Basic Statistics for Economists

Basic Statistics for Economists

By

Natalia Kovtun

Cambridge
Scholars
Publishing

Basic Statistics for Economists

By Natalia Kovtun

This book first published 2022

Cambridge Scholars Publishing

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

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

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

TABLE OF CONTENTS

1. THE DEVELOPMENT OF STATISTICS AS A SCIENCE

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The major precursors to the development of statistics were:

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

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

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

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

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

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

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

2. THE METHODS AND MAIN CONCEPTS OF STATISTICS: STATISTICAL OBSERVATION

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

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

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

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

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

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

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

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

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

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

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

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

- Methods of mass observation;

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

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

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

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

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

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

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

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

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

2.3. Statistical indicators. System of indicators

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

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

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

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

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

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

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

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

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

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

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

2.4. Statistical observation

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3. SUMMARY AND GROUPING OF STATISTICAL DATA

3.1. Concept of summary and grouping

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

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

1) Material grouping;
2) Development and measurement of the indicators that characterize typical groups and sub-groups;
3) Computation of group and general summaries;
4) Presentation of the results in the form of statistical tables and figures.

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

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

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

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

Table 3.1. Grouping by hair color and eye color

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

3.2. Methodology of grouping

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

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

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

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

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

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

3.3. Regrouping of statistical data

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

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

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

We use frequencies to define relative density with the formula:

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

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

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

Table 3.2. Distribution of workers by percentage of output rate

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

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

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

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

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

Table 3.4. Grouping of workers by percentage of output rate

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

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

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

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

The results of regrouping are presented in Table 3.6.

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

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

3.4. Rules for creating statistical tables

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

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

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

There are certain rules on how to make tables:

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

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

Table №

HEADING

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

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

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

Figure 3.1. Model of a statistical table