *relative acceleration (deceleration):*

$$\Delta\% = \frac{\Delta Y_i}{\Delta Y_{i-1}}. \tag{10.6}$$

The value obtained is expressed as percentage points.

Growth points (percentage points) can be summed up: the sum of the growth points is equal to the total increment rate over the whole period.

So, *relative acceleration (deceleration)* is the increment rate of the absolute increment rate.

In order to compare the relative speed of two parallel processes or phenomena *the lead factor (K_{el})* is used.

In comparative analysis K_{el} is determined by the ratio of the growth rate of one indicator based on different objects or different indicators based on one object.

For multiple time series that present different economic phenomena over the same periods of time, K_{el} is equal to:

$$K_{el} = \frac{G_r \text{ greater in value for a given indicator in the appropriate time period}}{G_r \text{ smaller for the indicator being compared in the same time period}}. \tag{10.7}$$

The calculation of this indicator allows you to clearly identify for which time series the intensity of changes in levels is the highest.

Comparisons of increment rates of related indicators are used to determine the *coefficient of elasticity* (K_{el}):

$$K_{el} = \frac{\Delta G_{rx}}{\Delta G_{ry}}. \tag{10.8}$$

The coefficient of elasticity shows the percentage of change of the dependent variable Y with the change of the independent variable X by 1 %.

For example, if the price of the sales of goods increased by 13 % and demand decreased by 7 %, the price elasticity of demand for goods would be:

$$K_{el} = \frac{-7}{13} = -0.54 .$$

The degree of price elasticity of demand (direct demand elasticity) shows that the demand for this product increases by 0.54 % with a 1 % increase in the price of goods sold. In most cases, this indicator is negative because, according to the law of demand, price and demand change in the opposite direction.

In the study of time series, whose levels vary or fluctuate, there is a need to calculate a constant characteristic of the period considered. This characteristic is the *average level*.

The methods for calculating the mean levels of a time series depend on the statistical structure of the indicator.

The average values for time series are [34]:

the average level of a series;

the average absolute increment;

the average growth rate;

the average rate of increment.

In economic practice, the use of averages is sometimes a prerequisite for analytical evaluation. For example, agricultural production depends on the weather conditions in a given year, so comparing the indicators that characterize the change in production of those products during the year becomes inappropriate. It would be a good idea to compare average annual levels, average annual absolute gains, or growth rates over certain time periods.

The average annual figures are used if comparison of absolute data is impossible. For example, to determine the output per capita, it is necessary to divide the absolute output by the population, which is not a constant value for the analyzed period of time. Thus, at the beginning of 2018 the population of a settlement in a region was 257.8 thousand people, while at the end it was 260.0 thousand people. To determine the volume of production per capita in 2018, it is necessary to attribute the total production to the average number of inhabitants of the settlement, which is 258.9 thousand people.

According to the theory of averages, the calculation of average values must be carried out in homogeneous groups. For phenomena that develop over time, this means that the average of the series levels must characterize, to a certain extent, the time with the same conditions of development. The overall average over the entire period under study may be supplemented by averages over the individual intervals of that period.

The order of calculation of the average level for the interval and moment time series differs.

The average level of a series ($\overline{Y}$) *is calculated as follows:*

1) for *an interval time series*:

a) when the series levels are equally spaced from each other:

$$\overline{Y} = \frac{Y_1 + Y_2 + Y_3 + ... + Y_n}{n} = \frac{\sum Y_n}{n}, \qquad (10.9)$$

where Y_n is the series level for the n-th period;

n is the number of levels of the time series.

For example, in the last five years the profit of an enterprise (in million UAH) was 12.0 in the first year; 13.4 in the second year; 14.0 in the third year; 13.8 in the fourth year; 15.6 in the fifth year.

The average annual profit of the enterprise is:

$$\overline{Y} = \frac{\sum Y_n}{n} = \frac{12.0 + 13.4 + 14.0 + 13.8 + 15.6}{5} = 13.76 \text{ million UAH;}$$

b) when the levels of a time series are not equally spaced from each other:

$$\overline{Y} = \frac{\sum Y_n \times t_n}{\sum t_n}, \qquad (10.10)$$

where t_n is the length of time during which the level value Y_n is retained.

For example, the average number of employees in a firm was 250 in the first quarter, 245 in the second quarter, 233 in the second half year.

The average number of employees in the firm in the given year is:

$$\overline{Y} = \frac{\Sigma Y_n \times t_n}{\Sigma t_n} = \frac{250 \times 3 + 245 \times 3 + 233 \times 6}{12} = 240.25 \text{ people;}$$

2) for *a moment time series*:

a) with equal time intervals between the dates:

$$\overline{Y} = \frac{\frac{1}{2}Y_1 + Y_2 + ... + \frac{1}{2}Y_n}{n-1}, \qquad (10.11)$$

where n is the number of levels of the time series.

This kind of average is called a *simple chronological average.*

For example, the balance of goods in a trading network amounted to 140 000 UAH as of January 1; 132 000 UAH as of April 1; 150 000 UAH as of July 1; 147 000 UAH as of October 1; 142 000 UAH as of January 1 next year.

The average quarterly amount of product balance is:

$$\overline{Y} = \frac{\frac{1}{2}Y_1 + Y_2 + ... + \frac{1}{2}Y_n}{n-1} = \frac{\frac{1}{2}140 + 132 + 150 + 147 + \frac{1}{2}142}{5-1} =$$

$$= 142.5 \text{ thousand UAH;}$$

b) with unequal time intervals between the dates:

$$\overline{Y} = \frac{(Y_1 + Y_2) \times t_1 + (Y_2 + Y_3) \times t_2 + ... + (Y_{n-1} + Y_n) \times t_{n-1}}{2(t_1 + t_2 + ... + t_{n-1})} = \frac{\sum (Y_n + Y_{n+1}) \times t_n}{2 \sum t_n}, \ (10.12)$$

where Y_n is the level of the moment time series as of the date n;

t_n is the period of time between the dates (or the period during which the level Y_n is unchanged).

This kind of average is called the *weighted chronological average*. For example, in 2016 the number of company employees was 200 as of January 1; 195 as of March 1; 210 as of April 1; 215 as of September 1; 203 as of January 1, 2017.

The average annual number of employees in the company in 2016 was:

$$\overline{Y} = \frac{\Sigma(Y_n + Y_{n+1}) \times t_n}{2\,\Sigma t_n} =$$

$$= \frac{(200 + 195) \times 2 + (195 + 210) \times 1 + (210 + 215) \times 5 + (215 + 203) \times 4}{2 \times (2 + 1 + 5 + 4)} =$$

$$= 208 \text{ people.}$$

The general characteristics of the intensity of dynamics are the average absolute increment, the average growth rate, the average increment rate.

An *average absolute increment* ($\overline{\Delta}$) characterizes the average rate of the level increase (decrease) [5]. It is defined for moment and interval time series with equal time intervals between the dates as a simple arithmetic mean of chain increments:

$$\overline{\Delta} = \frac{1}{n}\Delta Y_{chain}, \qquad (10.13)$$

where n is the number of absolute chain gains.

Based on the fact that $\Sigma\Delta Y_{chain} = \Delta Y_{bas.}$, we get:

$$\overline{\Delta} = \frac{1}{n}(Y_n - Y_0), \qquad (10.14)$$

where Y_n is the last level of the time series;
Y_0 is the level taken as the base of comparison;
n is the number of levels in a series.

In the calculation of the average *growth rate* it should be borne in mind that the rate of development occurs according to the rules of compound percent, where increment accumulates on the previous increment. Therefore, the average growth rate is calculated using the formula of a geometric mean.

Average growth rate ($\overline{G_r}$) shows how many times on average each level of the series is greater (less) than the previous level [15]:

1) for time series with equal time intervals between the dates:

$$\overline{G_r} = \sqrt[n]{G_{r1} \times G_{r2} \times G_{r3} \times ... \times G_{rn}}, \qquad (10.15)$$

where $G_{r1}, G_{r2}, G_{r3}, ..., G_{rn}$ are chain growth rates;
n is the number of chain growth rates;

2) when only the start and end levels of the time series are known (with uneven time intervals between the dates):

$$\overline{G_r} = \sqrt[n-1]{\frac{Y_n}{Y_0}}, \qquad (10.16)$$

where n is the number of levels of the time series;

The *average rate of increment* is defined as:

$$\Delta\overline{G_r} = \overline{G_r} - 1(100). \qquad (10.17)$$

The considered average indicators of dynamics are simple, and they clearly interpret the result.

The average growth rate can be calculated on the basis of [34]:

chain growth rates;

final level for the whole period of growth rate;

final and basic levels of the series.

For example, at the beginning of the reporting year, the authorized capital of a company was 10 million UAH, and at the end it amounted to 11.025 million UAH. In the first half of the year, capital increased by 2.52 %, in the second half year it grew by 7.8 % with the total yearly increase of 10.25 %.

Then the average annual growth rate is:

$$\overline{G_r} = \sqrt{1.0252 \times 1.078} = \sqrt{1.1025} = \sqrt{\frac{11.025}{10.0}} = 1.05 \text{ or } 105\ \%,$$

that is, the average annual increase of the company authorized capital was 5 %.

Important concepts

Absolute acceleration (deceleration) is the difference between the subsequent and previous absolute increments.

Absolute increment is an indicator that shows the absolute rate of change of the series levels over a period of time.

Basic characteristics of *dynamics* are dynamics characteristics calculated by comparison with a constant base of comparison.

Relative acceleration (deceleration) is the increment rate of absolute increment.

Elements of a time series are 1) *time* (t), over which (or as of which) the numerical value is given; 2) *series level* (Y), a specific value of an indicator.

Interval time series are time series that characterize the phenomena over a period of time.

The coefficient of elasticity is a comparison of the rate of increment of related indicators.

Chain characteristics of dynamics are dynamics characteristics calculated by comparing the adjacent levels.

Moment time series are time series that characterize the state of the phenomenon at a certain point in time.

Comparability of data is a methodological basis for analysis of a time series.

Time series in statistics is the sequence of values of statistical indicators that characterize the change in the magnitude of a social phenomenon or process in time.

The average level of a series is the level of a series that summarizes the properties inherent in a time series over a certain period.

Growth rate is an indicator that characterizes the intensity of change in the levels of a series; it is a multiple ratio of levels in the form of a factor or percentage.

Increment rate is an indicator that characterizes the relative rate of growth.

Typical tasks

Task 1. We have some data on the growth rate of the output of an enterprise (Table 10.1). It is necessary to determine the average growth rate and draw conclusions.

Table 10.1

The growth rate of the output on a yearly basis

Years	2009	2010	2011	2012	2013	2014	2015	2016
Growth rate	1.2006	1.1059	1.1839	1.2909	1.2792	1.2326	1.3245	1.3179

The solution.

1. Find the average growth rate:

$$\overline{\mathrm{Gr}} = \sqrt[n]{G_{r1} \times G_{r2} \times ... \times G_{rn}} = \sqrt[8]{\begin{matrix} 1.2006 \times 1.1059 \times 1.1839 \times 1.2909 \times \\ \times 1.2792 \times 1.2326 \times 1.3245 \times 1.3179 \end{matrix}} = 1.2399 .$$

Thus, the annual output of the company increased by an average of 23.99 % in the period 2009 – 2016.

Task 2. We have some data on changes in the environmental protection costs of a production company (Table 10.2).

Table 10.2

Estimates of a time series

Years	The cost of environmental protection, million UAH (Y)
2010	11.2
2013	7.7
2014	7.3
2015	6.9
2016	7.0
2017	7.1

It is necessary to calculate the absolute and relative indicators of the dynamic distribution series. The results must be presented in a table form.

The solution.

Let's choose 2013 as the base of comparison, because there was an annual change in the series level (Y) that year.

Table 10.3 presents the input information for the calculations.

In the period of 2013 – 2017, the company cost of environmental protection was reduced by 0.6 million UAH, and annual comparisons with 2013 give negative indicators. Therefore, the chain absolute increase in environmental protection cost in 2016 and 2017 was equal to 0.1 million UAH. In 2013 – 2017, the environmental expenditures were reduced by 7.79 % (92.21 – 100). Only in 2016 and 2017, there was a slight annual increase of 1.45 % and 1.43 %, respectively.

Table 10.3

The input data of the time series

Years	Cost of environmental protection, million UAH (Y)	Estimated indicators						Absolute value of a 1 % increment, (A), million UAH
		Absolute increment (ΔY), million UAH		Growth rate (G_r), %		Increment rate (ΔG_r), %		
		chain	basic	chain	basic	chain	basic	
1	2	3	4	5	6	7	8	9
2010	11.2	–	–	–	–	–	–	–
2013	7.7	–	–	–	–	–	–	–
2014	7.3	−0.4	−0.4	94.81	94.81	−5.19	−5.19	0.077

Table 10.3 (the end)

1	2	3	4	5	6	7	8	9
2015	6.9	−0.4	−0.8	94.52	89.61	−5.48	−10.39	0.073
2016	7.0	0.1	−0.7	101.45	90.91	1.45	−9.09	0.069
2017	7.1	0.1	−0.6	101.43	92.21	1.43	−7.79	0.070

Thus, in 2014 (compared to 2013), the reduction of environmental expenditures by 1 % equalled 0.077 million UAH; in 2016 the increase in expenses by 1 % was 0.069 million UAH, respectively. Thus, over the considered period, environmental spending by the company tended to decrease.

Task 3. Using the information of task 2 about the company environmental protection costs (Table 10.4), we determine the average of the time series.

The solution.

The considered time series is interval by the nature of the indicator; so the average level of the series for 2013 – 2017 is:

$$\overline{Y} = \frac{\Sigma Y_n}{n} = \frac{7.7 + 7.3 + 6.9 + 7.0 + 7.1}{5} = \frac{36}{5} = 7.2 \text{ million UAH},$$

that is in 2013 – 2017, the average level of environmental protection costs of the company amounted to 7.2 million UAH.

The average absolute increase over the given period is:

$$\overline{\Delta} = \frac{1}{n}\Delta Y_{chain.} = \frac{(-0.4) + (-0.4) + 0.1 + 0.1}{4} = \frac{0.6}{4} = 0.15 \text{ million UAH},$$

that is, over the given period, there was an average annual absolute decrease in environmental protection expenditures by 0.15 million UAH.

The average growth rate is calculated over different time periods:

1) 2010 – 2013:

$$\overline{G}_r = \sqrt[n-1]{\frac{Y_n}{Y_0}} = \sqrt[4-1]{\frac{7,7}{1.2}} = \sqrt[3]{0.6875 =}\; 0.85, \text{ or } 85\ \%,$$

that is, over the given period, the annual reduction of environmental costs of the manufacturing firm was on average 15 %;

2) 2013 – 2017:

$$\overline{G_r} = \sqrt[n]{G_{r1} \times G_{r2} \times \ldots \cdot G_{rn}} = \sqrt[4]{0.9481 \times 0.9452 \times 1.0145 \times 1.0143} =$$

$$= \sqrt[4]{0.9221} = 0.97993 = 0.98 \quad \text{or } 98\ \%.$$

or:

$$\overline{G_r} = \sqrt[n-1]{\frac{Y_n}{Y_0}} = \sqrt[5-1]{\frac{7.1}{7.7}} = \sqrt[4]{0.9221} = 0.98 \text{ or } 98\ \%,$$

that is, in the period 2013 – 2017, the company average annual reduction of environmental costs was 2 %;

3) 2010 – 2017:

$$\overline{G_r} = \sqrt[n-1]{\frac{Y_n}{Y_0}} = \sqrt[8-1]{\frac{7.1}{11.2}} = \sqrt[7]{0.63393} = 0.937 \text{ or } 93.7\ \%,$$

that is, over the given period, the annual reduction in environmental protection costs was on average 6.3 %.

Task 4. We have some data on the amount of tax revenues in the period 2013 – 2017 (Table 10.4). It is necessary to analyze the nontax revenues using the absolute, relative, and average values. The results of the calculations should be presented in tabular form.

Table 10.4

The amount of the nontax revenues

Years	2013	2014	2015	2016	2017
Nontax revenues, million UAH	130	156	141	150	145

The solution.

The results of the calculations are presented in Table 10.5.

Table 10.5

The results of the calculations of the time series indicators

Years	Nontax revenues, million UAH	Absolute increment, million UAH		Growth rate		Increment rate,%		Absolute value of a 1 % increment, million UAH
		base	chain	base	chain	base	chain	
2013	130	0	–	1.000	–	0	–	–
2014	156	+26	+26	1.200	1.200	+20.0	+20.0	1.30 (↑)
2015	141	+11	−15	1.085	0.904	+8.5	−9.6	1.56 (↓)
2016	150	+20	+9	1.154	1.064	+15.4	+6.4	1.41 (↑)
2017	145	+15	−5	1.115	0.967	+11.5	−3.3	1.52 (↓)

Next we calculate:

1) the average annual level of nontax revenues as a simple arithmetic average of:

$$\bar{y} = \frac{130 + 156 + 141 + 150 + 145}{5} = \frac{722}{5} = 144.4 \text{ million UAH};$$

2) the average annual increment rate:

a) depending on the base or chain absolute increments:

$$\bar{\Delta} = \frac{145 \quad 130}{5 \quad 1} = 3.75 \text{ million UAH;}$$

$$\bar{\Delta} = \frac{+26 \quad 15 \quad 5}{4} = 3.75 \text{ million UAH.}$$

3) the average annual growth rate according to the formula of a simple geometric mean:

a) depending on the base or chain absolute increments:

$$\overline{G_r} = \sqrt[4]{1.115} = 1.028;$$

$$\overline{G_r} = \sqrt[4]{1.2 \times 0.904 \times 1.064 \times 0.967} = 1.028.$$

b) the average annual growth increment:

$$\overline{Gr} = (1.028 - 1) \times 100 = 2.8\,\%$$

or

$$\overline{Gr} = 102.8 - 100 \ = 2.8\,\%.$$

Thus, it can be concluded that in 2014, compared to 2013, nontax revenues increased by 26 million UAH, and in 2015, compared to 2014, they decreased by 15 million UAH. On the whole, in the period from 2013 to 2017, nontax revenues increased by 15 million UAH (or 1.115 times, or + 11.5 %). During the same period of time, the average annual revenues amounted to 144.4 million UAH and their average annual absolute increment amounted to 3.75 million UAH. That is, in the period 2013 – 2017, on average, a yearly volume of nontax revenues increased by 2.8 %.

Reference to laboratory work

The guidelines for performing laboratory work on the topic "Acquisition of skills in the analysis of dynamics in MS Excel" are presented in [100]. The laboratory work is aimed at gaining practical skills in the analysis of time series and economic interpretation of calculated indicators with the help of MS Excel.

Questions for self-assessment

1. What is time series and what is their role in statistical analysis?
2. What are the elements of a time series?
3. What kinds of time series do you know? Give examples.
4. What conditions should be followed in the construction of time series?
5. What kind of averages are used to calculate the average of the moment and interval time series?
6. What are base and chain indicators of dynamics?
7. Name the types of indicators of dynamics and explain how they are calculated.
8. How is the average growth rate in the time series calculated?
9. What types of time series are there depending on the registration of facts?
10. What are the rules for constructing a time series?

Questions for critical rethinking (essays)

1. If the rate of development within the considered period is different, is the acceleration or deceleration of dynamics measured by comparison of the same characteristics of rate?

2. What interconnected characteristics should be used to evaluate the properties of dynamics? Explain the essence of the choice of indicators for evaluation.

3. Summarize the properties inherent in a time series by calculating different averages.

4. What are the causes of incomparability of time series levels? Give examples.

5. Explain the relationship of the absolute increment and the rate of increment. Prove that the absolute value of one percent of increment is one hundredth of the level taken as the base of comparison.

11. Analysis of development trends and fluctuations

Basic questions:

11.1. The techniques for identifying the main trend of development in a time series.

11.2. Interpolation and extrapolation.

11.3. Factor analysis of time series.

11.4. Analysis of seasonal fluctuations.

11.1. The techniques for identifying the main trend of development in a time series

Time series characterize the processes of development of socio-economic phenomena. These processes are both inherently dynamic and inertial. Dynamics is manifested in changes, fluctuations in the levels of a time series, while inertia – in the constancy of the shaping factors, direction and intensity of development. A time series contains the remnants of the past, the foundations of the present and the beginnings of the future. First of all, it is necessary to distinguish between the factors of evolutionary nature that exert a permanent influence and determine the general direction of a phenomenon, its long-term evolution. Continuous factors exert a decisive influence on the phenomenon under study and form the main, as a rule, long-term trend of development in the time series. The second group of factors is made up of *oscillatory factors*. The influence of these factors manifests itself periodically – it causes recurrent fluctuations in the levels of a time series –

seasonal, cyclical (for example, cycles of economic conditions). Schematically, cyclic oscillations can be represented as a sine wave (the value of the sign first increases, reaches a certain maximum, then decreases, reaches its minimum, again increases and so on). The other factors affecting a time series are factors that cause irregular fluctuations in the levels that cause random short-term changes in the levels of the dynamics. In turn, these factors are divided into: those that cause sporadic changes in levels (war, environmental disasters, epidemics, etc.); random, that is of little influence [6; 26; 34]. These are minor factors that lead to random multidirectional changes in levels.

Any socio-economic process (Y) is considered as a function of time (t). Of course, time is not a factor in a particular socio-economic phenomenon, it only accumulates all the conditions and causes that determine this phenomenon. Thus, any time series can theoretically be represented as the following components [22; 26; 34]:

a trend – the main long-term trend of development of a time series (to increase or decrease its levels) f(t);

cyclic (periodic) oscillations C(t);

seasonal fluctuations S(t);

random fluctuations E(t).

The relationship between these components can be represented additively (cumulatively), or multiplicatively (by product), or in a combined way:

$Y_t = f(t) + C_t + S_t + E_t$ if the nature of cyclic and seasonal fluctuations remains constant;

$Y_t = f(t) \times C_t \times S_t \times E_t$ if the nature of cyclical and seasonal fluctuations remains constant only in relation to the trend;

$Y_t = f(t) + C_t \times S_t + E_t$, a combined relationship between the components.

Consider, for example, the difference between additive and multiplicative seasonal components. The sales schedule for children's toys is likely to have annual peaks in November – December and another one – significantly lower in height – in the summer months that is the time of vacations. This seasonal pattern will be repeated every year. By its nature, the seasonal component can be *additive* or *multiplicative*. So, every year the volume of sales of some specific toy can increase in December by 3 million UAH. Therefore, you can take into account these seasonal changes by adding 3 million UAH to your December forecast. Here we have an additive seasonality. However, in December, sales of some toys may increase by 40 %, i.e.

multiply by a factor of 1.4. This means that if the average sales volume of this toy is small, the absolute increase (in monetary terms) in this volume in December will also be relatively small (but in percentage terms it will be constant). If the toy sells well, the absolute sales growth (in UAH) will be significant. Here again the sales volume increases the number of times equal to a certain factor, and the seasonal component (a multiplicative component in its nature) in this case equals 1.4. If you look at the graph of time series, the difference between these two types of seasonality will manifest itself as follows: in the additive case, the series will have constant seasonal fluctuations the value of which will not depend on the total level of values of the series; in the multiplicative case, the magnitude of seasonal fluctuations will vary depending on the overall level of the series values.

The above example can be extended to illustrate the notion of additivity and multiplicity of the trend-cyclic components. In the case of toys, the fashion trend can lead to a steady increase in sales (for example, it may be a general trend towards educational toys). Like the seasonal component, this trend can be inherently additive (annual sales increase by 3 million UAH) or multiplicative (annual sales increase by 30 % or 1.3 times). In addition, sales may include cyclical components. Let us reiterate that the cyclic component differs from the seasonal component in that it usually has a longer time length and manifests itself at irregular intervals. For example, some toys may have a surge of demand during the summer season (for example, an aggressively advertised doll depicting a popular cartoon character). As in the previous cases, such a cyclical component can change the sales volume additively or multiplicatively.

A peculiarity of studying the development of socio-economic phenomena over time is that in some time series the general trend of development (an increase or a fall of an indicator) manifests itself immediately through visual inspection of information, by plotting the source data, while in other time series the trend is not directly detected.

The study of the trend includes two main stages [34]:

a time series is checked for a trend;

aligning of the time series and direct identification of the trend is conducted with extrapolation of the results.

There are three types of trends in the socio-economic dynamics: *a middle-level trend*, *a dispersion trend, an autocorrelation trend*.

The middle level trend is analytically expressed by means of a mathematical function around which the actual levels of the phenomenon under study vary. In this case, the values of the trend at some points in time will be the mathematical expectation of a time series. Often, the middle level trend is called a deterministic (non-random) component of a time series.

The dispersion trend is the trend of change in the deviation between the empirical levels and the deterministic component of a series.

The autocorrelation trend is the trend of change of relationship between separate levels of a time series. Graphically, this change is not traceable.

The initial step in the identification and analysis of a trend is testing the hypothesis of the trend. There are about a dozen criteria for checking the trend. Let's consider some of them [22; 26; 34].

1. *Checking the significance of the difference between the averages*. A time series is split into two equal or almost equal parts. They test the hypothesis that there is a difference in the mean values: H_0: $\overline{y}_1 = \overline{y_2}$. Since the number of members of the series being analyzed is usually small, they use small-sample theory to test the hypotheses. The test is based on the Student's t_α-criterion. If $t \geq t_\alpha$, the hypothesis of no trend is rejected, and vice versa: if $t < t_\alpha$, the hypothesis (H_0) is accepted. Here t is the calculated value found for the data being analyzed; t_α is the tabular value of the criterion with the error probability level equal to α. In the case of equality or if there is no significant difference in the dispersions of two populations under study, the calculated value of t is determined by the dependence:

$$t = \frac{\overline{y_1} - \overline{y_2}}{\delta\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{11.1}$$

where $\bar{y}_1, \bar{y}_2$ are averages of the first and second half of the series;

n_1 and n_2 are the number of observations in these series;

σ is standard deviation of the difference of mean values due to the dependency:

$$\sigma = \sqrt{\frac{(n_1 - 1)^2 \times \delta_1^2 + (n_2 - 1)^2 \delta_2^2}{n_1 + n_2 - 2}}. \tag{11.2}$$

The dispersions for the first and second parts of the series are calculated according to the formula:

$$\sigma_i^2 = \frac{\sum(y_i - \overline{y_i})^2}{n-1}. \tag{11.3}$$

Testing the hypothesis of the equality of dispersions is carried out using the F-criterion, the main meaning of which is to compare the calculated relation with the table one. The calculated value of the criterion is determined by the formula:

$$F = \frac{\sigma_1^2}{\sigma_2^2}. \tag{11.4}$$

If the calculated value of F is less than the table value, with the given level of significance, the hypothesis about the equality of dispersions is accepted. If F is greater than the table value, the hypothesis of equality of dispersions is rejected and the dependence for the calculation of t is considered unsuitable for use. To fulfill the conditions for equality of dispersions, they determine the value of t_α and test the hypothesis (H_0). To do this, the theoretical value of t_α is determined by the number of degrees of freedom equal to $n_1 + n_2 - 2$. The considered method gives positive results for series with a monotonous trend. When a time series changes the general direction of development, the turning point of the trend is close to the middle of the series. Therefore, the average of the two segments will be close, and the check may not indicate the presence of a trend.

2. The *Foster – Stuart method for checking the trend*. This method, in addition to determining the presence of a trend, reveals a trend of dispersion of time series levels, which is important to know for the analysis and forecasting of economic phenomena.

The null hypothesis H_0 confirms the trend. In statistics, the criteria look like:

$$s = \sum_{i=2}^{n} S_i ; \tag{11.5}$$

$$d = \sum_{i=2}^{n} d_i , \tag{11.6}$$

where $d_i = u_i - l_i$; $S_i = u_i + l_i$.

If $x_i > x_{i-1},...,x_1$, $u_i = 1$, otherwise $u_i = 0$; if $x_i < x_{i-1},...,x_1$, $l_i = 1$; otherwise $l_i = 0$.

Statistics S are used to test the trend in dispersions, statistics d are used to detect the trend in averages.

It is obvious that $0 \le S \le n-1$; $-(n-1) \le d \le n-1$.

In the absence of a trend the values: $t = d/f$

$$\bar{t} = \frac{S - f^2}{l}, \text{ where } l = \sqrt{2\ln n - 3.4253}, \; f = \sqrt{2\ln n - 0.8456}$$

have a Student distribution with v = n degrees of freedom. The formulas for f and l are used for n > 50, their values for n < 50 are given in Table 11.1.

Table 11.1

The constants f and l of the Foster – Stuart criterion

n	10	15	20	25	30	40	45	50
f	1.964	2.153	2.279	2.373	2.447	2.509	2.561	2.606
l	1.288	1.521	1.677	1.791	1.882	1.956	2.019	2.072

If $|t|, |\bar{t}| > t_{\frac{1+\alpha}{2}}$, with confidence probability α the null hypothesis H_0 of the existence of a trend is accepted, otherwise hypothesis H_0 is rejected; $(t_\gamma – \gamma)$ is Student's quantile of the distribution.

After a trend has been established, a time series is described. Not only numerical values of levels, but also their sequence are important in the study of dynamics. As a rule, time intervals between the levels are the same (a day, a decade, a calendar month, a quarter, a year). Having taken any interval per unit, we write the sequence of levels as follows: y_1, y_2, y_3, ..., y_n.

Depending on the statistical nature of the level indicator there are primary and derivative time series, series of absolute, average and relative values. In terms of time, time series are divided into interval and moment ones. The moment level fixes the state of the phenomenon at a certain point in time t (for example, the number of employees at the beginning of the year, students as of September 1, etc.). In an interval series, the level is an aggregated result of the process which depends on the length of the time interval: electricity production per year, fishing per season. Note that, in contrast to the moment series, derivatives calculated on the basis of intervals depend on the length of time (average daily or average annual electricity production per capita).

Socio-economic processes are dynamic if they show a constant change in the levels of the time series. Along with dynamics they have inherent inertia: the mechanism of formation of phenomena and the nature of development (pace, direction, oscillations) remain. In the case of significant inertia of the process and stability of the complex of conditions of its development, it is possible to expect in the future the same properties and nature of the development that were revealed in the past. The dialectical unity of variability and constancy, dynamism and inertia shapes the nature of dynamics, making possible statistical forecasting of socio-economic processes [34].

When studying the patterns of socio-economic development, statistics performs a number of tasks: measures the intensity of dynamics, identifies and describes trends, assesses structural shifts, constancy and series fluctuations; identifies the factors that cause the change.

A prerequisite for the analysis of any time series is the comparability of the statistics that produce this series. The incomparability of the data may be due to various reasons:

changes in the methodology of registration and calculation of the indicator, including the use of different units for measurement;

changes in the structure of the population as well as territorial changes;

various critical moments of data logging or the length of the periods to which they belong; changing prices for cost indicators.

Comparability of data is ensured at the stages of data collection and processing. They also use special techniques to convert the data to a comparative form – the "statistical keys" of closing of time series.

Suppose that the monthly levels of raw material consumption for production in the first six months are not comparable, since in April the order of recording changed (Table 11.2). There are two ways to break intermittency:

the method of relative levels when the April level is taken as the base of comparison for each series. Two series of relative levels are combined into one;

the method that is based on the ratio of April levels: 55 : 50 = 1.1. Multiplying the levels of the first series by this factor, we obtain the only closed (comparable) time series for the whole period (the last graph of Table 11.2).

Table 11.2

Closing of time series

Months	Volume, tons		The series is closed	
	Former order of recording	New recording procedure	Relative values, %	Absolute values, ton
1	40	–	80	44.0
2	45	–	90	49.5
3	48	–	96	52.8
4	50	55	100	55.0
5	–	58	105	58.0
6	–	60	109	60.0

An important task of statistics in the analysis of time series is to determine the main trend of development inherent in a particular time series.

The main trend of development of a time series is the change that determines the general direction of development. This is a systematic component of long-term action. In some cases, the overall trend can be clearly seen in the dynamics of the indicator being considered, in other cases it might not be revealed due to significant random fluctuations. For example, at some points in time, strong fluctuations in retail prices may obscure the trend towards increasing or decreasing of this indicator. Therefore, *three groups of methods* are used in statistics to identify the main trend [34]:

the method of enlarging of intervals;

the methods of smoothing;

the method of analytical alignment.

1. *The method of enlarging of intervals* is based on the enlargement of time periods to which the levels belong. For example, you can convert a series of weekly data into a series of monthly dynamics, and replace a number of quarterly data with annual levels. The levels of a new series can be obtained by summing the levels of the original series or may represent the mean levels.

A common technique for detecting a development trend is to smooth a time series. The essence of different smoothing techniques is to replace the actual levels of a series by calculated levels that are less prone to fluctuations. This contributes to a clearer manifestation of the trend of development.

2. *The methods of smoothing* include the following methods: averaging, simple moving and weighted moving average.

2.1. The *left- and right-half averaging method*. A time series is divided into two parts, the arithmetic mean is found for each of them and a trend line is drawn through the points obtained.

2.2. The s*imple moving average method*. With this method, the output levels of a series are replaced by the average values obtained from a given level and several levels symmetrically arranged around it. The whole number of levels from which the average value is calculated is called the smoothing interval. However, the calculation of the average is carried out with the gradual exclusion of the smoothing interval of the first level of a time series from the given interval and the inclusion of the following one: $\bar{y}_1 = \frac{y_1 + y_2 + y_3}{3}$; $\bar{y}_2 = \frac{y_2 + y_3 + y_4}{3}$, etc.

The interval can be odd (3, 5, 7, etc. levels) or even (2, 4, 6, etc. levels).

With *odd smoothing*, the obtained arithmetic mean corresponds to the middle of the calculation interval, with even smoothing, the mean corresponds to the interval between the dates.

To smooth a time series with an even number of levels, perform an additional operation called centering, because, when calculating a moving average, for example, four levels $\bar{y}_1 = \frac{y_1 + y_2 + y_3 + y_4}{4}$, $\bar{y}_2 = \frac{y_2 + y_3 + y_4 + y_5}{4}$ its value refers to the point between the points in time when the actual levels of y_2 and y_3 were recorded:

$y_1, y_2, y_3, y_4, y_5, y_6$... are the input levels;

$\overline{y_1}, \overline{y_2}, \overline{y_3}$ … are the smoothed levels;

$\overline{y_{1ц}}, \overline{y_{2ц}}$ ... are the centered smoothed levels: $\overline{y_{1ц}} = \frac{\overline{y_1} + \overline{y_2}}{2}$, $\overline{y_{2ц}} = \frac{\overline{y_2} + \overline{y_3}}{2}$.

Let's consider the use of a moving average based on the example of a time series of sales of smartphones in a trading network in 2017 (Table 11.3):

$$y_1 = \frac{23 + 25 + 21}{3} = 23; \;\; y_2 = \frac{25 + 21 + 26}{3} = 24, \text{ etc.}$$

Table 11.3

The dynamics of sales of smartphones in a trading network in 2017

Month	Sold smartphones, thousand pieces	Three-level moving average	Four-level moving averages, not centered	Four-level moving averages, centered
1	23	–	–	
2	25	23	23.8	
3	21	24	25.0	24.4
4	26	25	24.8	24.9
5	28	26	26	25.8
6	24	27	26.8	27
7	29	27	27.3	27.5
8	28	29	27.8	28.4
9	30	29	29	29.3
10	29	30	29.5	30.1
11	31	31	30.8	–
12	33	–		–

In Fig. 11.1 the smoothing of highly fluctuating levels can be seen, which is done using the moving average.

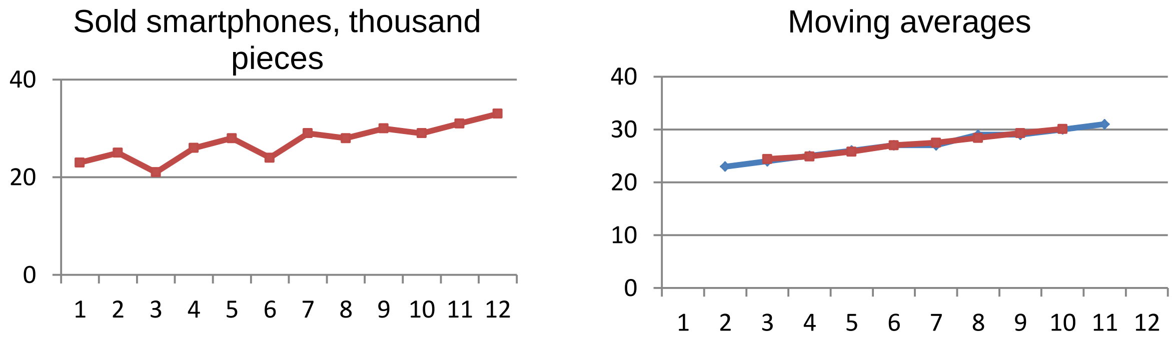


Fig. 11.1. **The dynamics of smartphone sales in a trading network in 2017**

2.3. The *weighted moving average* method. The main difference of this method from the previous one is that the levels included in the averaging interval are given different weights, because the approximation within the smoothing interval is carried out using the levels calculated by the n-th order polynomial:

$$\overline{y_i} = a_0 + a_1 \times x_i + a_2 \times x_i^2 + ... \,, \tag{11.7}$$

where i is the sequence number of the smoothing interval level.

The disadvantage of the simple moving average method is that the smoothed time series is reduced due to the inability to obtain smoothed levels for the beginning and end of the series. This drawback is eliminated by the use of the analytical alignment method to analyze the main trend.

3. *The method of analytical alignment* is the most effective way of determining the main trend of development of the phenomenon under study which manifests itself over time. Analytical alignment is the selection (for a given time series) of the theoretical curve which expresses the main features of the dynamics of the actual phenomenon, that is, describes the empirical data in the best way.

As a result of the alignment of a time series the most common, total series is obtained, which manifests itself in time as a result of all causal factors. Deviations of specific levels of a series from levels that correspond to the general trend, explain the effect of factors that appear by chance or cyclically. To propose a hypothesis of a possible type of development, you need to use a graphical method. Visual representation of the series being analyzed allows you to get an idea of the placement of empirical levels on the graph. This helps to better understand the specifics of changes in a time series. But the graphical method cannot give a generalized statistical estimate of the detected trend. The general view of the trend model is as follows:

$$Y_t = (t) + E_t \,, \tag{11.8}$$

where $(t)$is the level conditioned by the trend of development;

E_t is the accidental and cyclical deviation from the trend.

The choice of an adequate function is carried out by the method of least squares – the minimum of the deviations of the sum of squares between the theoretical and empirical levels of a series:

$$\sum(y_{t_i} - y_i)^2 \rightarrow \min \,, \tag{11.9}$$

where y_{t_i}, y_i are the theoretical and empirical levels of the series respectively.

The most important problem of analytical equalization is the selection of a mathematical function which calculates the theoretical levels of the trend. The correctness of the solution of this problem depends on the conclusions and forecasts about the trend patterns of the phenomenon under study.

The shape of the curve can be selected by analyzing the graphical representation of the levels of a time series. For this purpose it is advisable to use a graphical representation of smoothed levels in which random fluctuations are cancelled. The Fisher test (F) is used to estimate the proximity of the trend equation to the empirical time series. The actual (calculated) level of the F-criterion is compared with the theoretical (tabular) value:

$$F_{fact} = \frac{\eta_T^2}{1-\eta_T^2} \times \frac{n-m}{m-1} = \frac{V_1}{V_2}, \qquad (11.10)$$

where m is the number of parameters;

η_T^2 is the theoretical coefficient of determination.

$$\eta_T^2 = 1 - \frac{\sigma_{y-\hat{y}}^2}{\sigma_y^2}, \qquad (11.11)$$

where $\sigma_{y-\hat{y}}$ is the residual dispersion;

σ_y^2 is the total dispersion.

The residual dispersion is calculated by the formula:

$$\sigma_{y-\hat{y}}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}; \qquad (11.12)$$

the total dispersion:

$$\sigma_y^2 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n}. \qquad (11.13)$$

In order for the model to be reliable, the following conditions are required: $F_{fact} > F_{critical}$.

$F_{critical}$ is found using special Fisher distribution tables at $v_1 = m - 1$, $v_2 = n - m$ and the significance level of α.

For time series with small lengths and significant fluctuations the use of the analytic method of equalization with the time function is not recommended, because the approximation does not adapt to the conditions that change the formation of levels. With the emergence of new data it is necessary to build new models. Adaptive modeling and forecasting methods are used to smooth these types of time series. The exponential smoothing model is based on these methods. The time series is smoothed out by the weighted moving average in which the weights are distributed according to the exponential law.

11.2. Interpolation and extrapolation

Analysis of the dynamics of socio-economic phenomena, identification and characterization of the main trend of development give the basis for forecasting – determining the future size of the level of an economic phenomenon.

The forecasting process assumes that the pattern of development that has operated up to now (within a time series) will continue in the foreseeable future, that is, the forecast is based on extrapolation. Extrapolation to the future is called *prospective*, while extrapolation to the past is called *retrospective*.

Usually, when talking about extrapolation of time series, they often mean prospective extrapolation. Initial projections tend to be extrapolated to a trend. Different methods may be used, depending on the input information. The following elementary *extrapolation* methods can be mentioned: based on the average absolute gain, average growth rate, and *extrapolation* based on the application of the least squares method and the presentation of the evolution of a phenomenon in time in the form of a trend equation, that is, the mathematical function of the levels of a series level (y) on the time factor (t) [9; 26; 33; 34].

Prediction based on the average absolute gain can be fulfilled if there is a certainty that the general trend is linear, that is, the method is based on the assumption of a uniform change in the level (the uniformity means the stability of the absolute gain).

In this case, to get the forecast for i steps forward (i is the forecast period), it is enough to use the following formula:

$$\hat{y}_{n+1} = y_n + i \times \overline{\Delta}, \qquad (11.14)$$

where y_n is the actual value at the last n-th point of the series (the end level of the series);

$\hat{y}_{n+1}$ is a predictive estimate of the value (n + 1) of the series level;

$\overline{\Delta}$ is the value of the average absolute gain, calculated for a time series $y_1, y_2, y_3, ..., y_n$.

Forecasting based on the average growth rate can be made when there is reason to believe that the general trend of the series is characterized by an exponential curve. To find the predictive value for i steps forward, you must use the following formula:

$$\hat{y}_{n+1} = y_n \times \overline{K_p^i}, \tag{11.15}$$

where $\overline{K_p^i}$ is the average growth factor calculated for the series $y_1, y_2, y_3, ..., y_n$.

Prediction based on the analytical alignment is the most common prediction method. An analytical trend equation is used to obtain the forecast. To do this, it is sufficient to extend the conditional time value in the model. Formally, an extrapolation can be represented as definition of the function [34]:

$$Y_{t+v} = (Y_t^0, v), \tag{11.16}$$

where Y_{t+v} is the predictive value for the forecast period v;

Y_t^0 is the base of extrapolation, most often the last level of the series, determined by the trend equation.

Using the extrapolation method, two types of prediction are obtained: point and interval. Point forecast is a specific numerical value of the level in the forecast period (moment) of time. Interval prediction is a range of numerical values that approximately contains the predicted values of the levels [34].

Interval forecasts have significant advantages over point forecasts – they take into account the probability of forecasting. It is obvious that a point forecast is of low probability since the trend is characterized by fluctuations in levels and hence errors of parameters. The source of these errors is a limited number of observations, each containing a random component. The random component is observed outside the time series range as well, so it must be taken into account. To do this, they determine a confidence interval that with certain probability would set the limits of possible values Y_{t+v}. Point forecast is converted to interval forecast – the interval width depends on the variation of the levels of the time series around the trend and the probability of conclusion (1 − α):

$$Y_{t+v} = \pm t_{\alpha} \delta_{e}, \tag{11.17}$$

where t_{α} is the Student's confidence coefficient of distribution;

$$\delta_{e} = \sqrt{\frac{\sum(y_{i} - y_{t_{i}})^{2}}{n - m}}, \tag{11.18}$$

where δ_{e} is the residual root mean square deviation of the trend, adjusted to the number of degrees of freedom (n − m);

n is the number of levels of the basic time series;

m is the number of parameters of an adequate trend model.

In the practice of statistical trend analysis the following types of development of socio-economic phenomena in time are distinguished:

1) *uniform development*. This type of dynamics is characterized by constant absolute increments:

$$\Delta_{y} \cong \text{const}. \tag{11.19}$$

The main trend in this type of series is described by the equation of the linear function:

$$\hat{y}_{t} = a_{0} + a_{1}t, \tag{11.20}$$

where $\hat{y}_{t}$ is the theoretical value of the time series;

a_{0} and a_{1} are the equation parameters;

t is time (the sequential number of a period or a moment in time).

For linear dependence, the parameter a_{0} is usually not interpreted, but is sometimes regarded as a generalized initial level of the series; a_{1} is the tightness of relationship, that is, a parameter that shows how much the result will change as time changes by one. Thus, a_{1} can be presented as a constant theoretical absolute increment. The parameter a_{1} is the regression coefficient that determines the direction of development. If a_1 > 0, the levels of the time series increase uniformly, and with a_1 < 0 their uniform decrease is observed;

2) *equally accelerated (decelerated) development*. This type of dynamics is characterized by a constant development increase (slowdown) in time. The levels of such time series change with constant growth rates (Gr) [34]:

$$Gr_{const} \cong const. \tag{11.21}$$

The main trend in this kind of series is described by the 2nd order parabola:

$$\hat{y}_t = a_0 + a_1 t + a_2 t^2. \tag{11.22}$$

The parameters a_1, a_0 are identical to the parameters of the linear regression equation. The parameter a_2 characterizes a constant change in the intensity of development per unit of time. If a_2 > 0, there is an acceleration of development, and if $a_2 < 0$, the growth slows down;

3) *development with variable acceleration (deceleration)*. This kind of dynamics are described using the 3rd order parabola:

$$\hat{y}_t = a_0 + a_1 t + a_2 t^2 + a_3 t^3. \tag{11.23}$$

The parameter a_3 shows the change in acceleration. If $a_3 > 0$, acceleration increases, and if $a_3 < 0$, acceleration slows down.

4) *exponential development*. This type of dynamics is characterized by stable growth rates:

$$Gr_{ch} \cong const, \tag{11.24}$$

where Gr_{ch} is the growth rate (chain).

The main trend is described by the function:

$$\hat{y}_t = a_0 a_1^t, \tag{11.25}$$

where a_1 is the rate of growth (decrease) of the phenomenon under study per unit of time, that is, the intensity of development.

5) *development with the growth slowdown at the end of the period*. For this type of dynamics, the magnitude of the chain absolute increments is reduced in the final levels of a time series:

$$\Delta_{y_{ch}} \to 0. \tag{11.26}$$

The trend of development in such time series is described by means of a semilogarithmic function:

$$\hat{y}_t = a_0 + a_1 \lg t. \tag{11.27}$$

The following regression models are also most commonly used:

$$\text{hyperbola: } \hat{y}_t = a_0 + \frac{a_1}{t}. \tag{11.28}$$

This model is called the inverse model. It is usually used in cases where an unlimited increase in the explanatory variable (in this case, time t) asymptotically approximates the dependent variable Y to a certain limit:

$$\text{step function: } \hat{y}_t = a_0 \times t_1^{a}. \tag{11.29}$$

For analytical alignment, other mathematical functions can also be applied in time series. As noted above, when studying socio-economic phenomena one has to operate with a complex mechanism of interaction of the factors that shape the trend. Therefore, based on qualitative analysis, it is not always possible to draw reliable conclusions about the type of development in the form of an adequate mathematical function. At best, a working hypothesis about possible types of development can be put forward. But choosing a particular mathematical function on this basis is quite complicated. This is especially true for nonlinear functions whose theory is not well developed.

The practice of statistical study of a trend using modern computing facilities provides ample opportunity to find the most appropriate trend model. The performance of modern technologies with high memory capacity allows you to obtain all the necessary indicators for trend analysis, including those that are used to select an adequate mathematical function.

One of the indicators of the adequacy of the mathematical function used in the practice of statistical study of a trend is the *standardized error of approximation* [34]:

$$\sigma_{y_t} = \sqrt{\frac{\Sigma(y_{t_i} - y_t)^2}{n}}. \qquad (11.30)$$

The use of this approximation in the study of a trend is based on the fact that the most adequate function is considered to be a function in which the standardized error of approximation is minimal.

Let's consider an example of analytical alignment of a time series based on the example of the gross grain harvest in Ukraine. From the graphical analysis of the dynamics of gross harvest and identification of the trend using a 3-member moving average, it can be seen that the phenomenon is characterized by strong fluctuations and the trend is not clear. Therefore, we will perform an analytical alignment based on several functions and select the most appropriate one (adequate to real data) out of them.

Let's make preliminary analysis of the series with the help of chain increments, growth rates and increment rates (Table 11.4).

Table 11.4

The dynamics of gross grain harvest, absolute increment, growth rate and the increment rate

Years	Gross grain harvest, thou tons (y)	Absolute increment (chain) $\Delta_{y_{ch}}$	Growth rate Gr_{ch}	Increment rate Ir_{ch}, %
1	2	3	4	5
2001	38 674	–	–	–
2002	38 537	−137	1.00	−0.35
2003	45 623	7 086	1.18	18.39
2004	35 497	−10 126	0.78	−22.19
2005	33 930	−1 567	0.96	−4.41
2006	24 571	−9 359	0.72	−27.58
2007	35 472	10 901	1.44	44.37

Table 11.4 (the end)

1	2	3	4	5
2008	26 471	−9 001	0.75	−25.37
2009	24 581	−1 890	0.93	−7.14
2010	24 459	−122	1.00	−0.50
2011	39 706	15 247	1.62	62.34
2012	38 804	−902	0.98	−2.27
2013	20 234	−18 570	0.52	−47.86
2014	41 809	21 575	2.07	106.63
2015	38 016	−3 793	0.91	−9.07
2016	34 258	−3 758	0.90	−9.89
2017	29 295	−4 963	0.86	−14.49
2018	53 290	23 995	1.82	81.91

To calculate the parameters of the functions on the basis of the least squares method, the system of normal equations is used:

$$\begin{cases} na_0 + a_1\sum t = \sum y \\ a_0\sum t + a_1\sum t^2 = \sum yt \end{cases}. \qquad (11.31)$$

You can simplify the given system, by coding the time so that, $\sum t = 0$.

Then the equation system will look like:

$$\begin{cases} na_0 = \sum y \\ a_1\sum t^2 = \sum yt \end{cases}, \qquad (11.32)$$

where

$$a_0 = \frac{\sum y}{n}, \quad a_1 = \frac{\sum yt}{\sum t^2}. \qquad (11.33)$$

An example of coding for an even number of members of a time series is given in Table 11.5. If the number of levels in a series is odd, the center level of the series is coded as 0.

Table 11.5

Calculation of the trend equation

Years	Gross grain harvest, thousand tons (y)	t_i	t_i^2	$t_i\,y_i$	y_{t_i}
2001	38 674	−17	289	−657 458	33 864
2002	38 537	−15	225	−578 055	33 953
2003	45 623	−13	169	−593 099	34 043
2004	35 497	−11	121	−390 467	34 132
2005	33 930	−9	81	−305 370	34 222
2006	24 571	−7	49	−171 997	34 311
2007	35 472	−5	25	−177 360	34 400
2008	26 471	−3	9	−79 413	34 490
2009	24 581	−1	1	−24 581	34 579
2010	24 459	1	1	24 459	34 668
2011	39 706	3	9	119 118	34 758
2012	38 804	5	25	194 020	34 847
2013	20 234	7	49	141 638	34 937
2014	41 809	9	81	376 281	35 026
2015	38 016	11	121	418 176	35 115
2016	34 258	13	169	445 354	35 205
2017	29 295	15	225	439 425	35 294
2018	53 290	17	289	905 930	35 383
Σ	623 227	0	1 938	86 601	623 227

The parameters a are calculated as follows:

$$a_0 = 623\,227/18 = 34\,623.72;$$
$$a_1 = 86\,601/1938 = 44.686.$$

According to the calculated parameters, the trend model of the function is synthesized:

$$y_t = 34\,623.72 + 44.686\,t\,.$$

Next, the theoretical levels y_t for each year are calculated (a corresponding value of the year code is substituted for t):

$$Y_{2001} = 34\,623.72 + 44.686 \times (-17) = 33\,864;$$

$$Y_{2002} = 34623.72 + 44.686 \times (-15) = 33953;$$

and this way the calculations for each year are made.

$$Y_{2018} = 34623.72 + 44.686 \times 17 = 35383.$$

The correctness of the calculations is checked by the equation of the sums:

$$\sum y_i = \sum y_{t_i} .$$

The parameters of this regression equation can be interpreted as follows: a_0 = 34 623.72 thousand tons – this is the initial level of the gross harvest of cereals over the period considered; parameter a_1 of the trend model shows that the gross grain harvest in Ukraine increased on average by 44.686 thousand tons per year.

For a 2nd order parabola $\overline{y_t} = a_0 + a_1 t + a_2 t^2$, provided the conventional time-setting method is used, if $\sum t = 0$, the parameters of this function are given by the formulas:

$$a_0 = \frac{\sum t^4 \sum y - \sum t^2 \sum t^2 y}{n \sum t^4 - \sum t^2 \sum t^2}, \qquad (11.34)$$

$$a_1 = \frac{\sum yt}{\sum t^2}; \qquad (11.35)$$

$$a_2 = \frac{n \sum t^2 y - \sum t^2 \sum y}{n \sum t^4 - \sum t^2 \sum t^2}. \qquad (11.36)$$

As a result of finding the parameters we get:

$$y = 29\,437.64 + 44.685t + 48.167\, t^2.$$

For a 3rd order parabola $\overline{y_t} = a_0 + a_1 t + a_2 t^2 + a_3 t^3$, if you use the time-setting method, if $\sum t = 0$ the parameters of this function are given by the formulas:

$$a_0 = \frac{\sum t^4 \sum y - \sum t^2 \sum t^2 y}{n\sum t^4 - \sum t^2 \sum t^2}; \tag{11.37}$$

$$a_1 = \frac{\sum t^6 \sum ty - \sum t^4 \sum t^3 y}{\sum t^2 \sum t^6 - \sum t^4 \sum t^4}; \tag{11.38}$$

$$a_2 = \frac{n\sum t^2 y - \sum t^2 \sum y}{n\sum t^4 - \sum t^2 \sum t^2}; \tag{11.39}$$

$$a_3 = \frac{\sum t^2 \sum t^3 y - \sum t^4 \sum t\, y}{\sum t^2 \sum t^6 - \sum t^4 \sum t^4}. \tag{11.40}$$

As a result of obtaining the parameters we have:

$$y = 29437.64 - 60.729t + 48.167t^2 + 0.546t^3.$$

For finding the hyperbola equation $\hat{y}_t = a_0 + \frac{a_1}{t}$, time encoding in which $\sum t = 0$ is ineffective and only possible when the series is even, i.e. does not contain a time value equal to 0. The equation parameters are found from the following system of equations:

$$\begin{cases} \Sigma y = a_0 n + a_1 \Sigma \frac{1}{t} \\ \Sigma y \frac{1}{t} = a_0 \Sigma \frac{1}{t} + a_1 \Sigma \frac{1}{t^2} \end{cases},$$

where

$$a_0 = \frac{\Sigma y \Sigma \frac{1}{t^2} - \Sigma y \frac{1}{t} \Sigma \frac{1}{t}}{n\Sigma \frac{1}{t^2} - \Sigma \frac{1}{t} \Sigma \frac{1}{t}}; \tag{11.41}$$

$$a_1 = \frac{n\Sigma y \frac{1}{t} - \Sigma y \Sigma \frac{1}{t}}{n\Sigma \frac{1}{t^2} - \Sigma \frac{1}{t} \Sigma \frac{1}{t}}. \tag{11.42}$$

So, to find the parameter a_0, a_1 estimates, you need to find the following four sums:

$$\sum y;\ \sum y\frac{1}{t};\ \sum\frac{1}{t};\ \sum\frac{1}{t^2}\ .$$

The table of calculations is given in Table 11.6.

Table 11.6

The data for calculation of the parameters of the hyperbola equation

Years	Grain gross harvest, thousand tons (y)	t	1/t	$1/t^2$	y/t	y_{t_i}
2001	38 674	–17	−0.05882	0.00346	−2 274.94	34 506
2002	38 537	–15	−0.06667	0.00444	−2 569.13	34 491
2003	45 623	–13	−0.07692	0.00592	−3 509.46	34 470
2004	35 497	–11	−0.09091	0.00826	−3 227.00	34 442
2005	33 930	–9	−0.11111	0.01235	−3 770.00	34 402
2006	24 571	–7	−0.14286	0.02041	−3 510.14	34 339
2007	35 472	–5	−0.20000	0.04000	−7 094.40	34 225
2008	26 471	–3	−0.33333	0.11111	−8 823.67	33 959
2009	24 581	–1	−1.00000	1.00000	−24 581.00	32 629
2010	24 459	1	1.00000	1.00000	24 459.00	36 618
2011	39 706	3	0.33333	0.11111	13 235.33	35 289
2012	38 804	5	0.20000	0.04000	7 760.80	35 023
2013	20 234	7	0.14286	0.02041	2 890.57	34 909
2014	41 809	9	0.11111	0.01235	4 645.44	34 845
2015	38 016	11	0.09091	0.00826	3 456.00	34 805
2016	34 258	13	0.07692	0.00592	2 635.23	34 777
2017	29 295	15	0.06667	0.00444	1 953.00	34 757
2018	53 290	17	0.05882	0.00346	3 134.71	34 741
Σ	623 227	0	0.00000	2.41190	4 810.3403	623 227

Let's find the parameters a_0, a_1 and synthesize the trend model:

$$Y_t = 34\,623.7 + 1994.42 / t.$$

After construction of the regression equation, the approximation error (σ) for the trends is calculated (Tables 11.7 – 11.8).

Table 11.7

The data for calculation of the approximation error of linear trend equations and parabola

Years	Grain gross harvest, thousand tons (y)	Theoretical levels of the linear trend (y_{t_i})	The squares of deviations of the actual levels from the theoretical ones $(y_{t_i} - y_i)^2$	Theoretical levels of the 2nd order parabola (y_{t_i})	The squares of deviations of the actual levels from the theoretical ones $(y_{t_i} - y_i)^2$
2001	38 674	33 864	23 136 100	38 674	15 401 838
2002	38 537	33 953	21 013 056	38 537	1 140 925
2003	45 623	34 043	134 096 400	45 623	74 406 018
2004	35 497	34 132	1 863 225	35 497	522 125
2005	33 930	34 222	85 264	33 930	985 906
2006	24 571	34 311	94 867 600	24 571	47 804 356
2007	35 472	34 400	1 149 184	35 472	25 538 768
2008	26 471	34 490	64 304 361	26 471	10 667 375
2009	24 581	34 579	99 960 004	24 581	23 620 794
2010	24 459	34 668	104 223 681	24 459	25 720 055
2011	39 706	34 758	24 482 704	39 706	94 105 339
2012	38 804	34 847	15 657 849	38 804	63 023 467
2013	20 234	34 937	216 178 209	20 234	141 055 291
2014	41 809	35 026	46 009 089	41 809	65 085 917
2015	38 016	35 115	8 415 801	38 016	5 100 801
2016	34 258	35 205	896 809	34 258	15 217 313
2017	29 295	35 294	35 988 001	29 295	135 739 128
2018	53 290	35 383	320 660 649	53 290	84 128 641
Σ	623 227	623 227	1 212 987 986	623 227	829 264 056
δ	x	x	8 209	x	6 788

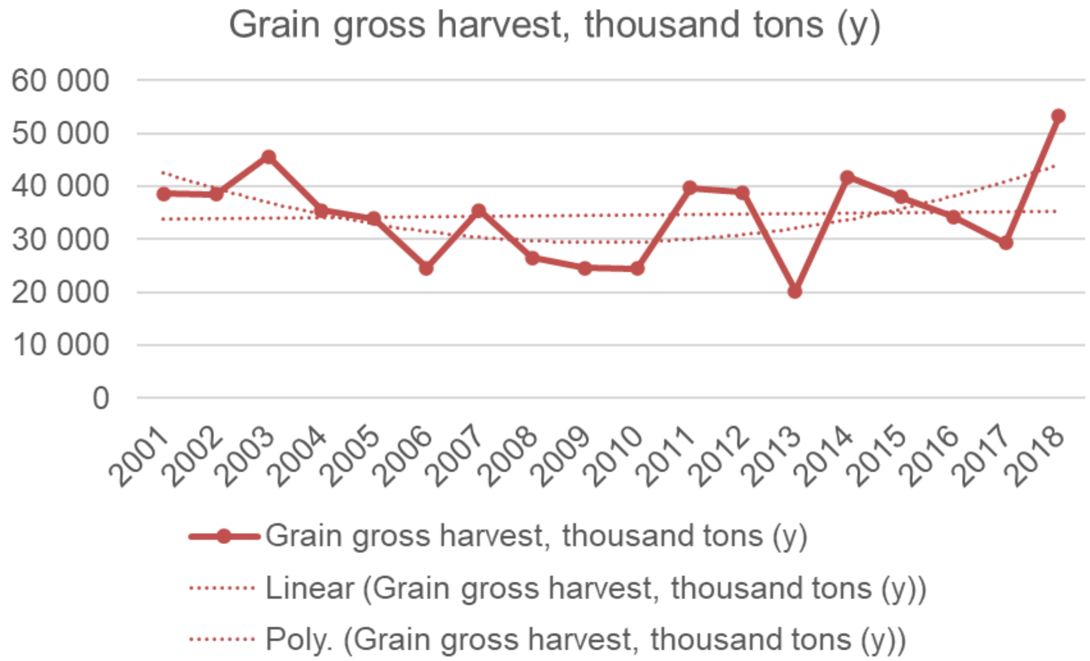
The data for calculating the error of approximation of the 3rd order and hyperbola equations are given in Table 11.8.

Table 11.8

The data for calculation of the approximation error of the 3rd order parabola and hyperbola equations

Years	Grain gross harvest, thousand tons (y)	Theoretical levels of the 3rd order parabola (y_{t_i})	The squares of deviations of the actual levels from the theoretical ones $(y_{t_i} - y_i)^2$	Theoretical levels of the hyperbola (y_{t_i})	The squares of deviations of the actual levels from the theoretical ones $(y_{t_i} - y_i)^2$
2001	38 674	41 707	9 199 905	34 506	17 368 860
2002	38 537	39 343	649 587	34 491	16 372 050
2003	45 623	37 168	71 495 156	34 470	124 382 595
2004	35 497	35 207	84 099	34 442	1 112 157
2005	33 930	33 488	195 689	34 402	222 898
2006	24 571	32 036	55 720 694	34 339	95 410 023
2007	35 472	30 877	21 112 091	34 225	1 555 411
2008	26 471	30 039	12 727 713	33 959	56 068 892
2009	24 581	29 546	24 651 140	32 629	64 775 210
2010	24 459	29 426	24 667 371	36 618	147 844 677
2011	39 706	29 704	100 045 762	35 289	19 514 059
2012	38 804	30 406	70 518 546	35 023	14 298 943
2013	20 234	31 560	128 280 766	34 909	215 345 030
2014	41 809	33 191	74 272 414	34 845	48 492 781
2015	38 016	35 325	7 241 902	34 805	10 310 310
2016	34 258	37 989	13 916 824	34 777	269 505
2017	29 295	41 208	141 916 832	34 757	29 829 985
2018	53 290	45 009	68 571 377	34 741	344 063 884
Σ	623 227	623 227	825 267 868	623 227	1 207 237 269
δ	x	x	6 771	x	8 190

The actual data and data aligned according to a straight line and the 3rd order parabola are graphically presented in Fig. 11.2.



Note: Poly. stands for Polynomial

Fig. 11.2. **The dynamics of grain gross harvesting in Ukraine in 2001 – 2018**

According to the summary data – graphs and standard error of approximation – it turns out that the most appropriate model is a 3rd order parabola.

11.3. Factor analysis of time series

An important place in the study of the dynamics of socio-economic phenomena belongs to factor analysis which is used to study the influence of individual factors on the quantitative and qualitative changes in the phenomenon over time. To perform factor analysis of time series, statistics uses a number of methods and techniques, namely: bringing the time series to one base, comparing several parallel series of dependent and independent variables, enlargement of periods, dividing the studied population into qualitatively homogeneous groups and subgroups, i.e. combining groupings, the use of variance and correlation methods of analysis.

If there is a need to compare the relative rate of change (rate of growth) of different phenomena or indicators, the most common method is *bringing the time series to one base*.

For this purpose, the indicators of time series are expressed in % to the first level of the series (basic growth rates are calculated), and then the coefficients of time lead or lag are calculated.

These coefficients compare the relative rate of time series of the same content across different entities (regions, countries, etc.) or different content across the same entity. For example, over the past three years, the capital-labor ratio in one industry increased by 50 %, in another one it increased by 25 %. The coefficient of time lead of the growth rate of the capital-labor ratio in the first branch compared to the second one is 1.50 : 1.25 = 1.20.

You can compare the dynamics of the capital-stok ratio and productivity in each industry. If the capital-stok ratio has increased by 25 % and labor productivity by 37.5 %, the time lead factor of the labor productivity growth is 1.375 : 1.250 = 1.10.

As to the increment rates, the ratios are only used for the related x and y indicators. This kind of relation is called the *empirical coefficient of elasticity* [34]; it shows the percentage change of y with the change of x by 1 %. For example, the price of product A increased by 2 %, and demand decreased by 4 %. The price elasticity of demand for this product was −4 / +2 = −2, that is, with a price increase of 1 % the demand for the product decreases by 2 %.

Actual levels of time series under the influence of various factors vary, deviating from the main trend of development. In some series fluctuations are systematic, regular in nature, repeated at certain intervals of time, in others they are not of this nature and therefore are called random. Systematic and random fluctuations can be combined in a particular series.

The simplest estimation of the systematic fluctuations are the *coefficients of nonuniformity*, which are calculated by the ratio of the maximum and minimum levels of the time series to the average. The greater the nonuniformy of the process, the greater the difference between the two coefficients [6; 9; 33; 34].

For example, drinking water consumption per day is 7200 m^3, an average per hour is 7200 : 24 = 300 m^3. The highest level of water consumption is in the period from 20:00 to 21:00 amounting to 381 m^3, the lowest consumption is in the period from 2:00 to 3:00 a.m. which makes 165 m^3.

The nonuniformity coefficients are as follows:

K_{max} = 381 : 300 = 1.27;

K_{min} = 165 : 300 = 0.55.

The fluctuation amplitude of 72 points [100 (1.27 – 0.55)] indicates a significant irregularity in water consumption during the day.

11.4. Analysis of seasonal fluctuations

The main tasks of statistical study of fluctuations of socio-economic processes are as follows:

measurement of the fluctuation intensity;

study of the type of fluctuations;

decomposition of complex fluctuations into heterogeneous components;

study of time changes in fluctuations and the dynamics of fluctuations;

studying the variation of fluctuations in a spatial or other population of objects;

study of factors of fluctuations and their statistical and mathematical modeling.

Separate socio-economic processes are characterized by seasonal ups and downs, for example, production and processing of agricultural products, uneven loading of transport, fluctuations in demand for goods, etc. *Seasonal fluctuations* are detected and analyzed on the basis of series of monthly or quarterly data.

The nature of seasonal fluctuations is described by a *"seasonal wave"* formed by seasonality indices. In the time series, which do not show a clear trend of development, *seasonality indices* are the ratio of the actual monthly (quarterly) levels y_t to the monthly (quarterly) average over a year $\bar{y}$, %, determined by the formula [34]:

$$I_s = 100\frac{y_t}{\bar{y}}. \qquad (11.43)$$

Depending on the method of alignment of the original data the following methods of calculation of the seasonality index are used: based on a simple average (the method of a constant average); based on the moving average; analytical alignment (the method of a variable average). Seasonality indices show how many times the actual level of a series at a moment in time or time interval is greater than the average.

The *method of a constant average* is used for series with an unexpressed main development trend (no trend). The index is calculated according to the formula [33, 34]:

$$I_s = \frac{Y_i}{\bar{y}}. \qquad (11.44)$$

The method of a variable average is used for series with a pronounced main trend of development, and in this case the index is calculated by the formula:

$$I_s = \frac{Y_i}{Y_t}. \qquad (11.45)$$

In the analysis of seasonality, the levels of a time series show the evolution of a phenomenon on a monthly (quarterly) basis over one or more years. Seasonal fluctuations may be subject to random deviations. To eliminate the deviations the averaging of individual indices of the same intraannual periods of the analyzed time series is conducted. Therefore, for each period of the annual cycle, generalized indicators are determined as average seasonality indices:

$$\overline{I_s} = \frac{\sum i_s}{n}. \qquad (11.46)$$

A set of average seasonality indices of similar periods make a model of a seasonal wave. If, when constructing a seasonal wave model, random fluctuations are eliminated, the sum of the average seasonality indices of the similar periods = 1200 % if the levels were taken for a month, and 400 % if the levels were quarterly. If this condition is not fulfilled, the model should be adjusted. To do this, they calculate the correction factor:

$$K_{cor} = \frac{1200}{\sum I_{s(average)}} \text{for monthly data;}$$

$$K_{cor} = \frac{400}{\sum I_{s(average)}} \text{ for quarterly data.}$$

All calculated average seasonality indices are adjusted for the value of this coefficient.

Let's consider the procedure for calculating a seasonal wave based on the example of electricity consumption by utilities of a region (Table 11.9).

Table 11.9

Monthly dynamics of electricity consumption

Month	Electricity consumed, million kWh, y_t	Seasonality index I_s, %	$I_s - 100$	$(I_s - 100)^2$
January	172	111.7	11.7	136.89
February	161	104.5	4.5	20.25
March	158	102.6	2.6	6.76
April	151	98.0	−2.0	4.00
May	147	95.5	−4.6	20.25
June	130	84.4	15.6	243.36
July	124	80.5	−19.5	380.25
August	146	94.9	−5.1	26.01
September	149	96.8	−3.2	10.24
October	155	100.6	0.6	0.36
November	168	109.1	9.1	82.81
December	187	121.4	21.4	457.96
Total	1848	100	0	1389.14

Average monthly consumption $\bar{y} = \frac{1848}{12} = 154$ (million kWh). Seasonality indices fluctuate from 121.4 % in December ($\frac{187}{154} \times 100$) to 80.5 % in July $\frac{124}{154} \times 100$. The amplitude of seasonal fluctuations is $R_t = 121.4 - 80.5 = 40.9$. The nature of the seasonal wave is schematically illustrated in Fig. 11.3.

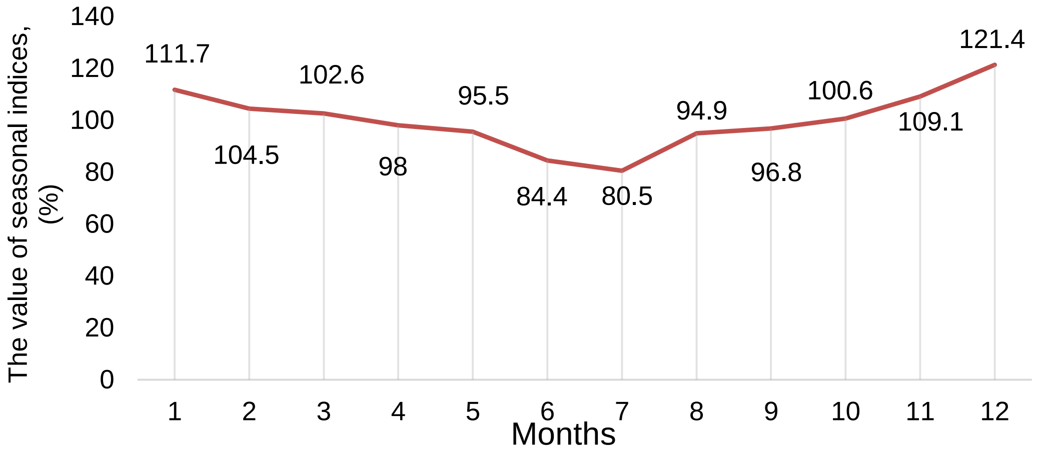


Fig. 11.3. **The seasonal wave of electricity consumption**

To compare the intensity of seasonal fluctuations of different phenomena or the same phenomenon in different years, the following generalized characteristics of the variation of seasonality indices are used:

mean linear deviation $\bar{I}_s = \frac{1}{12}\sum_1^{12}\left| I_s - 100 \right|$;

or standard deviation $\sigma_t = \sqrt{\frac{1}{12}\sum_1^{12}(I_s - 100)^2}$.

In the time series of electricity consumption (see Table 11.8), the mean square deviation is:

$$\sigma_t = \sqrt{\frac{1389.14}{12}} = 10.8 \text{ p. p.}$$

If there is a trend, a preliminary smoothing or smoothing of the time series is performed, the theoretical levels for each month (quarter) of the year are determined, and the seasonality index is calculated as the ratio of the actual levels of the series y_t to the theoretical levels Y_t, i.e. $I_c = 100\frac{y_t}{Y_t}$.

Important concepts

Amplitude (*swing*) is the difference between the largest algebraic trend deviation and the least algebraic deviation.

Analytical alignment is the selection of a theoretical curve for a given time series which expresses the main features of dynamics of the actual phenomenon, that is, describes the empirical data the best.

Aligning the time series by enlarging the intervals is converting the levels of the series into series of longer periods.

Extrapolation is the extension of the patterns of dynamics of development of the investigated phenomenon revealed in the analysis of time series to the future.

The *weighted moving average* is a flowing average calculated in such a way that different weights are given to the levels included in the averaging interval, since the approximation within the smoothing interval is carried out using the levels calculated in the n-th order polynomial.

Smoothing is a method of detecting the main trend of a time series.

Seasonality index is an indicator of seasonal fluctuations that shows how many times the actual level of a series at a time moment or time interval is greater than the average level.

Interval prediction is a range of numerical values that approximately contains the predicted level value.

Nonuniformity ratio is the ratio of the maximum and minimum levels of a time series to the average.

The *coefficient of stability* is a value that is a complement to the coefficient of fluctuations to one.

Time series level fluctuations are deviations of the series levels from a trend.

The method of averaging over the left and right halves is a method in which the time series is divided into two parts, the arithmetic mean is found for each of them and a trend line is drawn on the graph.

The *fluctuation period* is the duration of the cycle.

Forecasting is the assumption that the pattern that acts within a time series, and acts as the basis of forecasting, will remain.

A simple moving average is the arithmetic mean which is calculated with gradual exclusion of the first level of a time series from the given interval of smoothing and inclusion of the next one.

Equally accelerated (decelerated) development is a type of dynamics with constant increment rates.

Uniform development is a type of dynamics with constant absolute increments.

Development with slowing growth at the end of a period is the type of dynamics with which the magnitude of the chain absolute increments decreases at the end levels of a time series.

Exponential development is the type of dynamics with stable growth rates.

Development with variable acceleration (deceleration) is the type of a time series described by the parabola of the 3rd order.

Seasonal fluctuations are repetitive, persistent intraannual fluctuations.

A *trend* is the direction of development, the main long-term trend of a time series development.

Point forecast is a specific numerical value of the level in the forecast period.

Frequency is the number of cycles per unit of time.

Typical tasks

Task 1. We have some data on the profit of an enterprise (Table 11.10). You need to determine the missing level of the time series and draw conclusions.

Table 11.10

Enterprise profit data

Years	2012	2013	2014	2015	2016	2017	2018
Enterprise profit, million UAH	3.75	4.05	–	4.27	4.82	4.99	5.12

The solution.

1. The insufficient level is calculated by the method of averaging of the closest values: $Y_i = \frac{Y_{i-1} + Y_{i+1}}{2}$;

$$Y_{2014} = \frac{4.05 + 4.27}{2} = 4.16 \text{ (UAH million).}$$

2. The insufficient level is calculated by the method of the average absolute increment:

$$Y_i = Y_{i-1} + \overline{\Delta};$$

$$\overline{\Delta} = \frac{Y_n - Y_1}{n-1} = \frac{5.12 - 3.75}{7-1} = 0.23;$$

$$Y_{2014} = 4.05 + 0.23 = 4.28 \text{ (UAH million).}$$

3. The insufficient level is calculated by the method of the average growth rate:

$$Y_i = Y_{i-1} \times \overline{G}_r;$$

$$\overline{G}_r = \sqrt[n\text{-}1]{\frac{Y_n}{Y_1}} = \sqrt[7-1]{\frac{5.12}{3.75}} = 1.053.$$

$$Y_{2014} = 4.05 \times 1.053 = 4.26.$$

Conclusion. The calculation of the insufficient level of the time series by three ways shows that the level of profit of the enterprise in 2014 was in the range from 4.16 mln UAH to 4.28 mln UAH.

Task 2. We have some data on the dynamics of sales of tourist vouchers (Table 11.11).

Table 11.11

The dynamics of sales of tourist vouchers

Quarters	Sold vouchers		
	2016	2017	2018
1st	170	161	159
2nd	175	215	242
3rd	190	191	207
4th	171	186	195

To analyze the dynamics of sales of tourist vouchers during the year, it is necessary to determine the seasonality index and draw conclusions.

The solution.

1. Let's determine the growth rate on a yearly basis (Table 11.12).

Table 11.12

The calculation table

Year	Number of sold vouchers	Growth rate, %	
		chain	base
2016	706	–	–
2017	753	106.66	106.66
2018	803	106.4	113.74

We see that the series has a clear upward trend. To analyze the dynamics of the year in the series with the upward trend, the study of seasonality relies on the moving average method:

$$i_s = \frac{y_I}{Y_i}.$$

For each period of the annual cycle, generalized indicators are determined in the form of average seasonality indices:

$$\overline{I_s} = \frac{\sum\limits_{i=1}^{n} \sum i_{s_i}}{n}.$$

2. Let's define the theoretical values (y_t). To do this, we have to calculate the parameters a_0 and a_1 based on the auxiliary table (Table 11.13).

Table 11.13

The auxiliary table

Periods	Actual value, y_i	t	t^2	$y_i \times t$
2016				
1st	170	−5.5	30.25	−935
2nd	175	−4.5	20.25	−787.5
3rd	190	−3.5	12.25	−665
4th	171	−2.5	6.25	−427.5
2017				
1st	161	−1.5	2.25	−241.5
2nd	215	−0.5	0.25	−107.5
3rd	191	0.5	0.25	95.5
4th	186	1.5	2.25	279
2018				
1st	159	2.5	6.25	397.5
2nd	242	3.5	12.25	847
3rd	207	4.5	20.25	931.5
4th	195	5.5	30.25	1072.5
Total	2262	0	143	459

So, $a_0 = \frac{2262}{12} = 188.5$; $a_1 = \frac{459}{143} = 3.21$.

The straight line equation will look like this: $y_t = 188.5 + 3.21 \times t$.

3. Let's define the aligned values of the series and calculate the seasonality indices (Table 11.14).

Table 11.14

Calculation of the aligned values of the series and the seasonality index

Periods	Actual value			Aligned values			Actual values in % of the aligned ones			The sum of percent ratio	Seasonality index
	2016	2017	2018	2016	2017	2018	2016	2017	2018		
1	170	161	159	170.8	183.7	196.5	99.5	86.2	80.9	266.6	88.9
2	175	215	242	174.1	186.9	199.7	100.5	115	121.2	336.7	112.2
3	190	191	207	177.3	190.1	202.9	107.2	100.4	102	309.6	103.2
4	171	186	195	180.5	193.3	206.2	94.7	96.2	94.6	285.5	95.2

Conclusion. The largest share of sales of tourist vouchers is in the second quarter.

Task 3. It is necessary to build a linear trend model of dependence of oil production in the region based on the time factor.

The solution.

The procedure for calculation of the parameters of a linear function is considered based on the example of the time series of oil production in the region (Table 11.15).

Table 11.15

The dynamics of oil production

Year	y_t, million tons	Δ_t	Time variable, t	$y_t \times t$	Y_t = 74.5 + 3.8 t
2012	63.5	–	−3	−190.5	63.1
2013	66.8	3.3	−2	−133.6	66.9
2014	71.0	4.2	−1	−71.0	70.7
2015	74.3	3.3	0	0	74.5
2016	76.9	2.6	1	76.9	78.3
2017	82.2	5.3	2	164.4	82.1
2018	86.8	4.6	3	260.4	85.9
Total	521.5	×	0	106.6	521.5

The chain absolute increments of the time series are almost stable, so the trend can be described by a linear function. Since the length of the series n = 7, $\Sigma t^2 = 7 (7^2 - 1) : 12 = 28$. The parameters of the trend equation are:

$$a = \Sigma y_t : n = 521.5 : 7 = 74.5;$$
$$b = \Sigma y_t t : \Sigma t^2 = 106.6 : 28 = 3.8.$$

The linear trend looks like this: $Y_t = 74.5 + 3.8 t$, that is, the average level of oil production is 74.5 million tons, the average annual production increment is 3.8 million tons.

The last column of the table gives theoretical levels for each year Y_t, that is, the expected levels of oil production in the t-th year, due to the main factors of the industry's development is $Y_1 = 74.5 + 3.8 (-3) = 63.1$ million tons for 2012, $Y_2 = 74.5 + 3.8 (-2) = 66.9$ million tons for 2013, etc.

The sums of the actual levels $\sum y_t$ and the levels calculated based on a linear trend of theoretical levels $\sum Y_t$ are the same: $\sum y_t = \sum Y_t = 521.5$ million tons.

Continuation of the identified trend beyond the time series is called extrapolation of the trend. This is one of the methods of statistical forecasting, the prerequisite of which is the invariance of the causal complex that forms the trend. The forecast, expected level Y_{t+v} depends on the forecasting base and the time advance period v. So, assuming that the conditions in which the trend of oil production was formed will not change in the near future, let's determine the forecast for 2020. The forecast is based on the theoretical level of 2018, with a time advance period of v = 2. Oil production is expected to reach 93.5 million tons in 2020:

$$Y_{t+v} = 85.9 + 3.8 \times 2 = 93.5.$$

The extrapolation method gives point prediction. In practice, as a rule, the confidence limits of the forecast level $Y_{t+v} \pm ts_p$ are determined, where s_p is a standard prediction error, Student's t-quantile distribution.

Task 4. On the basis of the data on the distribution of vouches on a monthly basis in the region (Table 11.16), it is necessary to determine the indicators that characterize the form of seasonal fluctuations and the indicators of the intensity of fluctuations.

Table 11.16

Monthly distribution of vouchers in the region

Months	1	2	3	4	5	6	7	8	9	10	11	12
Number of vouchers	1994	1870	1860	1980	1995	2010	2030	2500	2400	2350	2120	2030

The solution.

In order to determine the indicators that characterize the form of seasonal fluctuations, it is advisable to build a table of calculations (Table 11.17). The form of seasonal fluctuations is characterized by absolute deviations from the monthly average and seasonal fluctuation indices. The absolute deviations are found by subtracting the monthly average from the monthly levels: 1994 − 2095 = − 101; 1870 − 2095 = − 225, etc.

Table 11.17

The calculation table

Months	Number of vouchers	Deviations from the monthly average	Squares of deviations	Seasonal fluctuation indices, %
1	1994	−101	10 201	95.18
2	1870	−225	50 625	89.26
3	1860	−235	55 225	88.78
4	1980	−115	13 225	94.51
5	1995	−100	10 000	95.23
6	2010	−85	7 225	95.94
7	2030	−65	4 225	96.90
8	2500	405	164 025	119.33
9	2400	305	93 025	114.56
10	2350	255	65 025	112.17
11	2120	25	625	101.19
12	2030	−65	4 225	96.90
Total	25139		10 201	

Seasonal fluctuation indices are calculated as follows:

$$\frac{1994}{2095} \times 100\,\% = 95.2\,\%;\ \frac{1870}{2095} \times 100\,\% = 89.3\,\%.$$

The form of seasonal fluctuations can be presented graphically (Fig. 11.4).

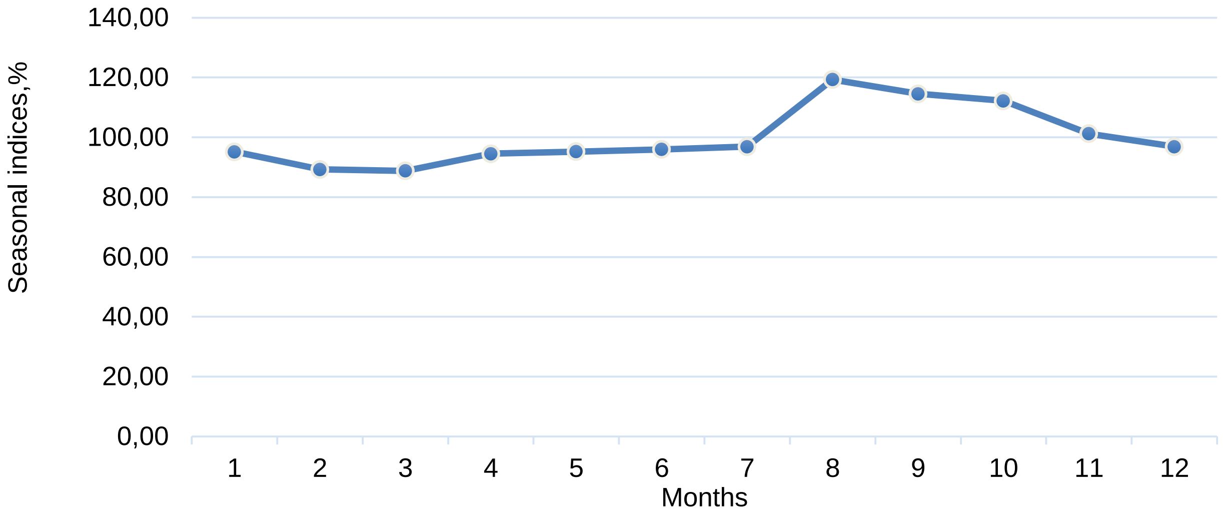


Fig. 11.4. **The seasonal fluctuation indices**

The graph clearly shows that growth of demand for vouchers is observed in the period from July to October while in winter months the demand declines, especially in January and March.

Reference to laboratory work

The guidelines for performing laboratory work on the topic "Analysis of trends of development and fluctuations" are presented in [100]. The laboratory work is aimed at acquiring the skills in extrapolation and interpolation in a time series, construction and analysis of models using MS Excel. The laboratory task is to find the levels of a time series by extrapolation using MS Excel; to build a forecast and find out within what range the forecast value will be; to build different dynamics models based on the time series data.

Questions for self-assessment

1. What are the techniques for aligning time series?

2. What is the essence of adjusting the time series to one base and closing the time series?

3. What is the essence of the method of enlargement of periods?

4. Describe the techniques for aligning the time series by the method of a moving average, the average absolute increment and the average growth rate.

5. What is the interpolation and extrapolation of time series, their meaning and application?

6. How is the forecast for the future made using the trend equation?

7. How are seasonal fluctuations measured in a time series?

8. How do you understand a trend of development? Give examples of trends.

9. What is the difference between smoothing and alignment of a time series? What methods are used in each case?

10. Describe the methods of interpolation of a time series.

Questions for critical rethinking (essays)

1. The features of the study of the dynamics of socio-economic phenomena.

2. Measurement of seasonal fluctuations in the time series of consumption of material goods by the population.

3. The use of trend models for prediction of educational trends in Ukraine and Europe.

4. Measurement of seasonal fluctuations in the field of agriculture.

5. Seasonality as one of the determining factors of tourist demand.

12. The index method

Basic questions:

12.1. Indices in statistical and economic analysis.

12.2. The aggregate index as the main form of the general index. Weighted average indices.

12.3. Indices with variable and constant weights.

12.4. Average indices.

12.5. Territorial indices.

12.6. The index factor analysis method.

12.1. Indices in statistical and economic analysis

In modern analytical practice, indices are regarded as a category that shows the relative change of complex phenomena, with individual elements which are directly incomparable. Therefore, the index, as a special relative value, is determined by the methodology of construction and the nature of the functions.

In the theory of research the *synthetic* and *analytical* functions of indices are considered.

According to the *synthetic function*, the peculiarity of the indices is the aggregation of heterogeneous units of a statistical population.

Analytical properties of indices allow a researcher to trace the influence of individual factors on the change of the phenomenon under consideration.

The complexity of representation of changes in socio-economic phenomena or processes implies a combination of these functions.

The fundamental difference of indices from relative and average values is a systematic approach to the study of phenomena based on an objective relationship between them.

In a systematic approach, relative values inevitably turn into indicators that characterize the dynamics of change in a unit of a phenomenon (individual indices). Individual indices are used to compute complex indices that show the dynamics of two or more interrelated phenomena. Considering a relative value as an individual index, a statistician only emphasizes that this value is intended for calculation of a complex index.

Indices do not necessarily determine the dynamics of the processes under study. Often, they are used to compare complex phenomena in statics (territorial indices) or to evaluate deviations in the levels of interrelated phenomena from a benchmark or an optimal variant.

The main point in the theory of index computation is the transition from the analysis of quantitative differences between the elements of the compared systems to the analysis of quantitative differences between the systems as a whole.

The ways of building an index depend on the content of the studied indicator, the methodology of calculation of initial statistical indicators, the availability of statistics and the purpose of the study. Indices are calculated based on the highest degree of statistical generalization, which results in summary and data processing [31; 40].

Thus, an index is a relative value that, based on the systematic nature of a phenomenon, characterizes its change in time and space, or the degree of deviation from a certain standard (norm).

The name of the index shows its meaningfulness, and the numerical value indicates the intensity of change or the degree of deviation of the phenomenon. For example, the consumer price index in the reporting year was 109.5 %.

The scorecard used to compute the indices has the following *conventional signs:*

q – the volume of production in natural units (physical volume of sales);

p – the price per unit;

Q or *(qp)* – the cost of production (turnover);

z – the unit cost;

zq – the total cost of production;

w – labor productivity;

f – average wage;

T – the number of employees (or hours worked);

t – the time spent per unit of output.

Each index includes data for two periods: *reporting* $(_1)$ – current (compared) and *base* $(_0)$ – used as a benchmark.

With *indices of dynamics* the base of comparison is an index of a certain previous period (moment) of time, with *territorial indices* it is an index of a certain region (object).

An index as a relative indicator can take the form of a ratio (when the base level is taken as one) or a percentage (when the base level is taken as 100). If the index is greater than 1 (or 100 %), the level of the investigated phenomenon increases, and if it is less than 1 (or 100 %), the level of the phenomenon decreases.

Any index consists of the following *elements*: the indexed value, the type (form), the weight of indices, the terms of calculation.

Depending on:

the indexed values the indices of the physical volume of products, price indices, labor productivity indices, etc. are built;

the form indices may be aggregate and average; depending on the type of the mean there are arithmetic mean, harmonic mean, geometric mean indices, etc.;

the choice of the index weight there may be simple (nonweighted) and weighted indices; weighted indices include constant and variable weight indices (revised as needed over time);

the calculation term, base indices (with a constant base of comparison) and chain indexes (with a variable base of comparison) are considered.

The techniques for construction and calculation of indices for comparison in time and space are the same.

All economic indices can be classified based on the following *characteristics*: the base for comparison, the degree of coverage of the elements of a population, the form of construction, the nature of the object of study, the type of statistical weights (comparisons), the composition of a phenomenon, the term of calculation.

Depending on *the base for comparison* indices are divided into:

dynamic, showing the change in the phenomenon taking into account the time factor (change in the value of the consumer basket of the population of a country in September compared to January of the current year);

territorial that show the result of comparisons in space (according to different entities, regions, countries). For example, the value of the consumer basket in Ukraine compared to the corresponding indicator in Bulgaria;

intergroups that characterize the deviation of a phenomenon from a certain reference level (the standard value of the indicator, the minimum or maximum value is accepted as the base for comparison).

The benchmark is crucial for justification of the index. The choice of the comparison base is determined by the purpose of the study.

According to the degree of coverage of the elements of a population (the level of aggregation of information) indices are divided into:

individual, characterizing the change in time or the ratio in the space of one individual phenomenon (change in the price of a product of a certain kind, change in the volume of sales of a product of a certain kind);

consolidated as being the ratio of levels of a complex phenomenon, not directly measurable between each other (change in the number of goods sold across different product groups).

Consolidated indices, in turn, are subdivided into:

group (subindices), if a relative change in a certain group constituting the population is studied (a change in the price of a group of food or nonfood products);

total, if for the given conditions a consolidated index is determined (a change in the price of a group of food and nonfood products).

A prerequisite for determining *an individual index* is the maximum homogeneity of the phenomenon (object) for which the index is calculated.

The *main element* of the index relation is the *indexable value* – the value of the characteristic of a statistical population, the change of which is the object of study [31; 34; 40].

Individual indices are denoted by the symbol (i). For example: the individual index of *the physical volume* of sales of goods (production of goods) (i_q):

$$i_q = \frac{q_1}{q_0}, \qquad (12.1)$$

where q_1 and q_0 are the volume of sales of a particular type of product in the current and base periods in natural units of measurement.

This index shows: how many times the volume of sales (output) of a unit of goods increased (decreased) in the reporting period compared to the baseline; the percentage of the increase (decrease) in the volume of sales of a certain product (output);

the individual *price index* (i_p) is calculated as:

$$i_p = \frac{p_1}{p_0}. \qquad (12.2)$$

This index describes the change in the price of a unit of a particular product in the current period (p_1) compared to the base period (p_0).

Consolidated indices are denoted by the symbol (I). The method for construction of consolidated indices is more complicated than for the individual ones.

The problems of computing consolidated indices relate to:

comparison of noncomparable numerator and denominator;

the choice of a commensurator (an indicator that allows you to reduce a set of heterogeneous elements to a comparable form) and weight (an indicator that determines the importance of the commensurator in a certain population);

the order of indexing and fixing of the commensurator and weight.

If comparability is ensured by means of comparable aggregates (the product of conjugate values – the indexed value and its weight), the order of indexing and fixing of the weights is determined by the appropriate indexing system: base weighted (Laspeyres) or current weighted (Paasche) [31; 34].

Group indices represent patterns of development of individual parts of a phenomenon. In such indices their relationship is shown by the method of groupings.

Depending on *the form of computation* indices are divided into:

aggregate, which is the main form of total indices;

weighted average, which, depending on the form of the average, is divided into:

arithmetic;

harmonic.

The choice of the index calculation f*orm* depends on:

the goals of the research;

the economic nature of the indicator being analyzed;

the available information.

According to the nature of the research object, total indices are subdivided into the indices of:

quantitative indicators containing the characteristic of changes in the volume of a certain phenomenon which is expressed in the respective units of measurement (the index of sales volume of the US dollar on the currency exchange);

qualitative indicators containing a characteristic of change in a qualitative feature which represents the peculiarities of the development of a phenomenon (the Euro exchange rate index).

Such distribution is based on the type of the indexed value.

Depending on the type of weight indices are:

simple (unweighted);

balanced, which, in turn, are subdivided into:

indices with *constant* (unchanged) weights;

indices with *variable weights* (which can be viewed for some time).

The most difficult part of building an index is choosing the index weight and calculating more accurately the weight of each group, and sometimes each unit that is included in the indexed set. The system of such weights should represent the model of the structure of the socio-economic phenomenon, whose dynamics is numerically expressed in the index [31; 40]. For example, the price index weights should present the commodity structure of the trading turnover.

Depending on *the composition of a phenomenon* indices are of:

variable composition;

permanent (fixed) composition;

structural shifts.

As to *the term of calculation* indices are divided into:

basic (with a constant time base);

chain where the numerical values of the indexed value in each current period are compared with their values in the previous period (indices with a variable base).

12.2. The aggregate index as the main form of the total index. Weighted average indices

The economic content of the index determines the method of calculation.

The method of constructing an aggregate index provides a solution to the following problems:

what value will be indexed;

on what composition of heterogeneous elements of the phenomenon it is necessary to calculate the index;

what will serve as a weight for calculation of the index.

Depending on the nature of the research object, volume and quality indicators are determined.

The group of volume indicators includes indices of the physical volume of products, national income, retail turnover, consumption and more. Volume calculation is taken as the basis for calculation of these indices.

The *group of qualitative indicators* includes indices of the cost of production, labor productivity, prices, etc. Qualitative indicators are used as a basis for calculation of these indices.

In order to ensure comparison of the components of indexed values, a measurement indicator – weight – is introduced into the calculation, which allows you to ensure quantitative comparability for the reporting and base periods.

When choosing the weight of the index it is necessary to be guided by the following *rule:* if a quantitative index is constructed, the weight is taken over the base period; to build a quality score index, weight is determined over the reporting period.

The most typical index of quantitative indicators is the volume index (physical volume).

The aggregative *physical volume index* has the following form:

$$I_q = \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0}, \qquad (12.3)$$

where $\Sigma q_1 p_0$ is the product cost in the reporting period at basic prices;

$\Sigma q_0 p_0$ is the product cost in the base period.

This index shows how many times the cost of products has increased (decreased) provided the growth (decrease) of the production volume or the percentage of the increase (decrease) in the product cost due to changes in the physical volume of its production.

If the object of study is an individual enterprise, the index is determined based on the totality of the goods produced. When the object of study is an industry, the index is calculated based on the totality of all goods produced in the industry, or their individual groups, depending on the purpose of the analysis. If the object of study is any region, the index is calculated based on the goods produced at the enterprises of that region.

In a market economy, a special place among quality indices is given to the price index.

The main content of *the price index* is the estimation of the dynamics of prices for goods of industrial and nonproductive consumption. In addition, the price index plays a role as a general measure of inflation in macroeconomic research. It is used for: adjustment of the statutory minimum wage; formation of tax rates; development of feasibility studies and projects; recalculation of the main indicators of the SNA from actual prices into comparable ones.

The aggregate formula of the total *price index* has the form:

$$I_p = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}, \tag{12.4}$$

where $\Sigma p_1 q_1$ is the cost of products in the reporting (current) period.

This index shows: how many times the cost of products (or turnover) increased (decreased) due to changes in prices; the percentage of the increase (decrease) in the cost of products (turnover) as a result of price changes.

Depending on the model of market economy development, different index systems are used [31; 40].

The basic index systems are as follows:

H. Paasche $$I_q = \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1};\ I_p = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1};$$

E. Laspeyres $$I_q = \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0};\ I_p = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0}.$$

The values of the price indices in the Paasche and Laspeyres systems do not coincide because they have different economic content.

The Paasche Price Index shows how much more expensive (cheaper) goods were in the reporting period than in the reference period.

The Laspeyres Price Index shows how many times the price of commodities in the base period went up (decreased) due to changes in their prices in the reporting period.

The Paasche index describes the impact of changes in prices on the value of goods sold in the reporting period. The Laspeyres index shows the impact of changes in prices on the value of goods sold in the base period.

The use of the Paasche and Laspeyres indices depends on the purpose of the study. If the analysis aims to determine the economic effect of changes in prices in the reporting period compared with the baseline, the Paasche index is used. It shows the difference between the actual cost of selling the goods in the reporting period ($\Sigma q_1 p_1$) and the estimated cost of selling the same goods at basic prices ($\Sigma q_0 p_1$)). If the purpose of the analysis is to determine the volume of sales in the future period of the same amount of goods as in the base period, but with new prices, the Laspeyres index is used.

Multiplying the prices by the corresponding amount of goods sold and summing up the products, we get an indicator of the total turnover. Its relative change in dynamics is characterized by *a consolidated index of value (turnover)* in actual prices:

$$I_{pq} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}. \qquad (12.5)$$

In the absence of a specific information base, which makes it impossible to carry out index analysis in aggregate form, indices can be constructed in the form of individual averages.

The weighted average index is the average of individual indices weighted based on the volumes that have the same dimension and are fixed at a constant level.

In the case of such index definition, a question arises as to the choice of the form of the mean and the choice of the system of weights.

The result obtained by constructing the averages of the individual indices should always coincide with the result of the aggregate index.

The average index is identical to the aggregate index.

Identity rules:

1) *the arithmetic mean of an index* will be identical to the aggregate index and give the same result when we take the products of the denominator of the aggregate index for the weights of the individual index.

The arithmetic mean of the physical volume is calculated as:

$$I_q = \frac{\Sigma i_q p_0 q_0}{\Sigma p_0 q_0}; \tag{12.6}$$

2) *the average harmonic index* will be identical to the aggregate index and give the same result when we take the products of the numerator of the aggregate index for the weights of the individual index.

The average harmonic price index is calculated as:

$$I_p = \frac{\Sigma p_1 q_1}{\Sigma \frac{p_1}{i_p} \times q_1}. \tag{12.7}$$

Computing average indices with other weights is not economically viable, though they can be formally determined.

That is, if instead of the indexed value in the numerator of the aggregate index its value from the formula of the corresponding individual index is substituted, the aggregate index is converted to the *arithmetic mean* (when quantitative indicators are dealt with), and if such a change is made in the denominator, we get the *average harmonic index* (in case with qualitative indicators).

Building indices is a process of finding the overall measure, the mean value of the change in a phenomenon, and the index method itself is a further development of the mean method.

In general, the classification of indices can be summarized as follows:

average harmonic indices with current weights correspond to aggregate indices with current weights;

mean arithmetic indices with basis weights correspond to aggregate indices with basis weights.

The concept of weight varies depending on the form in which the index is constructed: aggregate or average.

12.3. Indices with variable and constant weights

All indices considered measure the change in economic phenomena by comparing the data over two time periods. If there is a need to study the development of socio-economic phenomena over a period, they use a system of indices that consistently characterizes changes that occur over the selected time interval. The index system includes several components with their number being equal to the number of time periods included in the analysis minus one.

There are two options for building an index system [31; 34; 40]:

the indicators (levels) of each period included in the considered interval are compared with the level of one period selected for comparison. Indices included in such a system are called *base* indices;

the indicators (levels) are compared consecutively, each level is compared with the previous one. Such indices are called *chain indices*.

The system of chain and base indices may consist of either individual or general indices.

In statistical practice, the choice of an index system is determined by the nature of the tasks being solved. Base indices give a clearer picture of the general trend of change in the phenomenon under consideration, and chain indices give a more detailed picture of the consistent change in levels over time.

There is some connection between the *chain* and *base* indices:

for individual indices, the product of the chain indicators (growth factors with a variable base) will always be equal to the base index (the growth factor with a constant base);

the ratio of the two base individual indices (fixed-base growth ratios) gives a chain index (variable-base growth ratios).

Example 12.1. Let's build base and chain indices:

1) *for the physical volume individual index*:

a) *base indices* are: $i_q = \frac{q_1}{q_0}; \frac{q_2}{q_0}; \frac{q_3}{q_0}; \frac{q_4}{q_0} \ldots;$

b) *chain indices* are: $i_q = \frac{q_1}{q_0}; \frac{q_2}{q_1}; \frac{q_3}{q_2}; \frac{q_4}{q_3} \ldots.$

In the course of constructing a series of consolidated (total) indices, a specific question arises – the choice of the weights of these indices.

Indices can be calculated not only with *constant a*nd *variable* bases of comparison, but also with *constant* and *variable weights*.

2) *for the physical volume total index*:

a) *chain:*

with variable weight:

$$I_q = \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0};\ \frac{\Sigma q_2 p_1}{\Sigma q_1 p_1};\frac{\Sigma q_3 p_2}{\Sigma q_2 p_2};\ \frac{\Sigma q_4 p_3}{\Sigma q_3 p_3}; \qquad (12.8)$$

with constant weight:

$$I_q = \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0};\ \frac{\Sigma q_2 p_0}{\Sigma q_1 p_0};\ \frac{\Sigma q_2 p_0}{\Sigma q_1 p_0};\ \frac{\Sigma q_2 p_0}{\Sigma q_1 p_0}; \qquad (12.9)$$

b) *base:*

with variable weight:

$$I_q = \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0};\ \frac{\Sigma q_2 p_1}{\Sigma q_0 p_1};\ \frac{\Sigma q_3 p_2}{\Sigma q_0 p_2};\ \frac{\Sigma q_4 p_3}{\Sigma q_0 p_3}; \qquad (12.10)$$

with constant weight:

$$I_q = \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0};\ \frac{\Sigma q_2 p_0}{\Sigma q_0 p_0};\ \frac{\Sigma q_3 p_0}{\Sigma q_0 p_0};\ \frac{\Sigma q_4 p_0}{\Sigma q_0 p_0}. \qquad (12.11)$$

For general indices, the considered dependence will only occur if a number of total indices are calculated with the same weights.

A system of constant weight indices is usually built for quantitative indicators (volume indices); the system of indices with variable weights – for qualitative indicators (price indices, cost, time spent per unit of production).

Indices with constant weights are so-called *standardized indicators* that are calculated for the same "standard" structure of a set of phenomena. They allow you to eliminate the influence of the structure change on the indexed value dynamics.

12.4. Average indices

The dynamics of indicators with the levels expressed by average values is affected by the change in the structure of the phenomenon under consideration. The change in the structure of the phenomenon means the change in the proportion of individual units of the population from which the averages are formed, in their total number. For example, the average cost of a product in an industry is not only influenced by the change in the product cost at the industry enterprises, but also by the change in the share of enterprises with different cost in the total output of the product. The dynamics of average per capita income depends on the change in the income of each person and the change in the number of people with higher (or lower) incomes in the total population.

Thus, the change in the average value of an indicator is influenced by two factors at the same time: the change in the indexed values and the change in the structure of the phenomenon. Finding the quantitative impact of these factors is based on the construction of indices of the average level of the indicator [31; 40].

Analysis of the dynamics of the average level of indicators is carried out on the basis of building a system of interdependent indices. The index that characterizes the change in the average level of the indicator due to the change of all factors as a whole is equal to the product of the indices-multipliers, each of which characterizes the change of only one factor and thus the impact of this change on the dynamics of the average value of the indicator.

The variable composition index is calculated as the ratio of the average levels of the index for the current and base periods.

The general scheme of determining the index of variable composition is:

$$I_{v.c\bar{x}} = \frac{\bar{x}_1}{\bar{x}_0} = \frac{\Sigma x_1 f_1}{\Sigma f_1} : \frac{\Sigma x_0 f_0}{\Sigma f_0}, \qquad (12.12)$$

where x_1, x_0 are the levels of the indicator, which is indexed in the reporting and baseline periods respectively;

f_1, f_0 are the weights (frequencies) of the indicator, which is indexed in the reporting and baseline periods respectively.

This index shows the impact of two factors simultaneously on the overall dynamics of the average.

The change in the average level of the indicator due to the change of the indexed (quality factor) is determined by *the index of constant (fixed) composition:*

$$I_{c.c..\,\bar{x}} = \frac{\Sigma x_1 f_1}{\Sigma f_1} : \frac{\Sigma x_0 f_1}{\Sigma f_1}. \tag{12.13}$$

The change in the mean level due to the frequency change (quantitative factor) is determined by the *structural shift index:*

$$I_{s.s.\,\bar{x}} = \frac{\Sigma x_0 f_1}{\Sigma f_1} : \frac{\Sigma x_0 f_0}{\Sigma f_0}. \tag{12.14}$$

As the weights (frequencies) of indices of averages, along with absolute indices f, relative indices (shares) d can be used. Then the dynamics indices of the averages will be of the form:

$$I_{\bar{x}} = \frac{\Sigma x_1 d_1}{\Sigma x_0 d_0}; \tag{12.15}$$

$$I_{\bar{x}} = \frac{\Sigma x_1 d_1}{\Sigma x_0 d_1}; \tag{12.16}$$

$$I_{\bar{x}} = \frac{\Sigma d_1 x_0}{\Sigma d_0 x_0}, \tag{12.17}$$

where d_1, d_0 are the shares of units with a defined attribute value in the population in the reporting and baseline periods, respectively, ($\Sigma d = 1$)).

There is a correlation between the mean indices [31; 34]:

$$I_{v.c.} = I_{c.c.} \times I_{s.s.}. \tag{12.18}$$

Thus, determining a particular average level of an indicator, you can get a characteristic of its dynamics. The first step in obtaining the correct characteristics of the dynamics of a certain average is to understand its economic content.

So, the average of:

the wages of the company's employees are determined by the ratio of the wage bill for a particular research period in the company to the average number of employees in that company during the survey period;

the price of sales of a certain type of goods in the city is determined by the ratio of the total proceeds from the sale of this type to the amount of goods sold for the whole period of the study;

the productivity of employees of a manufacturing firm is determined by the ratio of output produced by all employees of the firm to their number.

A variety of mean indices are territorial indices, where average levels are compared across individual territories or objects.

12.5. Territorial indices

Statistics is a widely used method of comparing indicators across enterprises, cities, economic districts, oblasts, and countries. Summarizing indicators that characterize the correlation of the levels of complex economic phenomena in space (i.e. in terms of territories and objects) are called *territorial indices.*

The construction of territorial indices has its own peculiarities compared to the indices that characterize the dynamics of phenomena. The calculation of individual territorial indices is easy enough, since the usual relative values of comparison are calculated. Thus, if the average price for the same type of products is compared in two regions with 250 UAH in the first region, and 270 UAH in the second one, comparison of the first indicator and the second one gives a territorial index of 0.93. That is, an individual index is obtained which shows that the average price in the first region is 7 % lower than in the second one.

The construction of total territorial indices raises the problem of choosing a comparison base and an object at the level of which the index weight should be fixed. In each case, it is solved based on the purpose of study [31; 34; 40].

The comparison of indicators can be carried out either across two territories (objects) or a number of territories (objects). In the first case, the base may be the indicator of any of the territories, while in the second case, the comparison base should be economically justified. Thus, if, for example, the profitability of a number of the same type of enterprises with approximately the same technical and economic conditions of production is compared, the enterprise which has the highest level of profitability should be taken as the base for comparison.

For *the computation of territorial indices* of qualitative indicators the following values are taken as weight:

the average of the quantitative indicator for the totality of units of the compared territories;

a quantitative indicator related to the territory where the intensive indicator is more economical;

a quantitative indicator that is accepted as a standard.

In constructing *territorial indices* of quantitative indicators, the average level of a qualitative indicator can serve as a weight:

for the territory for which comparison is made;

established for the territory accepted as a standard.

In dynamic comparisons, as has been noted, (for example, in the aggregate price index) the amount of products produced in the reporting period is taken as weight. But for territorial comparisons, the terms "reporting period" and "baseline period" are notional. When comparing region A with region B, the base is the price level in region B, and the quantity of production in region A should be taken as weight.

In this case, the price index will look like:

$$I_p = \frac{\Sigma p_A q_A}{\Sigma p_B q_A}. \qquad (12.19)$$

However, it is not necessary to compare region A with region B. On the same basis, one can compare region B with region A. For such a comparison, the price level in region A will be the base, and the quantity of products in region B will be the weight of the "reporting period". Therefore, the price index will look like this:

$$I_p = \frac{\Sigma p_B q_B}{\Sigma p_A q_B}. \qquad (12.20)$$

Such differences in territorial indices arise due to the fact that the compared populations differ in the range of products, structure of production, structure of fixed assets, composition of employees, etc. In these cases, to form objective conclusions and an unambiguous answer, it is necessary to align the populations in structure.

Statistics offers different ways of solving this problem, including standard weights. Its essence is that the value of the indexed parameter is not weighted by the weights of one territorial unit, but by the weights of the economic district, region, country in which the territories (objects) are located.

So, assuming that q = $q_A + q_B$, whatever region is taken as the base for comparison, the results obtained will not contradict each other and the quality index (price index) will be of the form:

$$I_p = \frac{\Sigma p_A q}{\Sigma p_B q}. \tag{12.21}$$

In the territorial indices of quantitative indicators, you can choose as weight the average levels of the corresponding qualitative indicators (prices, cost, investment, labor, yield, etc.), calculated on average over the compared territories:

$$I_q = \frac{\Sigma q_A \overline{p}}{\Sigma q_B \overline{p}}, \tag{12.22}$$

where $\overline{p} = \frac{p_A q_A + p_B q_B}{q_A + q_B}$.

Territorial indexes are widely used to make international comparisons.

12.6. The index factor analysis method

Indices are used to determine both the dynamics of a complex phenomenon and the impact of individual factors on this phenomenon. The index method allows us to evaluate the influence of individual factors in both relative and absolute terms.

In statistical practice, the following *rule of factor analysis* is adopted: if an effective indicator can be presented as a product of a quantitative and qualitative factor, by determining the influence of the quantitative factor on the change in the effective indicator, the qualitative factor is fixed at the level of the base period; if the influence of the qualitative indicator is determined, the quantitative factor is fixed at the level of the reporting period [31; 40].

The most common index systems are:

physical volume (I_q) and price (I_p) indices which are factor indices relative to the product value index:

$$I_{pq} = I_p \times I_q; \tag{12.23}$$

cost (I_z) and product physical volume (I_q) indices which are factor indices relative to the index of production costs:

$$I_{zq} = I_z \times I_q; \tag{12.24}$$

workforce (time spent) (I_T) and wage (I_f) indices which are factor indices relative to the wage bill index:

$$I_F = I_T \times I_f \tag{12.25}$$

or

$$\frac{\Sigma f_1 T_1}{\Sigma f_0 T_0} = \frac{\Sigma f_0 T_1}{\Sigma f_0 T_0} \times \frac{\Sigma f_1 T_1}{\Sigma f_0 T_1}; \tag{12.26}$$

indices of the number of workers (I_T) and labor productivity (I_W) which are factors in relation to the index of the output:

$$I_Q = I_T \times I_w \tag{12.27}$$

or

$$\frac{\Sigma W_1 T_1}{\Sigma W_0 T_0} = \frac{\Sigma W_0 T_1}{\Sigma W_0 T_0} \times \frac{\Sigma W_1 T_1}{\Sigma W_0 T_1}; \tag{12.28}$$

fixed assets ($I_{\overline{\overline{F}}}$) and return on assets ($I_v$) indices which are factors relative to the output index:

$$I_Q = I_{\overline{\overline{F}}} \times I_V \tag{12.29}$$

or

$$\frac{\Sigma V_1 \overline{\overline{F}}_1}{\Sigma V_0 \overline{\overline{F}}_0} = \frac{\Sigma V_0 \overline{\overline{F}}_1}{\Sigma V_0 \overline{\overline{F}}_0} \times \frac{\Sigma V_1 \overline{\overline{F}}_1}{\Sigma V_0 \overline{\overline{F}}_1}. \tag{12.30}$$

These index systems are called *multiplicative models* [31; 34].

The index systems under consideration are two-factor systems. But, if the effective variable depends on three or more factors, multifactor index models are constructed.

There are two methods of decomposition of the general index into factors:

the method of isolated study of factors;

the chain substitution method (interrelated factor study).

Since in the economic practice the phenomena under consideration are complex and, as a consequence, interrelated, it is advisable to use the second method, which requires a correct location of factors in the structure of the model of the dependent variable (for example, W = a × b × c × d).

The qualitative factor should be placed first in the model. Increasing the chain of factors by one factor (for example, a × b) each time should lead to an indicator with a real economic sense.

To determine the impact of the first factor, all other factors are taken at the reporting period level (in the numerator and denominator).

To determine the impact of the second factor, the first factor is kept at the base period level, the third and all subsequent factors are kept at the reporting period level, etc.

For example, w = a × b × c × d. From here, sequential factor decomposition will look like:

$$I_w = \frac{W_1}{W_0} = \frac{a_1b_1c_1d_1}{a_0b_0c_0d_0} = I_a \times I_b \times I_c \times I_d \qquad (12.31)$$

or

$$\frac{a_1b_1c_1d_1}{a_0b_0c_0d_0} = \frac{a_1b_1c_1d_1}{a_0b_1c_1d_1} \times \frac{a_0b_1c_1d_1}{a_0b_0c_1d_1} \times \frac{a_0b_0c_1d_1}{a_0b_0c_0d_1} \times \frac{a_0b_0c_0d_1}{a_0b_0c_0d_0} = \frac{a_1b_1c_1d_1}{a_0b_0c_0d_0}. \qquad (12.32)$$

Index systems are used to determine, in absolute terms, the change in a complex phenomenon due to the influence of individual factors. The calculations associated with the specified definition, in absolute terms, of changes in the effective indicator due to individual factors, is called the decomposition of *absolute increment (reduction*) according to factors.

So, the considered four-factor index system (12.31 – 12.32) can be represented in absolute terms:

$$\begin{gathered} a_1b_1c_1d_1 - a_0b_0c_0d_0 = \\ = (a_1 - a_0)\, b_1c_1d_1 + (b_1 - b_0)\, a_0c_1d_1 + (c_1 - c_0)\, a_0b_0d_1 + (d_1 - d_0)\, a_0b_0c_0. \end{gathered} \qquad (12.33)$$

This kind of model is called *additive*.

Important concepts

Aggregate index is the main form of constructing aggregate indices.

Index weight is a value used to measure indexed values.

Index is a relative value that, based on the systematic consideration of phenomena, characterizes their change in time, space, or degree of deviation from a certain standard (norm).

The variable composition index is the dynamics of average levels of the indicator.

The index of constant (fixed) composition is an index that shows how the average level of the indicator changes due to changes in the qualitative factor (indexable value).

The structural shift index is an index that shows how the indicator average level changes with the change in the quantitative factor (its structure).

Index method is a complex characteristic of a relative change in time, space or in comparison with any standard of the phenomena which due to the presence of functional dependence between them are presented by a system of interrelated indicators based on the principle of presentation of an integral result through its components.

Indexed value is a feature whose change is being studied (price of goods, share price, time spent on work, volume of goods sold).

Factor analysis rule says that if the effective indicator can be presented as a product of quantitative and qualitative factors, when determining the influence of the quantitative factor on the change in the effective indicator, the qualitative factor is fixed at the level of the base period; if the influence of the qualitative indicator is being determined, the quantitative factor is fixed at the level of the reporting period.

The weighted average index is the average of individual indices weighted by volumes that have the same dimension and are fixed at a constant level.

Territorial index is a generalizing indicator that characterizes the correlation of levels of complex economic phenomena in space (in terms of territories or objects).

Typical tasks

Task 1. An analytical firm performs the work of assessment of the value of services provided by social workers (Table 12.1). The impact of the change in the number of social work services on the total time spent on providing these services by all employees should be determined.

Table 12.1

Performance indicators of a firm providing social services

Types of services	Total time spent on the service in 2018, thousand man-hour	Change in the number of services in 2019 compared to 2018, %
Symbols	$q_0 t_0$	i_q
Cleaning	16	−8
Buying products	10	+2

The solution.

Based on the economic content of the labor intensity indicator of the unit of work performed (t) as time spent (T) on the provision of services by social workers (q) we have a record $t = \frac{T}{q}$.

Then T = t × q for 2018 is the base period compared to 2019, and the total time spent this year is denoted by $q_0 t_0$.

The change in the number of provided services is presented in percentage. The base for comparison in percentage is 100 units. Then the change in the number of cleaning services for a certain period is: $\frac{100-8}{100} = \frac{92}{100} = 0.92$, in the purchase of products it is: $\frac{100+2}{100} = \frac{102}{100} = 1.02$.

If a question arises as to the total change in the number of services, a physical volume total index must be calculated. This index in aggregate form, according to the conditions of the problem, will look like:

$$Iq = \frac{\Sigma q_1 t_0}{\Sigma q_0 t_0}.$$

But there is no information on the time spent on services $q_1 t_0$ in 2019.

Then, if $i_q = \frac{q_1}{q_0}$, $q_1 = i_q \times q_0$.

Using the rules of index identity, let's convert the aggregate index to the average:

$$I_q = \frac{\Sigma q_1 t_0}{\Sigma q_0 t_0}.$$

Substitute $q_1 = i_q \times q_0$ in the expression $q_1 t_0$ and get a new index called the arithmetic mean: $I_q = \frac{\Sigma i_q q_0 t_0}{\Sigma q_0 t_0}$.

$$I_q = \frac{\Sigma i_q q_0 t_0}{\Sigma q_0 t_0} = \frac{0.92 \times 16 + 1.02 \times 10}{16 + 10} = \frac{24.92}{26} = 0.958 \text{ or } 95.8\ \%.$$

Consequently, the total time spent by the social services firm was reduced by 4.2 % due to changes in the number of services.

Task 2. The bank's analytical department examines deposit interest rates for legal entities and individuals (Table 12.2). The task of the department is to determine the dynamics of the average deposit rate in the banking institution.

Table 12.2

The data on the deposit interest rates

Depositors	The amount of deposits attracted, thousand UAH		Deposit rate, %	
	base period	current period	base period	current period
Symbols	f_0	f_1	x_0	x_1
Legal entities	960	1040	35	29
Individuals	210	360	22	18

The solution.

Mid-level dynamics is determined using variable composition, fixed composition and structural shift indices.

The average deposit rate variable composition index is:

$$I_{\bar{x}} = \frac{\bar{x}_1}{\bar{x}_0} = \frac{\Sigma x_1 f_1}{\Sigma f_1} : \frac{\Sigma x_0 f_0}{\Sigma f_0},$$

where $\bar{x}_1, \bar{x}_0$ is the average deposit rate in the reporting and base periods;

x_1, x_0 is the deposit rate in the reporting and base periods;

f_1, f_0 is the sum of deposits attracted in the reporting and base periods.

$$I_{\bar{x}} = \frac{0.29\times1040+0.18\times360}{1040+360} : \frac{0.35\times960+0.22\times210}{960+210} = 0.2617:0.3266=0.80$$

or 80 %.

The average deposit rate fixed composition index is:

$$I_{\bar{x}} = \frac{\Sigma x_1 f_1}{\Sigma f_1} : \frac{\Sigma x_0 f_1}{\Sigma f_1} = 0.2617 : \frac{0.35\times1040+0.22\times360}{1040+360} = 0.2617:0.3165=0.8268$$

or 82.68 %.

The structural shift index of the average deposit rate is:

$$I_{\bar{x}} = \frac{\Sigma x_0 f_1}{\Sigma f_1} : \frac{\Sigma x_0 f_0}{\Sigma f_0} = 0.3165:0.3266=0.9690 \text{ or } 96.6\ \%.$$

The interconnection of indices is: $I_{\bar{x}} = I_x \times I_f = 0.8268\times0.969=0.80$.

Therefore, the correlation of the indices ensures that the calculations are correct. The average deposit rate in the reporting period changed as compared with the base level by 20 % (100 – 80). This decrease was due to a decrease in the deposit rate for each group of depositors by 17.32 % (100 – 82.68) and a decrease in the amount of deposits attracted according to the types of depositors by 3.1 % (100 – 96.9).

Task 3. In order to study consumer demand for product A, a marketing survey is to be conducted (Table 12.3). It is necessary to determine in which of the regions and to what extent the level of prices for good A differs under different nomenclature.

Table 12.3

Basic market indicators for product A

Product nomenclature A	Region A		Region B	
	sold, kg	the price of 1 kg, UAH	sold, kg	the price of 1 kg, UAH
Symbols	q_A	p_A	q_B	p_B
CX	450	8.0	480	7.5
AH	240	5.0	210	4.5

The solution.

If in the price index the weights in the region that is compared with another one is fixed, you can build and calculate two indices:

$$I_p = \frac{\Sigma p_A q_A}{\Sigma p_B q_A} = \frac{8 \times 450 + 5 \times 240}{7.5 \times 450 + 4.5 \times 240} = \frac{4800}{4455} = 1.0774, \text{ or } 107.74\ \%,$$

that is, with regard to the structure of the nomenclature of goods A sold in region A, the price level in region A is higher by 7.74 % compared to region B.

However, comparing the prices in region B with the prices in region A, we get:

$$I_p = \frac{\Sigma p_B q_B}{\Sigma p_A q_B} = \frac{7.5 \times 480 + 4.5 \times 210}{8 \times 480 + 5 \times 210} = \frac{4545}{4890} = 0.9294, \text{ or } 92.94\ \%,$$

that is, with regard to the composition of commodity A sold in region B, prices in that region are lower than in region A by 7.06 %.

Thus, in each region, the price level for product A is different when the ratio of price levels is measured in relation to the product structure in the region being compared.

Task 4. The management of the production enterprise aims to investigate the efficiency of workers (Table 12.4). The economics department was tasked with determining how the output of the enterprise changed as a whole and due to the change in the number of employees and their productivity.

Table 12.4

The data on the products made by production units

Production subdivision	Number of products produced, units		Number of workers, people	
	base period	current period	base period	current period
Symbols	q_0	q_1	T_0	T_1
1	4060	3800	200	150
2	2500	2000	100	100

The solution.

Let's determine the levels of productivity (W) for each of the production subdivisions of the enterprise in the reporting and base periods.

Production subdivision No. 1: Production subdivision No. 2:

$$W_1 = \frac{q_1}{T_1} = \frac{3800}{150} = 25.33\ ; \qquad W_1 = \frac{2000}{100} = 20\ ;$$

$$W_0 = \frac{q_0}{T_0} = \frac{4060}{200} = 20.30. \qquad W_0 = \frac{2500}{100} = 25.$$

Now, let's calculate the indices.

If $W = \frac{q}{T}$, $q = W \times T$, then $I_q = I_w \times I_t$,

or $\frac{\Sigma W_1 T_1}{\Sigma W_0 T_0} = \frac{\Sigma W_1 T_1}{\Sigma W_0 T_1} \times \frac{\Sigma W_0 T_1}{\Sigma W_0 T_0}$.

According to the example

$I_q = I_w \times I_t = 1.045 \times 0.850 = 0.89$ or 89 %.

Thus, at the enterprise as a whole, production output reduced by 11 %: with the reduction in the number of employees, production reduced by 15 %, and with the increase of labor productivity, it increased by 4.5 %.

Absolute increments due to individual factors are calculated as the difference between the numerator and the denominator of the respective factor indices.

So, the total absolute increment is defined as:

$$\Delta q = W_1 T_1 - W_0 T_0 = q_1 - q_0.$$

$$\Delta q_{(w)} = W_1 T_1 - W_0 T_1 = (W_1 - W_0) \times T_1,$$

$$\Delta q_{(T)} = W_0 T_1 - W_0 T_0 = (T_1 - T_0) \times W_0.$$

Then we have $\Delta q = \Delta q_{(w)} + \Delta q_{(T)}$.

Let's determine the total absolute increment and its increment due to the influence of individual factors in the production subdivision No. 1:

$\Delta q = 3800 - 4060 = -260$ units, that is, in the reporting period, subdivision No. 1 produced 260 units less as compared to the baseline period. This decrease is due to:

1) the change in labor productivity:

$\Delta q_{(w)} = (25.33 - 20.30) \times 150 = 755$ units, that is, with the increase in labor productivity, the output increased by 755 units;

2) the change in the number of employees:

$\Delta q_{(T)} = (150 - 200) \times 20.3 = -1015$ units, that is, with reduction in the number of workers, 1 015 units of production were not made.

Let's check the relationship: 755 + (−1015) = −260 units.

In the given period, the manufacturing enterprise reduced its production by 11 % which is 260 units in absolute terms. The change in the number of employees contributed to a decrease in the production of products by 15 %, i.e. by 1 015 units, while the increase in the labor productivity of workers by 4.5 % increased the output by 755 units.

Reference to laboratory work

Guidelines for carrying out the laboratory work on the topic "Statistical data analysis by the index method using MS Excel" [100]. The laboratory work is aimed at acquiring statistical analysis skills using the index method in MS Excel.

Questions for self-assessment

1. What is the essence of the index?
2. What are the criteria by which indices are classified?
3. What is a prerequisite for building individual indices?
4. Determine the meaningful nature of aggregate indices.
5. What is the difference between the Paasche and Laspeyres index systems?
6. Under what conditions are weighted average indices applied?
7. Under what conditions can the relationship between chain and base indices be determined?
8. What are the differences in the construction of dynamics indices and territorial indices?
9. What is the difference between the indices of variable and fixed composition?

10. How is the absolute increase in the effective indicator due to individual factors of multipliers determined?

Questions for critical rethinking (essays)

1. Relative, average values and indices: differences in the use in economic research.

2. Basic index systems: conditions of combination in applied research.

3. Features of determination of weights depending on the form of construction of indices.

4. The practice of applying territorial indices.

5. Rules for factor decomposition in the construction of multifactor models.

References

1. Акімов О. В. Статистика в малюнках та схемах : навч. посіб. – Київ : ЦУЛ, 2007. – 168 с.

2. Аксьонова І. В. Історія статистики : конспект лекцій для студентів спеціальності 7.050 110 усіх форм навчання / І. В. Аксьонова, О. В. Авраменко. – Харків : Вид-во ХДЕУ, 2002. – 172 с.

3. Бек В. Л. Теорія статистики: курс лекцій : навч. посіб. / В. Л. Бек. – Київ : ЦУЛ, 2003. – 412 с.

4. Головач А. В. Статистичне забезпечення управління економікою: прикладна статистика : навч. посіб. / А. В. Головач, В. Б. Захожай, Н. А. Головач. – Київ : КНЕУ, 2005. – 334 с.

5. Гончарук А. Г. Основи статистики : навч. посіб. / А. Г. Гончарук. – Київ : ЦНЛ, 2004. – 148 с.

6. Горкавий В. К. Статистика : навч. посіб. / В. К. Горкавий. – Київ : Алерта, 2012. – 608 с.

7. Горна М. О. Статистика для маркетологів : дистанційний курс Moodle для студентів спеціалізацій "Маркетинг", "Підприємництво, торгівля та біржова діяльність" / М. О. Горна. – Київ : КНЕУ, 2018 – 140 с.

8. Григорович Б. А. Технології візуалізації даних / А. Г. Григорович, Б. А. Григорович // International Academy Journal Web of Scholar. – 2018. – Vol.1, No. 4 (22). – P. 23–28.

9. Григорук П. М. Багатовимірне економіко-статистичне моделювання : навч. посіб. / П. М. Григорук. – Львів : Новий Світ-2000, 2006. – 148 с.

10. Двігун А. О. Статистика : навч. посіб. для студентів вищих навчальних закладів / А. О. Двігун, П. А. Борисенко, К. І. Дерев'янко. – Запоріжжя : Акцент Інвесттрейд, 2012. – 307 с.

11. Дегтяр А. О. Статистичні методи в державному управлінні : навч. посіб. / А. О. Дегтяр, М. В. Гончаренко. – вид. 4-те випр. і допов. – Харків : Вид-во ХарРІ НАДУ "Магістр", 2018. – 176 с.

12. Економічна статистика : навч. посіб. / В. М. Соболєв, Т. Г. Чала, О. С. Корепанов та ін. ; за ред. В. М. Соболєва. – Харків : ХНУ ім. В. Н. Каразіна, 2017. – 388 с.

13. Єріна А. М. Організація вибіркових спостережень : навч. посіб. / А. М. Єріна. – Київ : КНЕУ, 2004. – 137 с.

14. Єріна А. М. Статистика : підручник / А. М. Єріна, З. О. Пальян. – Київ : КНЕУ, 2010. – 351 с.

15. Єріна А. М. Статистичне моделювання та прогнозування : підручник / А. М. Єріна, Д. Л. Єрін. – Київ : КНЕУ, 2014. – 170 с.

16. Єріна А. М. Теорія статистики : практикум / А. М. Єріна, З. О. Пальян. – Київ : Товариство "Знання", 2008. – 255 с.

17. Ковтун Н. В. Теорія статистики : підручник / Н. В. Ковтун. – Київ : Знання, 2012. – 399 с.

18. Кулинич О. І. Теорія статистики : підручник / О. І. Кулинич, Р. О. Кулинич. – 6-те вид., перероб. і допов. – Київ : Знання, 2013. – 239 с.

19. Лугінін О. Є. Статистика : підручник / О. Є. Лугінін, С. В. Білоусова. – Київ : ЦНЛ, 2006. – 580 с.

20. Манцуров І. Г. Статистика економічного зростання та конкурентоспроможності країни / І. Г. Манцуров. – Київ : КНЕУ, 2006. – 388 с.

21. Марець О. Р. Представлення статистичної інформації за допомогою графічного методу / О. Р. Марець, О. М. Вільчинська // International Scientific Journal. – 2015. – No. 9. – P. 118–125.

22. Мармоза А. Т. Практикум з теорії статистики : навч. посіб. / А. Т. Мармоза. – 3-тє вид., випр. – Київ : Ельга ; Ніка – Центр, 2007. – 348 с.

23. Мармоза А. Т. Теорія статистики / А. Т. Мармоза. – Київ : Ельга ; Ніка – Центр, 2003. – 392 с.

24. Матковський С. О. Теорія статистики : навч. посіб. / С. О. Матковський, О. Р. Марець. – Київ : Знання, 2010. – 535 с.

25. Моторин Р. М. Статистика для економістів : навч. посібник / Р. М. Моторин, Е. В. Чекотовський. – Київ : Знання, 2009. – 430 с. + компакт-диск.

26. Опря А. Т. Математична статистика : навчальний посібник для студентів економічних спеціальностей сільськогосподарських вузів / А. Т. Опря. – Київ : Урожай, 1994. – 206 с.

27. Опря А. Т. Статистика (модульний варіант з програмованою формою контролю знань) : навч. посіб. / А. Т. Опря. – Київ : ЦУЛ, 2012. – 448 с.

28. Осауленко О. Г. Статистичний словник // НТК статистичних досліджень: Словник / О. Г. Осауленко, О. О. Васєчко, М. В. Пугачова. – Київ : ДП "Інформ.-аналіт. агенство", 2012. – 250 с.

29. Сєрова І. А. Організація статистичних спостережень : конспект лекцій / І. А. Сєрова, І. В. Аксьонова. – Харків : ХНЕУ, 2008. – 236 с.

30. Статистика : навч.-метод. посіб. для самост. вивч. дисц. / А. М. Єріна, Р. М. Моторін, А. В. Головач та ін. ; за заг. ред. А. М. Єріної, Р. М. Моторіна. – Київ : КНЕУ, 2002. – 448 с.

31. Статистика (модульний варіант з програмованою формою контролю знань) : навч. посіб. / А. Т. Опря, Л. О. Дорогань-Писаренко, О. В. Єгорова та ін. ; за ред. А. Т. Опрі. – 2-ге вид., перероб. і допов. – Київ : ЦНЛ, 2017. – 536 с.

32. Статистика : підручник / Р. Я. Баран та ін. – Чернівці : Наші книги. – 2008. – 240 с.

33. Статистика : підручник / С. С. Герасименко, А. В. Головач, А. М. Єріна та ін. ; за наук. ред. д-ра екон. наук С. С. Герасименка. – 2-ге вид., перероб. і допов. – Київ : КНЕУ, 2000. – 467 с.

34. Статистика : навч. посіб. / за ред. О. В. Раєвнєвої. – Харків : ВД "ІНЖЕК", 2011. – 504 с.

35. Статистичні методи обробки та аналізу економічних даних : навч. посіб. для студ. вищих навч. закл. / за ред. Ю. В. Кулєшкова. – Кіровоград : КДТУ, 2003. – 137 с.

36. Стешенко В. С. Птуха М. В. // Енциклопедія історії України : у 10 т. / В. С. Стешенко, М. В. Птуха ; редкол. : В. А. Смолій (голова) та ін. ; Інститут історії України НАН України. – Київ : Наукова думка, 2012. – Т. 9 : Прил. – С. 58.

37. Тарасенко І. О. Статистика : навч. посіб. / І. О. Тарасенко. – Київ : ЦНЛ, 2006. – 344 с.

38. Теорія статистики : навч. посіб. / за ред. П. Г. Вашківа. – Київ : Либідь, 2001. – 320 с.

39. Теорія статистики : навч. посіб. / М. В. Макаренко, І. М. Гойхман, О. О. Гладчук та ін. ; за ред. М. В. Макаренко. – Київ : Кондор, 2010. – 236 с.

40. Ткач Є. І. Загальна теорія статистики : підручник [для студ. вищ. навч. закл.] / Є. І. Ткач, В. П. Сторожук. – 3-тє вид. – Київ : ЦУЛ, 2017. – 442 с.

41. Тринько Р. І. Основи теоретичної і прикладної статистики : навч. посіб. / Р. І. Тринько, М. Є. Стадник. – Київ : Знання, 2011. – 400 с.

42. Уманець Т. В. Загальна теорія статистики : навч. посіб. / Т. В. Уманець. – Київ : Знання, 2006. – 239 с.

43. Штагрет А. М. Статистика : [навчальний посібник] / А. М. Штагрет. – Київ : ЦНЛ, 2005. – 232 с.

44. Щурик М. В. Статистика : навч. посіб. / М. В. Щурик. – [2-ге вид., оновл. і допов.]. – Львів : Магнолія, 2006, 2009. – 546 с.

45. Аксянова А. В. Теория и практика статистики : учеб. пособ. / А. В. Аксянова, Н. Н. Валеев, А. М. Гумеров. – Москва : КолосС, 2008. – 284 с.

46. Балдин К. В. Общая теория статистики : учебное пособие / К. В. Балдин. – Москва : Дашков и К°, 2010. – 312 с.

47. Васильева Э. К. Статистика : учебник / Э. К. Васильева, В. С. Лялин. – Москва : Юнити, 2007. – 399 с.

48. Вуколов Э. А. Основы статистического анализа. Практикум по статистическим методам и исследованию операций с использованием пакетов STATISTICA и EXCEL : учеб. пособ. / Э. А. Вуколов. – Москва : Форум, 2010. – 464 с.

49. Глинский В. В. Статистический анализ : учебное пособие для студентов вузов экономического профиля / В. В. Глинский, В. Г. Ионин. – 3-е изд., перераб. и доп. – Москва : ИНФРА-М, 2002. – 241 с.

50. Гмурман В. Е. Теория вероятностей и математическая статистика : учеб. пособ. для вузов / В. Е. Гмурман. – 12-е изд. – Москва : Высшее образование, 2006. – 479 с.

51. Годин А. М. Статистика : учебник / А. М. Годин. – Москва : Дашков и К°, 2009. – 460 с.

52. Громыко Г. Л. Теория статистики : учебник / Г. Л. Громыко. – Москва : ИНФРА-М, 2010. – 479 с.

53. Дианов Д. В. Прикладная статистика : учебник / Д. В. Дианов. – Москва : Элит, 2006. – 768 с.

54. Дмитриева И. А. Общая теория статистики : учеб. пособ. / И. А. Дмитриева, С. Н. Лысенко. – Москва : Вузовский учебник, 2009. – 219 с.

55. Едронова В. Н. Общая теория статистики / В. Н. Едронова, М. В. Едронова. – Москва : Юристъ, 2001. – 511 с.

56. Елисеева И. И. Статистика : учебник / И. И. Елисеева. – Москва : Юрайт, 2010. – 565 с.

57. Илышев А. М. Общая теория статистики : учебник / А. М. Илышев. – Москва : Юнити – ДАНА, 2008. – 535 с.

58. Ковалевский Г. В. Статистика : учебник / Г. В. Ковалевский. – Харьков : ХНАГХ, 2012. – 444 с.

59. Колков С. В. Статистика : учеб. пособ. / С. В. Колков, К. Э. Плохотников. – Москва : Флинта, 2008. – 288 с.

60. Костин В. Н. Статистические методы и модели : учеб пособ. / В. Н. Костин, Н. А. Тишина. – Оренбург : ГОУ ОГУ, 2004. – 138 с.

61. Лезина М. Л. Статистика : учеб. пособ. для вузов / М. Л. Лезина. – Москва : Элит-2000, 2008. – 368 с.

62. Лялин В. С. Статистика: теория и практика в Excel : учеб. пособ. / В. С. Лялин. – Москва : Финансы и статистика, 2010. – 448 с.

63. Маличенко И. П. Общая теория статистики: практикум с решением : учеб. пособ. / И. П. Маличенко. – Ростов-на-Дону : Феникс, 2010. – 282 с.

64. Мхитарян В. С. Статистика : учебник / В. С. Мхитарян. – Москва : Academia, 2010. – 272 с.

65. Назаров М. Г. Общая теория статистики : учебник для вузов / М. Г. Назаров. – Москва : Омега-Л, 2010. – 410 с.

66. Назаров М. Г. Практикум по общей теории статистики : учеб.-метод. пособ. / М. Г. Назаров. – Москва : Кнорус, 2010. – 178 с.

67. Орлов А. И. Прикладная статистика : учебник для вузов / А. И. Орлов. – Москва : Экзамен, 2006. – 672 с.

68. Петров Л. Ф. Методы динамического анализа экономики / Л. Ф. Петров. – Москва : Инфра-М, 2010. – 238 с.

69. Практикум по статистике : учеб. пособ. / А. П. Зинченко и др. – Москва : КолосС, 2001. – 392 с.

70. Просветов Г. И. Статистика: задачи и решения : учеб.-практ. пособ. / Г. И. Просветов. – Москва : Альфа-Пресс, 2008. – 496 с.

71. Протасов К. В. Статистический анализ экспериментальных данных / К. В. Протасов. – Москва : Мир, 2005. – 142 с.

72. Рудакова Р. П. Практикум по статистике / Р. П. Рудакова, Л. Л. Букин, В. И. Гаврилов. – Санкт-Петербург : ЛГУ им. А. С. Пушкина, 2006. – 251 с.

73. Салин В. Н. Статистика : учеб. пособ. / В. Н. Салин, Э. Ю. Чурилова, Е. П. Шпаковская. – Москва : КноРус, 2008. – 296 с.

74. Сигел Э. Ф. Практическая бизнес-статистика / Э. Ф. Сигел ; пер. с англ. – 4-е изд. – Москва : ИД "Вильямс", 2004. – 1051 с.

75. Статистика : учебник / под ред. И. И. Елисеевой. – Москва : Проспект, 2010. – 448 с.

76. Статистика : учебник / под ред. В. Г. Ионина. – Москва : Инфра-М, 2008. – 445 с.

77. Статистика : учебник / под ред. И. И. Елисеевой. – Москва : Проспект, 2005. – 444 с.

78. Статистика : учеб. пособ. / под ред. В. М. Симчеры. – Москва : Финансы и статистика, 2008. – 368 с.

79. Тарновская Л. И. Статистика : учеб. пособ. для студ. высш. учеб. заведений / Л. И. Тарновская. – Москва : Academia, 2008. – 320 с.

80. Теория статистики с элементами эконометрики : в 2 ч. Ч. 2 : Учебник для академического бакалавриата / отв. ред. В. В. Ковалев. – Москва : Изд-во Юрайт, 2017. – 348 с.

81. Теория статистики : учебник / под ред. Р. А. Шмойловой. – 5-е изд. – Москва : Финансы и статистика, 2007. – 656 с.

82. Berenson M. Applied statistics. A first course / M. Berenson. – New Jersey : Englewood Cliffs, 1988. – 324 p.

83. Business statistics: a decision-making approach / D. F. Groebner, Patrick W. Shannon, P. C. Fry, K. D. Smith. – New Jersey : Prentice Hall, 2008. – 1040 p.

84. Dowdy S. Statistics for research / S. Dowdy, S. Weardon, D. Chilko. – 3rd ed. – Hoboken, N.J. : Wiley Interscience, 2004. – 627 p.

85. Few S. Show Me the Numbers: Designing Tables and Graphs to Enlighten / S. Few. – Oakland, California : Analytics Press, 2004. – 263 p.

86. Graham U. A dictionary of statistics: Invaluable reference work, covering all areas of statistics / U. Graham, I. Cook. – 2nd ed. – Oxford ; New York : Oxford UP, 2006. – 490 p.

87. Keller G. Instructor's solutions manual for statistics for management and economics / G. Keller, B. Warrack. – 5th ed. – Duxbury : Pacific Grove, Ca., 2000. – 738 p.

88. Keller G. Statistics for Management and Economics / G. Keller, B. Warrack. – 5th ed. – Duxbury : Thomson Learning, 1999. – 920 p.

89. Leboucher L. Introduction à la statistique descriptive / L. Leboucher, M.-J. Voisin. – Toulouse : Cépaduès, 2013. – 206 p.

90. Monino J.-L. Statistique descriptive / J.-L. Monino. – 5th ed. – Malakoff Cedex : Dunod, 2017. – 343 p.

91. Py B. Statistique descriptive : Nouvelle méthode pour bien comprendre et réussir / B. Py. – 5-ème ed. – Paris : Economica, 2007. – 350 p.

92. Simon P. The Visual Organization: Data Visualization, Big Data, and the Quest for Better Decisions / P. Simon. – Wiley and SAS Business Series. – New Jersey : Wiley, 2014. – 240 p.

93. Steele J. Beautiful Visualization: Looking at Data through the Eyes of Experts (Theory in Practice) / J. Steele, N. Iliinsky. – Boston : O'Reilly Media Inc., 2010. – 375 p.

94. Офіційний сайт Головного управління статистики в Харківській області. – Режим доступу : http://kh.ukrstat.gov.ua.

95. Офіційний сайт Державної служби статистики України. – Режим доступу : www.ukrstat.gov.ua.

96. Офіційний сайт Міністерства економічного розвитку і торгівлі України. – Режим доступу : http://me.gov.ua.

97. Офіційний сайт Національного банку України. – Режим доступу : https://bank.gov.ua.

98. Про внесення змін до деяких законодавчих актів України у зв'язку з прийняттям Закону України "Про інформацію" та Закону України "Про доступ до публічної інформації" [Електронний ресурс] : Закон України № 1170-VII від 27.03.2014 р. – Режим доступу : https://zakon.rada.gov.ua/laws/show/1170-18.

99. Про державну статистику [Електронний ресурс] : Закон України № 2614-XII від 17.09.1992 р. – Режим доступу : https://zakon.rada.gov.ua/laws/show/2614-12.

100. Статистика [Електронний ресурс] : методичні рекомендації до лабораторних робіт для студентів усіх спеціальностей першого (бакалаврського) рівня / уклад. О. В. Раєвнєва, І. В. Аксьонова, І. А. Сєрова та ін. – Харків : ХНЕУ ім. С. Кузнеця, 2019. – 104 с. – Режим доступу : http://repository.hneu.edu.ua/handle/123456789/21537.

101. Guillot C. Statistiques Descriptives [Електронний ресурс] / C. Guillot. – Режим доступу : http://www.mghassany.com/documents/stat-desc-fall-16-17.pdf.

102. Introduction to R. Grammar of Graphics [Electronic resource]. – Access mode : https://ramnathv.github.io/pycon2014-r/visualize/ggplot2.html.

103. Mazerolle F. Statistique descriptive [Електронний ресурс] / F. Mazerolle. – Режим доступу : www.economie-cours.fr.

104. Tille Y. Resume du Cours de Statistique Descriptive [Електронний ресурс] / Y. Tille. – Режим доступу : http://www.cours_statistique_descriptive[1].pdf.

Annexes

Annex A

Table

The critical values of the correlation ratio η^2 and the coefficient of determination R^2 if $\alpha = 0.05$

k_2	k_1						
	1	2	3	4	5	6	7
5	0.569	0.699	0.764	0.806	0.835	0.854	0.872
6	0.500	0.632	0.704	0.751	0.785	0.811	0.831
7	0.444	0.575	0.651	0.702	0.739	0.768	0.791
8	0.399	0.527	0.604	0.657	0.697	0.729	0.754
9	0.362	0.488	0.563	0.618	0.659	0.692	0.719
10	0.332	0.451	0.527	0.582	0.624	0.659	0.687
12	0.283	0.394	0.466	0.521	0.564	0.600	0.630
14	0.247	0.348	0.417	0.471	0.514	0.550	0.580
16	0.219	0.312	0.387	0.429	0.477	0.507	0.538
18	0.197	0.283	0.345	0.394	0.435	0.470	0.501
20	0.179	0.259	0.318	0.364	0.404	0.432	0.468
22	0.164	0.238	0.294	0.339	0.377	0.410	0.439
24	0.151	0.221	0.273	0.316	0.353	0.385	0.414
26	0.140	0.206	0.256	0.297	0.332	0.363	0.391
28	0.130	0.193	0.240	0.279	0.314	0.344	0.371
30	0.122	0.182	0.227	0.264	0.297	0.326	0.353
40	0.093	0.139	0.176	0.207	0.234	0.259	0.282
50	0.075	0.113	0.143	0.170	0.194	0.216	0.235
60	0.063	0.095	0.121	0.144	0.165	0.184	0.202
80	0.047	0.072	0.093	0.110	0.127	0.142	0.156
100	0.038	0.058	0.075	0.090	0.103	0.116	0.128
120	0.032	0.049	0.063	0.080	0.087	0.098	0.109
200	0.019	0.030	0.038	0.046	0.053	0.060	0.067
400	0.010	0.015	0.019	0.023	0.027	0.031	0.034

Note: k_1 and k_2 are degrees of freedom.

Annex B

The table of values of F for significance level α = 0.05

k_2 k_1	1	2	3	4	5	6	8	12	24	∞
1	161.5	199.5	215.7	224.6	230.2	233.9	238.8	243.9	249.1	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	604	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.91
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.79
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.72	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.47	2.30	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
35	4.12	3.26	2.87	2.64	2.48	2.37	2.22	2.04	1.83	1.57
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.53
50	4.03	3.18	2.79	2.56	2.40	2.29	2.13	1.95	1.74	1.44
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
∞	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75	1.52	1.00

Annex C

The table of values of F for significance level α = 0.01

k_1 / k_2	1	2	3	4	5	6	8	12	24	∞
1	4052.1	4999.0	5403.5	5625.1	5764.1	5889.4	5981.3	6105.8	6234.2	6366.5
2	98.49	99.01	99.17	99.25	99.30	99.33	99.36	99.42	99.46	99.50
3	34.12	30.81	29.46	28.71	28.24	27.91	27.49	27.05	26.60	26.12
4	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.37	13.93	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.27	9.89	9.47	9.02
6	13.74	10.92	9.78	9.15	8.75	8.47	8.10	7.72	7.31	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47	6.07	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67	5.28	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11	4.73	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71	4.33	3.91
11	9.65	7.20	6.22	5.67	5.32	5.07	4.74.	4.40	4.02	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16	3.78	3.36
13	9.07	6.70	5.74	5.20	4.86	4.62	4.30	3.96	3.59	3.16
14	8.86	6.51	5.56	5.03	4.69	4.46	4.14	3.80	3.43	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67	3.29	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55	3.18	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.45	3.08	2.65
18	8.28	6.01	5.09	4.58	4.25	4.01	3.71	3.37	3.01	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.30	2.92	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23	2.86	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.17	2.80	2.36
22	7.94	5.72	4.87	4.31	3.99	3.75	3.45	3.12	2.75	2.30
23	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.07	2.70	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.03	2.66	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.32	2.99	2.62	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.29	2.96	2.58	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.26	2.93	2.55	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.23	2.90	2.52	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.20	2.87	2.49	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84	2.48	2.01
35	7.42	5.27	4.40	3.91	3.59	3.37	3.07	2.74	2.37	1.90
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66	2.29	1.82
45	7.23	5.11	4.25	3.77	3.45	3.23	2.94	2.61	2.23	1.75
50	7.17	5.06	4.20	3.72	3.41	3.19	2.89	2.56	2.18	1.68
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50	2.12	1.60
∞	6.64	4.60	3.78	3.32	3.02	2.80	2.51	2.18	1.79	1.00

The critical values of the selective correlation coefficient for different numbers of degrees of freedom and significance levels

Number of degrees of freedom k = n − 2	Significance level		Number of degrees of freedom k = n − 2	Significance level	
	a = 0.05	a = 0.01		a = 0.05	a = 0.01
1	0.997	0.999	24	0.388	0.496
2	0.950	0.990	25	0.381	0.487
3	0.878	0.959	26	0.374	0.478
4	0.811	0.917	27	0.367	0.470
5	0.754	0.874	28	0.361	0.463
6	0.707	0.834	29	0.355	0.456
7	0.666	0.798	30	0.349	0.449
8	0.632	0.765	35	0.325	0.418
9	0.602	0.735	40	0.304	0.393
10	0.576	0.708	45	0.288	0.372
11	0.553	0.684	50	0.273	0.354
12	0.532	0.661	60	0.250	0.325
13	0.514	0.641	70	0.232	0.302
14	0.497	0.623	80	0.217	0.283
15	0.482	0.606	90	0.205	0.267
16	0.468	0.590	100	0.195	0.254
17	0.456	0.575	125	0.174	0.228
18	0.444	0.561	150	0.159	0.208
19	0.433	0.549	200	0.138	0.181
20	0.423	0.537	300	0.113	0.148
21	0.413	0.526	400	0.098	0.128
22	0.404	0.515	500	0.088	0.115
23	0.396	0.505	1000	0.062	0.081

Annex E

Table E.1

Chaddock scale

η	0.1 – 0.3	0.3 – 0.5	0.5 – 0.7	0.7 – 0.9	0.9 – 0.99
Strength of relationship	Weak	Moderate	Noticeable	Close	Very close

Table E.2

The quantiles of X^2 distribution for α = 0.05

k	0.95	k	0.95
1	3.84	11	19.68
2	5.99	12	21.03
3	7.82	13	22.36
4	9.49	14	23.69
5	11.07	15	25.00
6	12.59	16	26.30
7	14.07	17	27.59
8	15.51	18	28.87
9	16.92	19	30.14
10	18.31	20	31.41

The critical values of Spearman's rank correlation coefficient for right-handed verification for α

Sample size	0.10	0.05	0.025	0.01	0.005	0.001
n = 4	0.8000	0.8000	–	–	–	–
5	0.7000	0.8000	0.9000	0.9000	–	–
6	0.6000	0.7714	0.8286	0.8857	0.9429	–
7	0.5357	0.6786	0.7450	0.8571	0.8929	0.9643
8	0.5000	0.6190	0.7143	0.8095	6.8571	6.9286
9	0.4667	0.5833	0.6833	0.7667	0.8167	0.9000
10	0.4424	0.5515	0.6364	0.7333	0.7818	0.8667
11	0.4182	0.5273	0.6091	0.7000	0.7455	0.8364
12	0.3986	0.4965	0.5804	0.6713	0.7273	0.8182
13	0.3791	0.4780	0.5549	0.6429	0.6978	0.7912
14	0.3626	0.4593	0.5341	0.6220	0.6747	0.7670
15	0.3500	0.4429	0.5179	0.6000	0.6536	0.7464
16	0.3382	0.4265	0.5000	0.5824	0.6324	0.7265
17	0.3260	0.4118	0.4853	0.5637	0.6152	0.7083
18	0.3148	0.3994	0.4716	0.5480	0.5975	0.6904
19	0.3070	0.3895	0.4579	0.5333	0.5825	0.6737
20	0.2977	0.3789	0.4451	0.5203	0.5684	0.6586
21	0.2909	0.3688	0.4351	0.5078	0.5545	0.6455
22	0.2829	0.3597	0.4241	0.4963	0.5426	0.6318
23	0.2767	0.3518	0.4150	0.4852	0.5306	0.6186
24	0.2704	0.3435	0.4061	0.4748	0.5200	0.6070
25	0.2646	0.3362	0.3977	0.4654	0.5100	0.5962
26	0.2588	0.3299	0.3894	0.4564	0.5002	0.5856
27	0.2540	0.3236	0.3822	0.4481	0.4915	0.5757
28	0.2490	0.3175	0.3749	0.4401	0.4828	0.5660
29	0.2443	0.3113	0.3685	0.4320	0.4744	0.5567
30	0.2400	0.3059	0.3620	0.4251	0.4665	0.5479

Index of entries

Content

НАВЧАЛЬНЕ ВИДАННЯ

Раєвнєва Олена Валентинівна
Аксьонова Ірина Вікторівна
Бровко Ольга Іванівна та ін.

СТАТИСТИКА

Навчальний посібник
(англ. мовою)

За загальною редакцією
д-ра екон. наук, професора О. В. Раєвнєвої

Самостійне електронне текстове мережеве видання

Відповідальний за видання *О. В. Раєвнєва*

Відповідальний редактор *М. М. Оленич*

Редактор *З. В. Зобова*

Коректор *З. В. Зобова*

Подано практико-орієнтований навчальний матеріал, спрямований на поглиблення спеціальних знань зі статистики та підвищення якості набуття студентами професійних компетентностей. Матеріал організований у вигляді взаємозв'язку компонентів триєдиної системи: прикладних теоретичних знань зі статистики, економіко-математичних методів оброблення статистичної інформації та інформаційної підтримки оброблення й аналізу даних, що відповідає сучасним вимогам підготовки компетентних фахівців у галузях народного господарства.

Рекомендовано для студентів усіх спеціальностей, аспірантів, викладачів, економістів, аналітиків, які проводять прикладні статистичні дослідження.

План 2020 р. Поз. № 28-ЕНП. Обсяг 377 с.

Видавець і виготовлювач – ХНЕУ ім. С. Кузнеця, 61166, м. Харків, просп. Науки, 9-А

Свідоцтво про внесення суб'єкта видавничої справи до Державного реєстру
ДК № 4853 від 20.02.2015 р.