**MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE**

**SIMON KUZNETS KHARKIV NATIONAL UNIVERSITY OF ECONOMICS**

***I. Serova***

# STATISTICS

**Summary of Lectures**

**Харків. Вид. ХНЕУ ім. С. Кузнеця, 2014**

UDC 311.2(075.8)
BBC 60.6я73
S49

Reviewer – Doctor of Science in Economics, Professor, Head of Management and Business Department of Simon Kuznets National University of Economics *T. Lepeyko.*

Затверджено на засіданні кафедри статистики та економічного прогнозування.

Протокол № 13 від 07.03.2014 р.

**Serova I.**

S49 Statistics : summary of lectures for full-time students of training directions 6.140103 "Tourism", 6.030601 "Management" of specialization "Business Administration" / I. Serova. – Kh. : Publishing House of S. Kuznets KhNUE, 2014. – 100 p. (English)

The methods of statistical methodology are considered. The theoretical material is supported by the description of practical application of statistical methods to the assessment of the economy state and development. Much attention is paid to the analysis of statistical data and the use of mathematical methods in economic research.

Tasks for independent work are provided which correspond to the topics of the curriculum and have an emphasis on calculations of different levels of complexity.

Recommended for students of training directions 6.140103 "Tourism", 6.030601 "Management".

Розглянуто прийоми статистичної методології. Теоретичний матеріал підкріплено описом практичного застосування статистичних методів для оцінювання стану та розвитку економіки. Значну увагу приділено проблемі аналізу статистичних даних та використанню математичних методів в економічних дослідженнях.

Наведено завдання для самостійної роботи, які відповідають темам навчального плану та мають розрахункове спрямування різного рівня складності.

Рекомендовано для студентів напрямів підготовки 6.140103 "Туризм", 6.030601 "Менеджмент".

**UDC 311.2(075.8)**
**BBC 60.6я73**

# Introduction

In today's workplace, students can have an immediate competitive edge over both new graduates and experienced employees if they know how to apply statistical analysis skills to real-world decision-making problems.

This summary of lectures is designed to provide an introductory statistics text for students who do not necessarily have an extensive mathematics background but who need to understand how statistical tools and techniques are applied to business decision-making.

The key principles of these lectures are as follows:

students need to be shown the relevance of statistics,

students need to be familiar with the software used in the business world,

students need to be given sufficient guidance on using software,

students need ample practice in order to understand how statistics is used in business.

Students need a frame of reference when learning statistics, especially when statistics is not their major.

That frame of reference for students should be the functional areas of business – that is, accounting, finance, information systems, management, and marketing.

The focus in teaching each topic should be on its application to business, the interpretation of results, the presentation of assumptions, the evaluation of the assumptions, and the discussion of what should be done if the assumptions are violated.

Introductory statistics courses should recognize that in business, spreadsheet software is typically available on a decision maker's desktop.

Both classroom examples and homework exercises should involve actual or realistic data as much as possible.

Students should work data sets, both small and large, and be encouraged to look beyond the statistical analysis of data to the interpretation of results in a managerial context.

The main goal of the academic discipline is to develop the necessary theoretical knowledge and practical skills for working under modern conditions of social and economic information.

After studying the material presented in the summary of lectures, students should acquire the following professional competencies.

1. The ability to apply scientific principles of information preprocessing according to a certain social and economic problem.

2. The ability to identify cause-and-effect relations between social and economic phenomena and to form an overall scheme of their assessment.

3. The ability to detect regularity and to determine the forms of its functioning in the existing statistical populations.

4. The ability to create a general system of statistical indicators to assess the social and economic problems at various levels.

5. The ability to develop a statistical observation plan in accordance with a certain social and economic situation.

6. The ability to identify and combine methods of statistical and economic analysis to assess social and economic content of the problem.

7. The ability to form representative research tools.

8. The ability to generate and define the main characteristics of the statistical distribution for a population.

9. The ability to apply methods of data visualization based on the analytical material availability.

10. The ability to generate analytical reviews.

11. The ability to use common software packages to obtain the generalized characteristics of social and economic phenomena.

***Analytical activities*** are crucial for bachelors in management and tourism.

This summary of lectures is a generalization of domestic and foreign scientists' researches in the field of statistical analytical activities. The material is presented in a set of topics.

In each topic the theoretical material is confirmed by practical examples.

The summary of lectures contains tasks for independent work of various levels of complexity.

The summary of lectures has a large number of visual materials that can help students to understand the present texts and promote mastering of the academic discipline.

"…our world has three different forms of lie:
the white lie, telling a lie and ***… statistics***"

Unknown member of the London Royal Scientific Society

# Topic 1. Methodological Principles of Statistics

- The subject and object of statistics
- The categories and concepts in statistics

## *1.1. The subject and object of statistics*

"… **statistics knows everything** …", may be because of it this academic discipline is obligatory for all economic specializations in all universities.

*Economics* is the basis for statistics.

*Mathematics* is an instrument for analysis of economic processes.

The main idea of this discipline is the following (Fig. 1.1).

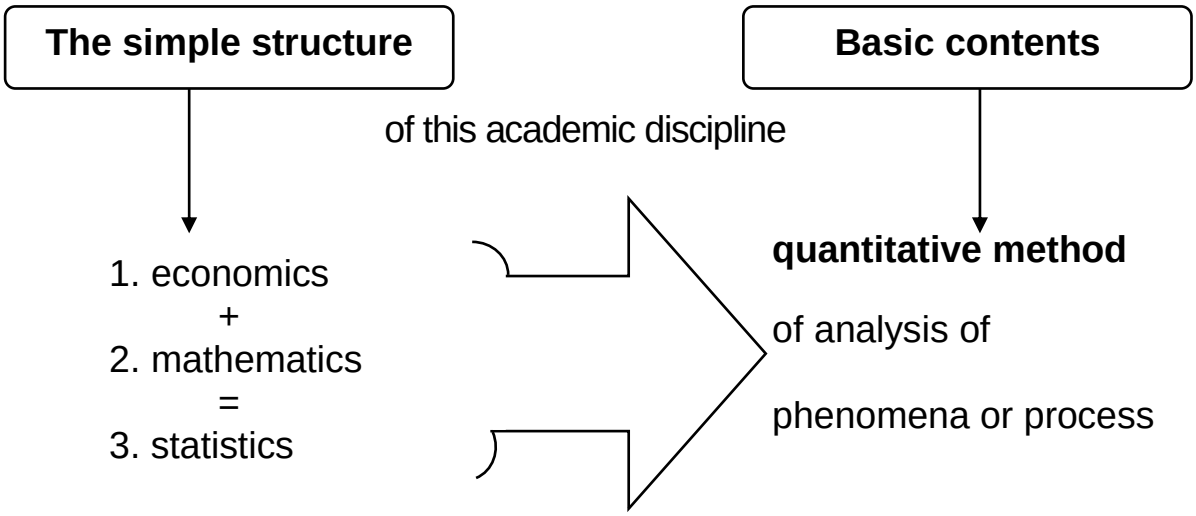



Fig. 1.1. **The main idea of Statistics**

In our country and analytical practice of other countries the contents of statistics is considered as both the method of accounting and the method of analysis.

The basic problems of the method of accounting are data searching, data collection and levels of measurement.

The basic problems of the method of analyses are to select a method for a specific purpose, to explain this result.

The problems of the method of accounting are solved by descriptive statistics.

The problems of the method of analyses are solved by inferential statistics.

The subject-oriented approach to the contents of the academic discipline is as follows (Fig. 1.2).

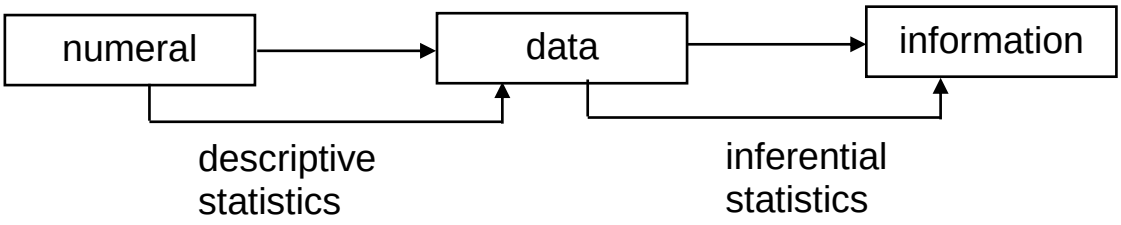



Fig. 1.2. **The subject-oriented approach**

Statistics works with numerals and uses descriptive statistics for transforming numerals into data, and statistics uses inferential statistics for transforming data into information.

The first topics of the course include the basic concepts and approaches for studying this academic discipline.

Statistics is a social science. Statistics studies the quantitative aspect of mass socio-economic phenomena and their qualitative characteristics. It studies their structure and distribution, location in space, direction and speed of change over time, trends and regularities, the density of interconnections and interdependence.

The **object** of this subject is society, phenomena and processes of public life.

The **subject** of statistics is quantitative and qualitative interconnections of mass phenomena at specific place and time.

The **peculiarities** of the subject are:

the study of mass socio-economic phenomena,

the description of quantitative aspects of these phenomena taking into account the place and time in which they occurred,

the identification and measurement of regularities of mass phenomena and processes.

**The tasks of the subject** are:

to study regularities of mass socio-economic processes;

to study parts of a population which consists of homogeneous units;

to study dynamics of phenomena.

## *1.2. The categories and concepts in statistics*

**The theoretical basis** of statistics includes categories and concepts of the economic theory.

The **law of large numbers –** when considering a large mass of phenomena occasional deviation is not taken into account.

*For example*, if all the students of the group regularly attend classes during the semester and the monitor of the group skipped classes once we will inform the Dean office that the attendance is 100 %.

**Variables are** characteristics of items or individuals and are what you analyze when you use a statistical method.

**Statisticians classify variables** as either being categorical or numerical and further classify numerical variables as having either discrete or continuous values.

**Categorical variables** (qualitative variables) have values that can only be placed into categories.

**Numerical variables** (quantitative variables) have values that represent quantities.

**Discrete variables** have numerical values that arise from a counting process.

**Continuous variables** produce numerical responses that arise from a measuring process.

Using levels of measurement is another way of classifying data.

There are four widely recognized **levels of measurement**:

nominal,
ordinal,
interval,
ratio scales.

A **nominal scale** classifies data into distinct categories in which no ranking is implied.

An **ordinal scale** classifies data into distinct categories in which ranking is implied.

An **interval scale** is an ordered scale in which the difference between measurements is a meaningful quantity but does not involve a true zero point.

A **ratio scale** is an ordered scale in which the difference between measurements involves a true zero point.

**Quantitative data** are measurements whose values are inherently numerical.

**Qualitative data** are data whose measurement scale is inherently categorical.

A **population** consists of all the items or individuals about which you want to draw a conclusion.

A **parameter** is a numerical measure that describes a characteristic of a population.

A **statistic** is a numerical measure that describes a characteristic of a sample.

A **sample** is the portion of a population selected for analysis.

**Statistical regularity** is manifested by the main causes of the phenomena in terms of the law of large numbers

**Statistical regularity** is consistency and order in mass processes or phenomena. This consistency manifests itself only in statistical **population.**

**Types of regularities:**

- **the regularity of development** (dynamics) of the phenomena,

e.g. the increase in the Earth population, decrease in the amount of manufactured goods;

- **the regularity of structural changes**,

e.g. the increase in the part of older population;

- **the regularity of distribution of population elements**,

e.g. the distribution of workers according to their salary (or qualification);

- **the regularity of phenomena interdependence**,

e.g. the dependence of the profit on the amount of sales, the dependence of birthrate on women's age.

**Statistical population** is a certain set of elements that are combined by the conditions of their development and existence.

The peculiarities of the statistical population are:

- qualitative homogeneity,
- the change of a sign within certain limits.

Statistics studies **homogeneous** population.

**The homogeneous** population is a population whose elements have similar properties and belong to the same type,

e.g. the data on employment of workers and equipment utilization can not be combined into a single set.

**Applying the concepts**

1. Go to the official site of the State Statistics Committee of Ukraine (www.ukrstat.gov.ua) and read today's top analysis.

Give an example of a numerical variable found in the poll.

Give an example of a categorical variable found in the poll.

2. For each of the following variables, determine whether the variable is categorical or numerical:

the number of emails received in a week,

a favorite department store,

the amount of time spent shopping in the bookstore,

the name of the Internet provider.

3. In a business publication such as *Statistics of Ukraine*, find a graph or chart representing quantitative data and qualitative data. Discuss how the data were gathered and the purpose of the graph or chart.

# Topic 2. Statistical Observation

• The contents of statistical observation as a method of information provision.

• The forms, types and methods of observation.

## *2.1. The contents of statistical observation as a method of information provision*

**Statistical observation** is the planned, scientifically organized mass registration of mass data about any social and economic phenomena and processes.

The **task** of statistics is the account of each population unit, and of individual values of the population variables.

The degrees of registration for statistical observation are:

**primary observation** is a check of the data, which concern an object of research,

**secondary observation** is an original collection of previously recorded and processed data.

Statistics works with data. **The data in statistics** is the mass system of the quantitative characteristics of the social and economic phenomena and processes.

If you want to have some information, you must consider 3 basic **levels of statistical research**:

- statistical observation,
- summarization and grouping of statistical data,
- analysis of statistical data.

When you consider the first level, you work with numerals, the second one deals with data, the third level involves information.

This distribution between the levels of statistical research and the levels of numeral generalization is quite conventional.

In our country statistics is official. The main principle of its organization is centralization.

We have strict subordination for all the levels of organization in official statistics.

In our country the highest level of official statistics is represented by the **State Statistics Committee of Ukraine.**

The regional **State Statistics Committee** is subordinated to the **State Statistics Committee of Ukraine.**

The district **Committee of Statistics** is subordinated to the regional **State Statistics Committee.**

The **positive** characteristic of this hierarchy is the obligatory provision of information by all manufacturing companies, organizations and institutions.

The **negative** characteristics of this hierarchy are the following:

the closed nature of data,

the legal vulnerability for special observation.

These **positive and negative** characteristics taken together pose a global problem for the analytical sphere.

The gist of this problem is as follows: the impossibility of obtaining reliable information, which can reflect the real situation.

The methods of solving this problem are:

selective availability of information for a wide range of users,

training specialists, who know how to combine different forms of monitoring statistics to obtain reliable information.

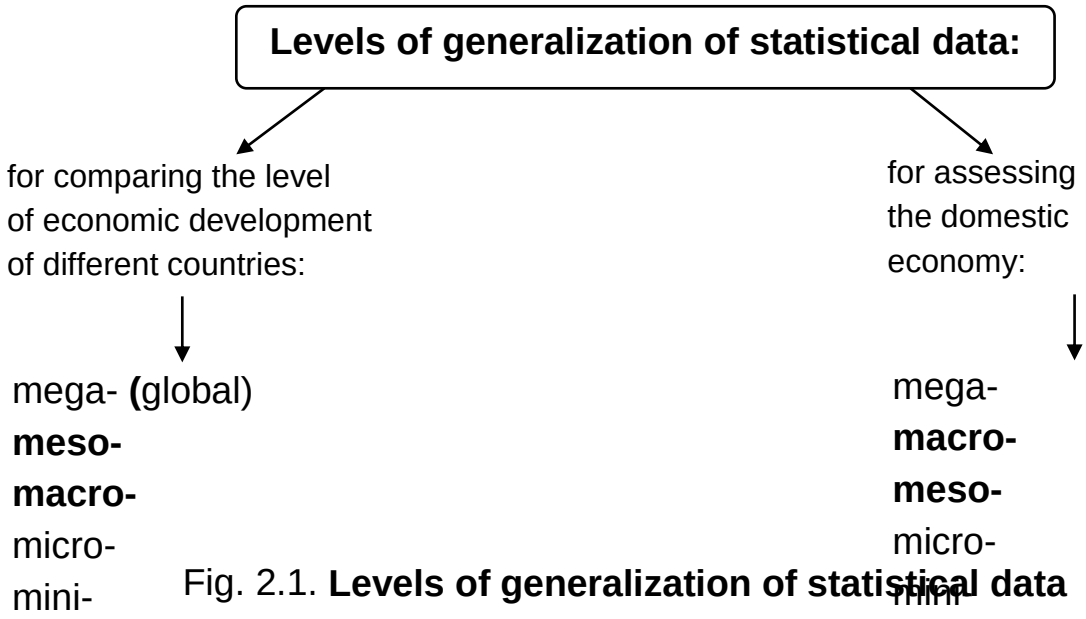



Fig. 2.1. **Levels of generalization of statistical data**

**The UN Statistics Division** generalizes data on the global level.

The **United Nations Statistics Division** (UNSD), under the United Nations Department of Economic and Social Affairs, serves as the central mechanism within the Secretariat of the United Nations. The Division compiles and disseminates global statistical information, develops standards and norms for statistical activities, and supports countries' efforts to strengthen their national statistical systems.

The Division regularly publishes data updates, including the Statistical Yearbook and World Statistics Pocketbook, and books and reports on statistics and statistical methods.

**Eurostat** and the **Interstate Statistical Committee of the CIS** generalize data on the mesolevel.

**Eurostat** is a Directorate-General of the European Commission located in Luxembourg. Its main responsibilities are to provide the European Union with statistical information at the European level and to promote the harmonization of statistical methods across the member states of the European Union. The organizations in the different countries which actively cooperate with Eurostat are summarized under the concept of the European Statistical System.

General statistical activities related to the European Statistical system are:

- coordination and governance of the European Statistical System;
- statistical methodological coordination and research;
- statistical quality and reporting.

**The Interstate Statistical Committee of the CIS** is an interstate authority, acting under the CIS.

**The aims of activity of this organization** are to coordinate national statistical services, to facilitate the organization of information exchange and analysis of social economic development of nations and to develop general recommendations on statistics.

**The State Statistics Committee of Ukraine** generalizes data on the **macrolevel**.

**The State Statistics Committee of Ukraine** is the government agency responsible for collection and dissemination of statistical information in Ukraine.

**The State Statistics Committee of Ukraine** is the central authority with a special status. The basic document regulates legal relations in the sphere of the state statistics, under the Law of Ukraine "About the State Statistics", which defines the rights and functions of bodies of the state statistics, organizational bases of realization of the state statistical activity with the purpose of reception of objective statistical information.

Manufacturing companies, organizations, institutions generalize data on the microlevel.

The structural units of manufacturing companies, organizations and institutions generalize data on the minilevel.

The regional level in these hierarchies is different. For the domestic economy the regional level is the level of a region.

If you compare the level of economic development of different countries, Eurostat and the **ISC of** the **CIS** are considered to be a mesolevel (or the regional level).

The requirements for statistical data are defined as follows:

- **completeness** which is sufficient coverage in the data scope and contents;
- **timeliness** which is timely receipt of the data by the user;
- **reliability** which means matching the actual data to the state;
- **uniformit**y in the space and time of the data, for example, by the structure of population, units of measurement, methods of collecting and **processing**, territorial belonging of the units;
- **accessibility** which implies an opportunity to get acquainted with or to receive the data.

The **statistical observation plan** is the basis for any statistical research.

The statistical observation plan includes two groups of questions:

1) **methodological** questions,

2) **organizational** questions.

The **first group** includes the following ones:

**Why is this observation made?**

This question defines the purpose of the observation.

The purpose of the observation is to collect statistical data to generalize the features of a phenomenon (or a process), as well as to define corresponding regularities.

**What is the object of the observation?**

This question defines the object of the observation and the unit of the observation.

**The object** of the observation is the statistical population of the phenomena that must be researched.

The **unit of the observation** is the unit from which we receive information.

**What do we observe?**

This question defines the object and the unit of population.

A **unit of population** is the primary element of the observed object. The element possesses the signs, which are recorded.

*e.g.* If the object of the observation is the performance of the students of your course, then the performance of one separate student will be a unit of population but the performance of a group will be a unit of the observation.

If the object of the observation is the performance of your group, then the performance of a separate student will be both the unit of population and the unit of the observation.

- **What is the source of information during this observation?**

The information source is the society and the processes taking place in it.

- **What questions must be answered?**

The **observation program** is a set of questions which must be answered during the observation.

It is necessary to use **statistical tools** to conduct a statistical **observation.**

**Statistical tools** are:

- **a statistical blank form** for a registration record,
- **a manual** that is a document that explains the procedure for data registration.

**Organizational questions** for the observation plan.

- **Who conducts this observation?**

This question defines the subject of the observation (the definition of oversight and personnel).

- **Where does the observation take place?**

This question defines the place of the observation.

- **When does the observation take place?**

This question defines the observation time.

During the observation it is necessary to define the objective and subjective time.

The **observation time** is the time when observation is made.

If the object of the observation is a **process** it is necessary to choose a **period of time.** That is the period during which the data are accumulated and collected.

If the object of the observation is the condition of a phenomenon it is necessary to choose **a point in time** – the date when the condition of a phenomenon is observed.

The **subjective time** is the time during which the data are processed.

*e.g.* Information on the number of employees at the company is collected during February, and is sent to statistical bodies by the 15 of March. The period of February is an **objective time**, but the period from the first of March to the 15 of March, is a **subjective time.**

The first **Ukrainian Census** was carried out by the State Statistics Committee of Ukraine on the 5 of December 2001.

The critical moment of the population census was at midnight on the 5 of December, but the date was recorded in the period from the 5 of December to the 12 of December.

The critical moment, is an **objective time, but** the period from the 5 of December to the 12 of December is a **subjective time**.

- **What is the financial and technical provision for the observation?**

This question defines the financial and technical resources.

- **How is the accuracy of the observation guaranteed?**

This question defines control systems: logical control and arithmetic control.

The **control system** of observation.

During the observation there can be two **types of errors**:

- **recording errors**,
- **sampling errors.**

## *2.2. The forms, types and methods of observation*

**The forms of observati**on in **our country** are:

**reporting;**

**special observation** which includes:

- census,
- registration (accounting),
- special surveys;

**statistical register**.

Every enterprise, institution or organization is required to submit timely records about their activities for a certain period of time.

**Special observation** is carried out if the information from the reporting forms is not enough to conduct a study.

**A register** is a list of students in a group.

**Types of observation and their classification**

**The first level** of classification is the classification by the **coverage of population units** which are considered as follows:

- the **running** observation involves each population unit,
- the **non-running** observation does not involve each population unit.

The types of **non-running** observation are:

the **sampling** observation,

the **monographic** observation,

the **study of the main mass data**,

the **monitoring**.

The following type of this classification is the classification according to the **data registration time.**

According to the **data registration time** an observation can be:

**current** which is conducted at all times, that is constantly;

**discontinuous** which is conducted at regular intervals or one-time.

According to **ways of obtaining information,** observation can be:

**direct;**

**documented;**

in the form of a **survey**.

**In international statistics** the following data collection methods are used:

experiments,

telephone surveys,

written questionnaires and surveys,

direct observations and personal interviews.

**Applying the concepts**

1. Give two examples of both the observed object and the unit of observation.

2. Give two examples of both the running observation and the discontinuous observation.

3. Give two examples of both the observed object and the unit of population.

4. Give two examples of both the non-running observation and the current observation.

# Topic 3. Summarization and Grouping of Statistical Data

- Statistical grouping is the basis for scientific data
- Methods of visualization for grouped data

## *3.1. Statistical grouping is the basis for scientific data*

The first phase of organizing and summarizing data is statistical summarization.

The statistical summarizing is the process of ordering, systematization, aggregation and scientific processing of the primary statistical data.

The types of summarization can be differentiated according to the degree of data processing. In statistical practice we use a **simple** summarization and **complicated** summarization.

**The simple** summarization is the calculation of the total results without dividing the population into groups.

**The complicated** summarization is the distribution of population into groups and subgroups according to certain essential signs, and also calculation of the total outcome and group outcomes.

The basic idea of statistical summarization is creating automated data banks, and defining technological schemes for processing the information.

The stages of statistical summarization are:

- the theoretical analysis of the phenomenon investigated,
- the development of a program for summarization.

The development of a program for summarization includes:

definition of grouping signs and the number of groups,

justification of a system of indicators for the group characteristic,

creating the table layout, which will reflect the results of summarization,

- the summarization process itself.

The complicated summarization is the grouping in its essence.

**Statistical grouping** is the distribution of the whole population of the researched social phenomena into types, groups and subgroups according to a certain important sign.

A **grouping sign** is a sign (quantitative or qualitative), according to which the grouping is conducted.

Classification is a statistical standard. The obligatory grouping is named classification.

The peculiarities of statistical classifications are:

common standards for the distribution of all population units;

in most cases a qualitative sign is a grouping sign;

its consistency for a certain period of time.

The following information is very important, because it is used both in our country and abroad.

**Types of statistical grouping** can be classified as follows:

- Depending on the number of grouping signs these types can be:

  simple,

  combinational,

  multi-dimensional.

- Depending on the analytical function these types can be:

***typological,*** that is grouping by a **qualitative** sign. The number of the groups is defined by the structure of the sign;

***structural*** grouping, that is grouping by a **quantitative** sign. Structural grouping studies the structure of a phenomena and structural shift;

***analytical*** grouping, that is grouping by a **quantitative** sign. Analytical grouping is defined by interconnection between a **factorial** sign and a **resultative** sign.

The factorial sign influences other signs and determines their changes.

The resultative sign is changed under the influence of a factorial sign.

***Analytical*** grouping is carried out according **to the factorial sign**. An average level of a resultative sign is determined for each group.

The **number of groups** and the **length of a period** must be determined for grouping.

If you make groupings according to qualitative sings the number of the groups equals the number of the sings, if you make groupings according to quantitative sings you must use **Sturge's** formula (rule):

$$n = 1 + 3{,}322 \lg N, \qquad (3.1)$$

where *n* is the number of groups,

*N* is the number of population units.

The use of this formula is reflected in Table 3.1. This table is more useful for mathematical calculations.

This information is used for measurement of current signs.

Table 3.1

**The dependence of the number of groups on the amount of population**

| **N** | 15 – 24 | 25 – 44 | 45 – 89 | 90 – 179 | 180 – 359 | 360 – 719 | 720 – 1439 |
|---|---|---|---|---|---|---|---|
| **n** | 5 | 6 | 7 | 8 | 9 | 10 | 11 |

If we have even distribution of a sign then the length of the period is defined according to the **formula**:

**The range of the interval (**or the length of the period**)** $h = \frac{x_{max} - x_{min}}{n}$,

where $x_{max}$ is the maximum value of a sign,

$x_{min}$ is the minimum value of a sign.

This formula is valid for studying the even distribution of a sign.

*For example*, there are 30 employees in a company. Each employee has a different record of service. The minimal record of service is 1 year, the maximal record of service is 25 years.

Your task is to make a grouping of employees according to their record of service.

If we study the record of service of 30 employees then the number of groups must be 6.

Then the r**ange of the interval** will be $h = \frac{x_{max} - x_{min}}{n}$, = (25 – 1) / 6 = 4.

The sign under study is expressed by *X*, but the frequency of this sign's occurrence is expressed by *f*.

**Grouping of employees according to their record of service.**

| Group number | Record of service (x) | Number of employees (f) |
|---|---|---|
| 1. | **1 – 5** | **2** |
| 2. | **5 – 9** | **8** |
| 3. | **9 – 13** | **4** |
| 4. | **13 – 17** | **1** |
| 5. | **17 – 21** | **12** |
| 6. | **21 – 25** | **3** |
| **Total** | | **30** |

You have received a distribution of employees according to their record of service**. You** created 6 groups with an equal and closed interval.

Which group must value 21 or 17 be included into?

There is a **rule** that the lower boundary of the closed interval is always included into this interval.

In our practical everyday activity we use another rule**.** If the interval is **closed** and value 25 is the maximal value of a sign, then value 25 is included into the $6^{th}$ group, and value 21 is included into the $5^{th}$ group.

If the interval is **open,** then value 21 is included into the $6^{th}$ group, and value 17 is included into the $5^{th}$ group.

**The types of intervals** can be:

**open** and **closed,**

**equal** and **unequal.**

Unequal intervals are mostly used in practice, but it is easier to work with equal intervals.

Only the **first** and the **last** intervals **can be open**.

*For example*, there are 30 employees in the company. 29 of them have a record of service from 1 to 25 years, but 1 person has his record of service of 40 years.

Then **it is not correct** to define the range of the interval as h = (40 – 1) / 6 = 39 / 6 = 9.

The employees who have their record of service from 1 to 25 years constitute the **main mass data**, so we will be creating a grouping according to this mass data, but you will open the last interval.

That is, h = (25 – 1) / 6 = 4.

You have the following grouping of employees according to their record of service.

| Group number | Record of service (x) | Number of employees (f) |
|---|---|---|
| 1. | 1 – 5 | 2 |
| 2. | 5 – 9 | 8 |
| 3. | 9 – 13 | 4 |
| 4. | 13 – 17 | 1 |
| 5. | 17 – 21 | 12 |
| 6. | 21 and more | 3 |
| **Total** | | **30** |

This approach helps us decrease errors during grouping.

For further calculations the open intervals must be closed.

The first interval must be closed according to the range of interval of the second one.

The last interval must be closed according to the range of the previous interval.

Primary summarizing of data into groups is named **primary grouping**.

For conducting comparative analysis it is sometimes necessary to change the number of groups and the range of the interval.

Regrouping the previously grouped data is named a **secondary grouping**.

There are two main ways to create new groups (secondary groups):

- agglomeration of intervals,
- regrouping according to shares.

## *3.2. Methods of the grouped data visualization*

The grouped data can **be represented** as**:**

- **tables,**
- **charts,**
- **distribution series.**

A statistical table always consists of a subject and a predicate.

A subject in the table is the object of study (researches).

The predicate in the table is the indicators which characterize the given object.

**Types of tables** can be:

- simple,
- group,
- combinational.

A simple table is a list of population elements (signs).

A group table is a grouping created according to a single sign.

A combinational table is a grouping created according to two and more signs.

A statistical table must always have its title.

The **layout** of a statistical table

Table 3.2

**The general title**

| SUBJECT | PREDICATE | | | | |
|---|---|---|---|---|---|
| | Titles of columns | | | | |
| A | 1 | 2 | 3 | 4 | 5 |
| Line titles | | | | | |
| | | | | | |
| | | | | | |
| Total | | | | | |

## Applying the concepts

1. There is some information about the work of 25 companies in the region.

Using this information, select from the list of primary data the indicators that **characterize the efficiency of the capital assets use.**

According to the received information provide the results as a table and explain the type of grouping and the type of the table that you used.

Analyse the data and make the necessary conclusions.

**Primary data**

| Ordinal number of a company | The average annual cost of capital assets, mln UAH | The loss of working time, thousand of man-days | Profit, ml UAH | Production output, mln UAH |
|---|---|---|---|---|
| 1 | 3.4 | 66.0 | 3.5 | 15.7 |
| 2 | 3.1 | 44.0 | 3.3 | 18.0 |
| 3 | 3.5 | 91.0 | 3.5 | 12.0 |
| 4 | 4.1 | 78.0 | 4.5 | 13.8 |
| 5 | 5.8 | 57.4 | 7.5 | 15.5 |
| 6 | 5.2 | 42.0 | 6.9 | 17.9 |
| 7 | 3.8 | 100.0 | 4.3 | 12.8 |
| 8 | 4.1 | 79.8 | 5.9 | 14.2 |
| 9 | 5.6 | 57.0 | 4.8 | 15.9 |
| 10 | 4.5 | 38.0 | 5.8 | 17.6 |
| 11 | 4.2 | 32.0 | 4.6 | 18.2 |
| 12 | 6.1 | 112.0 | 8.4 | 13.0 |
| 13 | 6.5 | 72.0 | 7.3 | 16.5 |
| 14 | 2.0 | 55.7 | 2.1 | 16.2 |
| 15 | 6.4 | 36.0 | 7.8 | 16.7 |
| 16 | 4.0 | 85.2 | 4.2 | 14.6 |
| 17 | 8.0 | 72.8 | 10.6 | 14.8 |
| 18 | 5.1 | 54.6 | 5.8 | 16.1 |
| 19 | 4.9 | 37.0 | 5.3 | 16.7 |
| 20 | 4.3 | 56.4 | 4.9 | 15.8 |
| 21 | 5.8 | 56.0 | 6.0 | 16.4 |
| 22 | 7.2 | 70.4 | 10.4 | 15.0 |
| 23 | 6.6 | 53.6 | 6.9 | 16.5 |
| 24 | 3.0 | 34.9 | 3.5 | 18.5 |
| 25 | 6.7 | 55.4 | 7.2 | 16.4 |

# Topic 4. Generalizing Statistical Indicators

• The essence and types of statistical indicators

## *4.1. The essence and types of statistical indicators*

The statistical indicator is a concept of analytical practice. This concept is a general description of phenomena and processes.

The indicators reflect qualitative and quantitative interrelation between phenomena and processes that are researched.

The general scheme of the indicator looks like a numeric value plus its interpretation.

**Indicator = numeric value + interpretation of this value.**

The phenomenon of correlation between the value and its interpretation is defined in this model (Fig. 4.1).

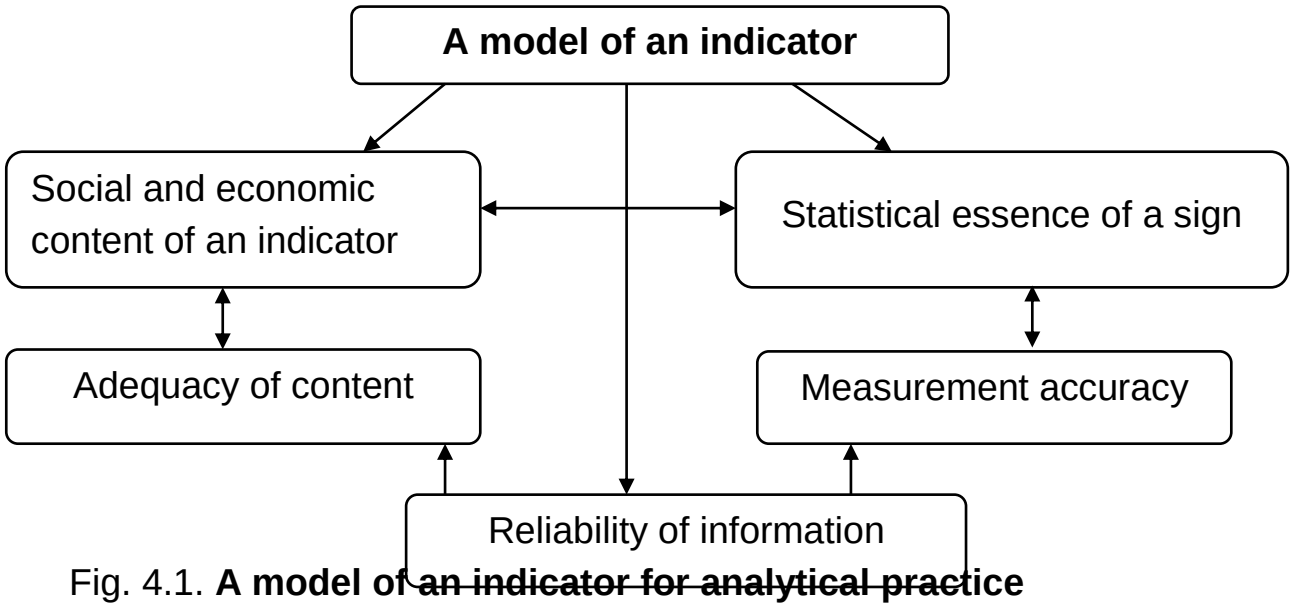



Fig. 4.1. **A model of an indicator for analytical practice**

**Types** of statistical indicators can be classified as follows**:**

**by the analytical function:**

- absolute values,
- relative values,
- average values;

**by the time of processing information:**

- moment,
- interval;

**by the method of calculation:**

- primary,
- secondary values.

***Inverted indicators*** are included into a separate group.

*For example:* the productivity of labor indicator and the labor-intensity indicator.

The **direct** indicator is the productivity of labor. The **inverted** indicator is labor intensity.

**Absolute statistical values c**haracterize numerical parameters of social economical phenomena. They always have units of measurement (figures).

The units of measurement of absolute statistical values are used depending on the purpose of the study and its essence.

**Units of measurement** of absolute values can be:

**natural** (tons, pieces, square meters);

**conventionally natural.**

Can you calculate all the bottles of vodka, which were manufactured by a distillery?

The answer to this question is contained in conventionally natural units.

Accounting in natural units is the most precise in the economic practice:

**combined units (**ton per kilometer, kilowatt per hour of electricity);

**labor units (*man-days*** meaning the number of workers multiplied by all the days of their work, ***man-hours*** meaning the number of workers multiplied by all the days of their work and their working hours);

**cost units** (pound, euro, tugrik).

The **relative statistical values** are a result of the ratio of two absolute statistical values.

Relative values can be expressed in the following **kinds:**

**coefficient** (factor) when a unity (figure one) is the basis for comparison;

**percent** when 100 is the basis for comparison;

**per one thousand** when 1 000 is the basis for comparison;

**per ten thousand** when 10 000 is the basis for comparison.

The **coefficients** (factors) and **percent** are used for solving economical problems. The **per one thousand** and **per ten thousand** are used for solving social-demographic problems.

But this typology is conventional.

The **types of relative** values can be classified as follows:

the relative value of **fulfillment of a plan** (or maintaining contract obligations),

the relative value **of a planned target**,

the relative value **of dynamics**.

This type of relative values can be used for solving economical problems.

The value, which is compared, has the name **currentvalue** (or value under review), and is marked **1**. The value, which is **compared to**, has the name **basic-value**, and is marked **0**.

**r.v.f.p.** = products factually manufactured in March ($y_1$)/products which were planned for March ($y_{pl}$),

**r.v.p.t.** = products which were planned for March ($y_{pl}$)/products factually manufactured in February ($y_0$),

**r.v.d**. = products factually manufactured in March ($y_1$)/products factually manufactured in February ($y_0$).

**The interrelation** between **relative values** is as follows:

**r.v.d**. = **r.v.f.p** · **r.v.p.t**

or

$$(y_1) / (y_0) = (y_1) / (y_{pl}) \cdot (y_{pl}) / (y_0)$$

The **relative value of comparison** characterizes the same values, which belong to different objects.

*For example*, the number of male students in the first group is compared to the number of male students in the second group.

The **relative value of structure** describes the ratio of a part to its integer.

*For example*, the number of male students in the group in comparison with the total number of students in this group.

The **relative value of coordination** describes the ratio of parts to each other.

*For example*, the number of male students in the group in comparison with the number of female students in this group.

The **relative value of intensity** characterizes the degree of extension of the phenomenon in a particular environment.

*For example,* GDP per capita, birth rate.

The **average value** ***($\bar{x}$)*** is a general quantitative characteristic sign of the statistical population. The average value **always** has a **bar** over a symbol**.**

The **average value** is an indicator of central tendency in the distribution of population.

The general representation of the average power – type value – is the following:

$$\overline{X_k} = \sqrt[k]{\frac{X_i^k}{n}}, \qquad (4.1)$$

where *k* is the power on which the choice of the formula depends.

Average values **can be** simple and weighted.

If we have **ungrouped** data we use **simple average** values, if we have **grouped** data we use **weighted** ones (values).

The choice of an average value formula depends on the economic contents of information.

**Forms** of average values are:

**An arithmetic average** value: $\overline{X_1} = \frac{\sum X_i}{n}$, where k = 1.

**A harmonic average** value: $\overline{X_1} = \frac{n}{\sum \frac{1}{X_i}}$, where k = -1.

**A geometric average** value: $\overline{X_0} = \sqrt[n]{X_1 \cdot X_2 \cdot ... X_n} = \sqrt[n]{\prod_1^n X_1}$, where k = 0.

**A square** average value: $\overline{X_2} = \frac{\sum X_i^2}{n}$, where k = 2.

**A chronological average** value: $\overline{x} = \frac{\frac{1}{2}x_1 + x_2 + x_3 + ... + \frac{1}{2}x_n}{n-1}$.

The average values are ranged according to *k* power.

**Majorant** properties of average values are:

$$\overline{X_2} > \overline{X_1} > \overline{X_0} > \overline{X_{-1}}. \qquad (4.2)$$

The **mathematical properties** of arithmetic average values are:

1. The sum of deviations of certain values of a sign from an average value is equal to "zero".
2. If each value of a sign is multiplied by the same number, the average value will change correspondingly.
3. If each value of a sign is added the same, the average value will change correspondingly.
4. If frequency of occurrence of a sign is multiplied by the same number, the average value will not change.
5. The sum of squared deviations of certain values of signs from the average value will be minimal.

The **arithmetic average** values can be **simple** and **weighted**.

The **formula of the simple** value is as follows:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \qquad (4.3)$$

where $x_i$ is a value of a sign for the $i^{th}$ element of the population,

$n$ is the number of elements in the population.

The **formula of the weighted** value is as follows:

$$\overline{X} = \frac{\sum X_i \cdot f_i}{\sum f_i}, \quad (4.4)$$

where $f_i$ is the weight of the $i^{th}$ group or the frequency of occurrence of a sign.

The **harmonic average** values can be **simple** values and **weighted** ones.

The **formula of the simple** value is as follows:

$$\overline{X} = \frac{n}{\sum \frac{1}{X_i}}. \quad (4.5)$$

The **formula of the weighted** value is as follows:

$$\overline{X} = \frac{\sum M_i}{\sum \frac{M_i}{X_i}}, \quad (4.6)$$

where $M_i$ is the weight which is equal to $x_i$ multiplied by $f_i$.

*For example*: There is a company which consists of three departments. The workers of these departments get their salary.

Table 4.1

**The same information about two months of company's work**

| Depart-ment | March | | April | |
|---|---|---|---|---|
| | Average salary, £ | Number of workers | Wage fund, £ | Average salary, £ |
| No. 1 | 265 | 10 | 5 600 | 280 |
| No. 2 | 510 | 18 | 5 100 | 510 |
| No. 3 | 390 | 22 | 7 055 | 415 |

Our task is **to define** the average salary within the company in March and in April.

You have some information about March. This information includes: the average salary of workers of each department and the number of workers in each department.

**Thus,** the average salary of a worker will be calculated as:

**wage fund / number of workers**.

To solve this problem**,** you must build a ***system of reasoning*****.**

This is a classical scheme for solving such problems.

The data about the wage fund in March is absent, but you can get this value multiplying the average salary of workers of each department by the number of workers in this department.

$$\frac{\text{wage fund}}{\text{number of workers}} = \frac{\text{average salary of workers of each department} \cdot \text{number of workers in this department}}{\text{number of workers in this department}},$$

or $\overline{X} = \frac{\sum X_i \cdot f_i}{\sum f_i}$.

The system of economic considerations will suggest the choice for formula calculating the average value.

Then, the average salary will be defined as **408.2** £.

You have information about April. This information includes: the average salary of workers in each department and the wage fund for each department.

Thus, the average salary of workers will be calculated as the wage fund divided into the number of workers.

You have the data about the wage fund in April, but we do not have the data about the number of workers. You can get this data as a ratio of the wage fund of a given department to the average salary of its workers.

$$\frac{\text{wage fund}}{\text{number of workers}} = \frac{\text{wage fund of each department}}{\frac{\text{wage fund of each department}}{\text{average salary of workers of each department}}},$$

or $\overline{X} = \frac{\sum M_i}{\sum \frac{M_i}{X_i}}$.

Then, the average salary will be defined as **377.8** £.

The system of considerations in the first case suggests that in this case we must use the formula of the **arithmetic average value,** in the second case we must use the formula of the **harmonic average** value.

According to the **time of processing information** there may be the following classification of values:

moment,

*e.g.* The population as on the 1 January 2011 is a **moment indicator**

interval,

*e.g.* The amount of a family income per month is an **interval indicator.**

According to the **method of calculation** there may be the following classification of values:

primary,

*e.g.* The amount of sales and the price per unit are **primary** indicators;

secondary,

*e.g.* The revenues from the sales is a **secondary** value (or indicator).

## Applying the concepts

1. Information about the country's external debt in conventional units (c.u.).

| Organizations-lenders | Period | |
|---|---|---|
| | Basic | current |
| **International financial organizations:** | 4 473 | 4 930 |
| **World Bank** | 1 586 | 2 019 |
| **IMF** | 2 790 | 2 800 |
| **European Bank** | 97 | 111 |

**Define** all the possible types of relative values.

2. A construction company has built three houses.

| Buildings | Total area, sq m | Living area, % | Average living space per inhabitant, sq m / person | Average market price of 1 square m of the total area, € / sq m |
|---|---|---|---|---|
| 1 | 1 975 | 75 | 40.5 | 380 |
| 2 | 2 000 | 70 | 30.0 | 360 |
| 3 | 1 870 | 65 | 50.0 | 400 |

**Calculate** averages values, which make sense, according to the given information.

3. We have some information about the sale of goods A in the markets of Kharkov.

| Market | The price of a unit, c.u. | The sales amount, thousand t | Revenues from sales, thousand c.u. | Areas served by the markets, are home for: | |
|---|---|---|---|---|---|
| | | | | population, thousand people | household, thousand |
| **X** | 38.0 | 2.4 | 91.2 | 12.6 | 4.5 |
| **Y** | 34.0 | 4.1 | 139.4 | 12.56 | 5.8 |
| **Z** | 35.0 | 3.9 | 136.5 | 20.46 | 6.2 |

**Calculate** the average price of the goods A in Kharkov as a whole.

**Make** an analytical report on the situation with the prices of goods A.

# Topic 5. Analysis of Distribution Series

- Characteristics of the distribution center
- Quintiles of distribution
- Measurement of variation
- Characteristics of the forms of distribution

## *5.1. Characteristics of the distribution center*

**A distribution series** is an ordered distribution of population units into groups according to a certain sign.

**A distribution** s**eries** characterizes the structure of a population according to a certain sign.

**Elements** of distribution series are:

**variant** *(xi)*, that is the value of a grouping sign which can vary;

**frequency** *(fi)* which shows how often individual values of a sign are repeated.

The values of the **variant** and the frequency determine the regularity of distribution of signs.

**The kinds of** distribution series are:

attributive distribution series;

variational distribution series;

combinational distribution series.

The **attributive** distribution **series** are distributed according to a qualitative sign (or a categorical variable).

The **variational** distribution **series are** distributed according to a quantitative sign (or a numerical variable).

The **combinational** distribution **series** are distributed according to both qualitative and quantitative signs (or both categorical and numerical variables).

In business the customers using statistical data are asking questions about numerical variables.

When summarizing numerical variables, you need to do more than just prepare the tables and charts.

The three **major properties** which describe a set of numerical data are:

**central tendency, variation, shape.**

The **central tendency** is the extent to which all the values group around a typical or central value.

The **variation** is the **value** of dispersion or scattering of values away from a central value.

The **shape** is the pattern of the distribution of values from the lowest value to the highest value.

**The** variation series **is the base** for these calculations.

## Measures of central tendency

In any analysis and/or interpretation, a variety of descriptive measures represent the properties of **central tendency, variation and shape. This variety** may be used to extract and summarize the salient features of a data set.

If these descriptive summary measures are computed from a sample of data, they are called **statistics. I**f they are computed from an entire population of data, they are called **parameters.**

**Statistics usually takes** samples rather than use the entire population.

Most sets of data show a distinct **tendency** to group around a central point. Thus, for any particular set of data, it usually becomes possible to select some typical value to describe the entire set.

When people talk about an "average value" or the "middle value" or the "most frequent value", they are talking informally about the mean, median and mode – three measures of central tendency.

Each one is calculated differently, but each one is a specific number resting near the middle of a larger group of data.

The mean (population or sample) is the balance point for data, so using the mean as a measure of centre generally makes sense. However, the mean is potential disadvantage: the mean can be affected by **extreme values.** There are many instances in business when this may occur. For example, in a population or sample of income data, there may be extremes on the high end that pull the mean upward from the center.

The **mode** and the **median in** a ***discrete*** **distribution series.**

The **median** is the middle value in a set of data that has been ranked from smallest to largest. Half the values are smaller than or equal to the median. Half the values are larger than or equal to the median.

To calculate the **median** for a set of data, you first rank the values from smallest to largest.

Median = n + 1/2 ranked value.

You compute the **median** value by the two following rules:

**Rule 1.** If there is an **odd** number of values in a data set, the median is the middle-ranked value.

**Rule 2.** If there is an **even** number of values in a data set, the median is the average of the middle-ranked values.

*For example*. You have some information about your morning make-up during 10 days (in minutes):

| | |
|---|---|
| Ranked value: | 29, 31, 35, 39, 39, 40, 44, 44, 52; 53. |
| Ranks: | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. |

You have an **even** number of values**.**

In this situation the **median** = **39.5.**

Because the result of dividing n + 1 by 2 is (10 + 1) / 2 = 5.5 for this sample of 10, you must use Rule 2 and the average between the fifth and sixth ranked values, 39 and 40. The median of 39.5 means that for half the days, the time for my make-up is less than or equal to 39.5 minutes, and for half the days, the time for my make-up is greater than or equal to 39.5 minutes.

The **mode** is the value in a set of data that appears most frequently. Like the median and unlike the mean, extreme values do not affect the mode. However, the mode is mostly used for descriptive purposes because it lacks the stability of other measures of central tendency

The mode is used as a measure of central location.

A common mistake is to state the mode as being the frequency of the most frequently occurring value.

If no value occurs more frequently than any other, the data set is said not to have a mode. The mode might be especially useful in describing the central location value for clothes sizes.

*For example*, shoes come in full and half sizes.

Consider the following sample data that have been sorted from low to high:

Shoes 7.5, 8.0, 8.5, 9.0, 9.0, 10.0, 10.0, 10.0, 10.5, 10.5, 11.0, 11.5.
(sizes)

The mean for these sample data is (7.5 + 8.0 + 11.5) / 12 = 9.63.

Although 9.63 is the numerical average, the mode is 10, because more people wore that size shoe than any other. In making a purchasing decision, a shoe shop manager would order more shoes of the modal size than of any other size. The mean does not reflect the purchasing decision.

**The mode and the median are an *interval* of a distribution series**

The **mode** of an ***interval*** of a **distribution series** has the following formula for calculation:

$$m_o = x_{mo} + i_{mo} \cdot \frac{f_{mo} - f_{mo-1}}{(f_{mo} - f_{mo-1}) + (f_{mo} - f_{mo+1})}, \quad (5.1)$$

where $x_{mo}$ is the lower limit of the modal interval;

$i_{mo}$ is the value of the modal interval;

$f_{mo}$ is the modal interval frequency;

$f_{mo-1}$; $f_{mo+1}$ is the frequency of the interval preceding the modal one and the frequency of the interval following the modal one.

***A modal*** interval is the interval which has the highest frequency.

The **median** of an ***interval*** of a **distribution series** has the following formula for calculation:

$$M_e = x_{me} + i_{me} \cdot \frac{\frac{\sum t}{2} - S_{me-1}}{f_{me}}, \quad (5.2)$$

where $x_{me}$ is the lower limit of the median interval;

$i_{me}$ is the value of the median interval;

$S_{me-1}$ is the cumulative frequency that precedes the median interval;

$f_{me}$ is the frequency (relative frequency) of the median interval.

The scheme for calculating the **mode** and the **median** in discrete distribution series is as follows:

*For example*, a sample survey of households in Kharkov was conducted (Table 5.1).

Table 5.1

**Primary data**

| **Number of children in the family** (x) | **0** | **1** | **2** | **3 and more** | **Total** |
|---|---|---|---|---|---|
| **Number of families** (f) | 13 | 21 | 10 | 6 | 50 |
| **Cumulative frequency** (S) | 13 | 34 | 44 | 50 | |

**A mode** is a sign that often occurs in a distribution **series**.

The greatest number of families is 21, so the mode equals to one **child in a family.**

***Conclusion***. A typical family for Kharkov is a family with one child.

To determine the median it is necessary to know its ordinal number in a ranked distribution series.

According to the geometric properties of the median, its number is calculated as:

$\sum f / 2 = 50 / 2 = 25$.

In this situation 25 is not the median value. It is its place in the ranked distribution series.

The first cumulative frequency, which absorbs the number 25 is 34, so the median is equal to one **child.**

***Conclusion***. In Kharkov, on average, 50 % families have up to one child and 50 % families have more than or equal to one child.

The scheme for calculating the **mode** and the **median in the interval** of a distribution series is as follows.

*For example*, you have some information about the distribution of banks' total assets (Table 5.2).

Table 5.2

**Primary data**

| **Total assets of banks, mln $** | **Up to 3** | **3 – 5** | **5 – 7** | **7 and more** | **Total** |
|---|---|---|---|---|---|
| **Number of banks** | 17 | 6 | 23 | 5 | 51 |
| **Cumulative frequency** | 17 | 23 | 46 | 51 | |

A modal interval is the interval, which has the highest frequency of occurrence of a sign.

In this example, the interval from 5 to 7 is the modal interval, because the largest frequency corresponds to it.

Then, **Mo = 5.97 mln $,**

**Me = 5.087 mln $,**

***Conclusion.*** The most typical banks for Kharkov are the ones whose total assets are 5.97 mln $. On average, 50 % banks in the city have assets worth up to 5.087 mln $.

Statistics works with the data that represent only a homogeneous population.

Table 5.3

**The descriptive measures of the center**

| Descriptive measures | Computation method | Data level | Advantages/disadvantages |
|---|---|---|---|
| **Mean** | Sum of values divided by the number of values | Ratio, interval | • numerical center of the data<br>• sum of deviations from the mean is zero<br>• sensitive to extreme values |
| **Median** | Middle value for data that have been sorted | Ratio, interval, ordinal | • not sensitive to extreme values<br>• computed only from the center values<br>• does not use information from all the data |
| **Mode** | Value that occurs most frequently in the data | Ratio, interval, ordinal, nominal | • may not reflect the center<br>• may not exist<br>• might have multiple modes |

There is a **theoretical rule which says** that a collection is considered to be homogeneous when there is approximate equality: $\bar{x} = M_o = M_e$.

In analytical practice the mode and the median are called the **structural averages.**

## *5.2. Quantiles of distribution*

Aside from the measures of central tendency, there are some useful measures of "noncentral" location. They are often used for summarizing or describing the properties of data. The measures are called **quantiles**.

**Quantiles** of the distribution are characteristic values of a sign that divide a population into equal parts:

the **median** divides the population into two parts,

the **quartile** divides the population into four parts,

the **decile** divides the population into ten parts,

the **percentile** divides the population into a hundred parts.

**Quartiles (Q)** split a set of data into four equal parts. The first **quartile** Q1, divides the smaller 25.0 % of the values from the other 75.0 % of the values that are larger. The second **quartile** Q2, is the median – 50 % of the values are smaller than the median and 50 % are larger. The third **quartile** Q3, divides the smaller 75.0 % of the values from the other 25.0 % that are larger.

Use the following **rules** to calculate the quartiles:

**Rule 1.** If the result is a whole number, then the quartile is equal to that ranked value.

*For example*, if the sample size n = 7, the **first quartile** is equal to (7 + 1) / 4 = second ranked value.

**Rule 2.** If the result is a fractional half (2.5, 4.5, etc.), then the quartile is equal to the average of the corresponding ranked value.

*For exampl*e, if the sample size n = 9, the **first quartile** is equal to (9 + 1) / 4 = 2.5 ranked value, halfway between the second ranked value and the third ranked value.

**Rule 3.** If the result is neither a whole number nor a fractional half, you round the result to the nearest integer and select that ranked value.

*For example*, if the sample size n = 10, the **first quartile** is equal to (10 + 1) / 4 = 2.75 ranked value. Round 2.75 to 3 and use the third ranked value.

The first **quartile** and the third **quartile** for **ungrouped data** are as follows**.**

Q1 = (n + 1) / 4 ranked value,

Q3 = 3(n + 1) / 4 ranked value.

The **first quartile** and the **third quartile** for **grouped data** are as follows.

The lower quartile allocates a quarter of the population with the lowest value sign:

$$Q_1 = x_Q + h \cdot \frac{\frac{1}{4}\sum f - S_{Q_1-1}}{f_{Q1}}. \qquad (5.3)$$

The upper quartile allocates a quarter of the population with the larger value sign:

$$Q_3 = x_Q + h \cdot \frac{\frac{3}{4}\sum f - S_{Q_3-1}}{f_{Q1}}, \tag{5.4}$$

where $X_{Q1}$, $X_{Q3}$ are lower limits of the quartile intervals,
$h$ is the quartile interval value,
$f_{Q1}$, $f_{Q3}$ are interval frequencies containing the first and third quartile,
$S_{Q1-1}$, $S_{Q3-1}$ are cumulative frequencies preceding the quartile intervals.

The **decile** divides the population into ten parts.
The **first decile** has its own formula for calculations:

$$D_1 = x_{D_1} + h \cdot \frac{\frac{1}{10}\sum f - S_{D_1-1}}{f_{D_1}}. \tag{5.5}$$

The **ninth decile**:

$$D_9 = x_{D_9} + h \cdot \frac{\frac{9}{10}\sum f - S_{D_9-1}}{f_{D_9}}, \tag{5.6}$$

where $X_{D1}$, $X$ are lower limits of the decile intervals,
$h$ is the decile interval value,
$f_{D1}$, $f_{D9}$ are intervals frequencies containing the first and ninth decile,
$S_{D1-1}$, $S_{D9-1}$ are cumulative frequencies, preceding the decile intervals.

**The percentile** location value has its own formula for calculations:

$$i = (p / 100) \cdot n, \tag{5.7}$$

where $p$ is the desired percent,
$n$ is the number of values in the data set.

If $i$ is not an integer, it must be approximated to the next highest integer. The next integer, larger than $i$ corresponds to the position of the $p^{th}$ percentile in the data set.

If $i$ is an integer, the $p^{th}$ percentile is the average of the values in the position $i$ and position $i + 1$.

The second quartile is the $50^{th}$ percentile and is also the median.

## *5.3 Measurement of variation*

Consider a situation involving two plants of a company. The division vice-president asks two plant managers to record their production output for 5 days. The resulting sample data is shown as follows.

Manufacturing output for Company

| **Plant A** | **Plant B** |
|---|---|
| **15** units, | **23** units, |
| **35** units, | **26** units, |
| **25** units, | **25** units, |
| **20** units, | **24** units, |
| **30** units. | **27** units. |

Instead of reporting these raw data, the managers reported only the mean and median for their data. The following are the computed statistics for the two plants:

**Plant A** and **Plant B** have the following information:

$\bar{x}$ = 25, $\bar{x}$ = 25,

Me = 25. Me = 25.

The division vice-president looked at these statistics and concluded:

1. Average production is the same at both plants.
2. At both plants, the output is at 25 units or more half the time and at 25 units or fewer half the time.
3. Because the mean and median are equal, the distribution of output at the two plants is symmetrical.
4. Based on these statistics, there is no reason to believe that the two plants are different in terms of their production output.

Thus, looking only at the measures of the data's central location can be misleading.

We can detect the **variation**. A set of data exhibits variation if all the data are not the same value.

In international statistics and national statistics **the measures of variation** are: ***range, interquartile range, variance.***

There are several different measures of variation that are used in business decision-making.

The simplest measure of variation is the **range**. The range is a measure of variation that is computed by finding the difference between the maximum and minimum values in a data set.

$$R = X_{max} - X_{min}. \tag{5.8}$$

A measure of variation that helps reduce (decrease) distortion of data because of extreme values is called the **interquartile range.**

The interquartile range is a measure of variation that is determined by computing the difference between the third and first quartiles.

$$IR = Q_3 - Q_1. \tag{5.9}$$

You see, the range is easy to compute and understand. The interquartile range is designed to overcome the range's sensitivity to extreme values. However, neither measure uses all the available data in its computation. Thus, both measures ignore potentially valuable information in data.

Two measures of variation that incorporate all the values in a data set are the variance and the standard deviation. These two measures are closely related.

The **population variance** is the average of the squared distances of the data values from the mean.

For **ungrouped data** you have the following formula of calculation:

$$\sigma^2 = \frac{\sum (x - \bar{x})^2}{n}. \tag{5.10}$$

For **grouped data** we have the following formula of calculation:

$$\sigma^2 = \frac{\sum (x - \bar{x})^2 \cdot f}{\sum f}. \tag{5.11}$$

**A simplified method** of calculation of population variance has the following formula:

$$\sigma^2 = \overline{x^2} - \bar{x}^{\,2}, \qquad (5.12)$$

$$\text{or} \quad \sigma^2 = \frac{\sum x^2 \cdot f}{\sum f} - \left( \frac{\sum x^2 \cdot f}{\sum f} \right)^2.$$

The population variance and dispersion are two synonymic terms.

The **kinds** of **population variance are:**

the total variance,
the within-group variance,
the average of the group variance,
the intergroup variance.

The **total variance** characterizes a sign variation under the influence of all the factors affecting the population.

$$\sigma^2 = \frac{\sum_1^n (x - \bar{x})^2 \cdot f_i}{\sum_1^n f_i}. \qquad (5.13)$$

The **within-group variance** describes the variation of the sign influenced by all the factors operating within a particular group.

$$\sigma_i^2 = \frac{\sum_1^n (x - \bar{x}_i)^2 \cdot f_i}{\sum_1^n f_i}. \qquad (5.14)$$

The **average of the group variance** (dispersion) characterizes a random variation that has appeared under the influence of occasional factors. It does not depend on the statement that formed the group basis.

$$\overline{\sigma_i^2} = \frac{\sum_1^n \sigma_i^2 \cdot f_i}{\sum_1^n f_i}. \qquad (5.15)$$

The **intergroup variance** describes the variation of a resulting sign under the influence of a grouping sign.

$$\delta^2 = \frac{\sum_{1}^{n}(x - \bar{x})^2 \cdot f_i}{\sum_{1}^{n} f_i}. \qquad (5.16)$$

The **rule** of summing **population variance** is $\sigma^2 = \overline{\sigma_i^2} + \delta^2$.

The **coefficient of determination** characterizes the degree of influence of a grouping sign on forming the total variance.

$$\eta^2 = \frac{\delta^2}{\sigma^2}. \qquad (5.17)$$

The coefficient of determination reflects the share of variation of the resulting sign under the influence of the factorial sign. If the connection between these signs is absent, the coefficient of determination is equal to zero. The coefficient of determination is equal to one if there is a functional connection.

The degree of connection between the grouping and resulting signs can be calculated as the **empirical correlation ratio.**

$$\eta = \sqrt{\eta^2} = \sqrt{\frac{\delta^2}{\sigma^2}}. \qquad (5.18)$$

If the empirical correlation ratio is equal to zero, then the grouping sign does not affect the formation of a total variation. If this sign is equal to one, then there exists a functional relationship.

The degree of connection between the grouping and resulting signs can be calculated by Cheddok's scale.

| η | 0.1 – 0.3 | 0.3 – 0.5 | 0.5 – 0.7 | 0.7 – 0.9 | 0.9 – 0.99 |
|---|---|---|---|---|---|
| **Relationship** | mild | moderate | considerable | close | very close |

The **standard deviation** is the positive square root of the variance. The **standard deviation** is in the original units (dollars, tons, square meters).

The **standard deviation** and variance offer alternatives to the range for measuring variation in data.

The **standard deviation** for **ungrouped data** and for **grouped data** is:

for **ungrouped data:**

$$\sigma = \sqrt{\frac{\sum_{1}^{n} (x_i - \bar{x})^2}{n}}, \tag{5.19}$$

for **grouped data**

$$\sigma = \sqrt{\frac{\sum_{1}^{n} (x_i - \bar{x})^2 \cdot f_i}{\sum_{1}^{n} f_i}}, \tag{5.20}$$

Unlike the previous measures of variation presented earlier the **coefficient of variation** is a relative measure of variation. It is always expressed as a percentage.

$$V_\sigma = \frac{\sigma}{\bar{x}} \cdot 100. \tag{5.21}$$

The **coefficient of variation** is used to compare the variations of different signs or a sign in different collections.

The coefficient of variation is the criterion of homogeneity when it is less than 33 % ($V\sigma < 0.33$).

The characteristics of the range, interquartile range, variance, and standard deviation are as follows:

- the more the data are spread out or dispersed, the larger the range, interquartile range, variance, and standard deviation are;
- the more the data are concentrated or homogeneous, the smaller the range, interquartile range, variance, and standard deviation;
- if all the values are the same, all the range, interquartile range, variance, and standard deviations equal zero;
- none of the values variation (the range, interquartile range, variance, and standard deviation) can ever be negative.

## *5.4. Characteristics of the forms of distribution*

The general idea of distribution of population units can be reflected in the graphs named a **histogram** and a **polygon.**

**A polygon** is used for visualizing ***discrete*** distribution series**.**

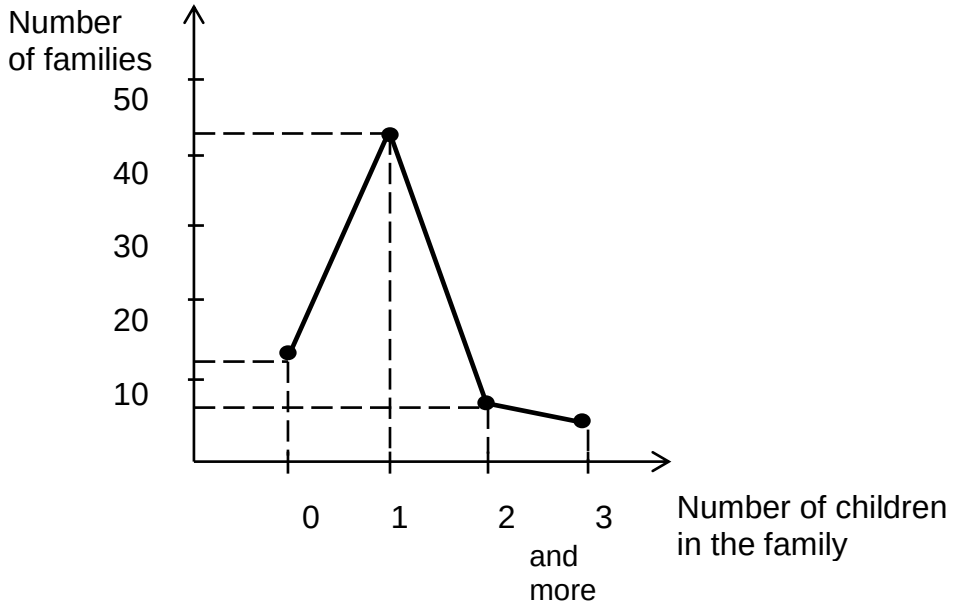



Fig. 5.1. **A polygon**

**A histogram** is used for visualizing ***interval*** distribution series.

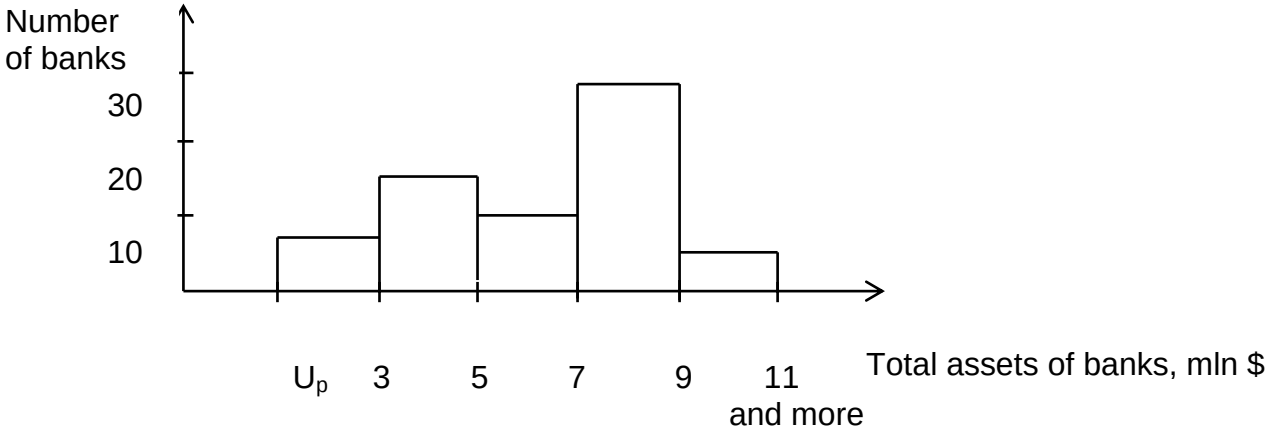



Fig. 5.2. **A histogram**

Frequency distributions are useful for analyzing large sets of data. They are presented in a table format and may not be as visually informative as a graph. If a frequency distribution has been developed from a quantitative variable, a frequency histogram can be constructed directly from the frequency distribution. In many cases, the histogram offers a superior format for transforming the data into useful information.

**A frequency histogram** is a graph of a frequency distribution with the horizontal axis showing a sign (or classes).

The vertical axis shows the frequency. The rectangles have a height equal to the frequency in each sign (class).

A **polygon** is a representation of the shape of particular distribution.

To complete a **polygon,** we connect the first and last midpoints with the horizontal axis so as to enclose the area of the observed distribution.

This situation is correct for the ***interval*** distribution series.

Polygons provide us with a useful visual possibility for comparing two or more sets of data.

A shape is a pattern of the distribution of data values throughout the entire range of all the values.

Distribution is either symmetrical or skewed.

In a symmetrical distribution, the values below the mean are distributed exactly as the values above the mean. In this case, the low and high values balance each other.

In a skewed distribution, the values are not symmetrical around the mean. This skewness results in an imbalance of low values or high values.

The ratio between the mean, median and mode was defined by Karl Pearson:

$$M_e = 1 / 3M_o + 2 / 3\bar{X}. \quad (5.22)$$

The **symmetric** distribution is a data set whose values are evenly spread around the center. For a symmetric distribution, the mean, median and mode are equal.

$$\bar{x} = M_e = M_o. \quad (5.23)$$

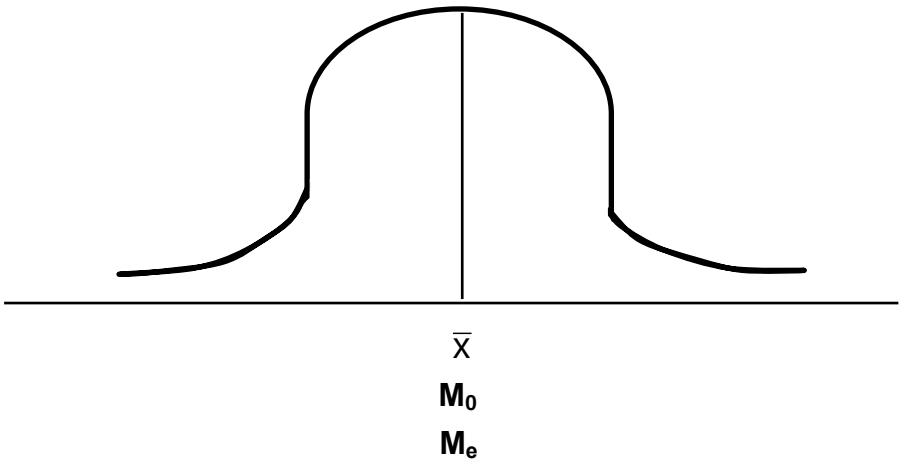



Fig. 5.3. **The symmetric distribution**

The **skewed** distribution is a data set that is not symmetric.

For a skewed distribution, the mean will be larger or smaller than the median and mode.

The **right-skewed** distribution is a data distribution where the mean for the data is larger than the median and mode:

$$\bar{x} > M_e > M_o \qquad (5.24)$$

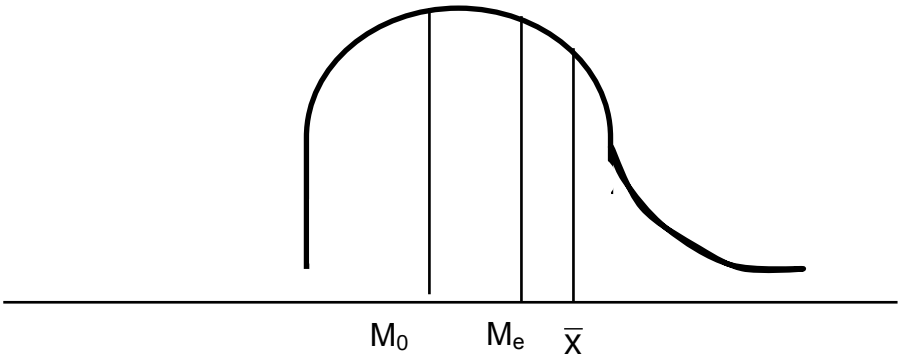



Fig. 5.4. **The right-skewed distribution**

The **left**-**skewed** distribution is a data distribution where the mean for the data is smaller than the median and mode:

$$\bar{x} < M_e < M_o \qquad (5.25)$$

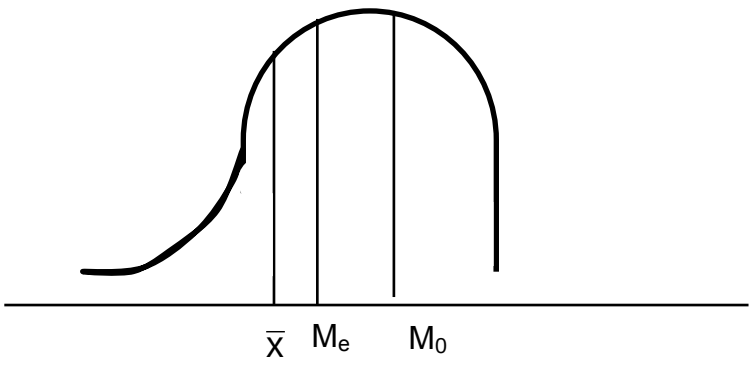



Fig. 5.5. **The left-skewed distribution**

**The central moment of distribution**

For defining an asymmetric distribution and leptokurtic or platykurtic distribution we must know the following formula of the **central moment of distribution**:

$$\mu_k = \frac{\sum_{1}^{n}(x - A)^k \cdot f_i}{\sum_{n}^{n} f_i}. \qquad (5.26)$$

The shape of a distribution is characterized by the following indicators:

**asymmetry coefficient:**

$$A_5 = \frac{\mu_3}{\sigma^3}. \qquad (5.27)$$

The **numerator** of this formula is the moment of the third order. The **denominator** is a standard deviation of the third power.

**excesses coefficient:**

$$E_5 = \frac{\mu_4}{\sigma^4}. \qquad (5.28)$$

The **numerator** of this formula is the moment of the fourth order. The **denominator** is a standard deviation of the fourth power.

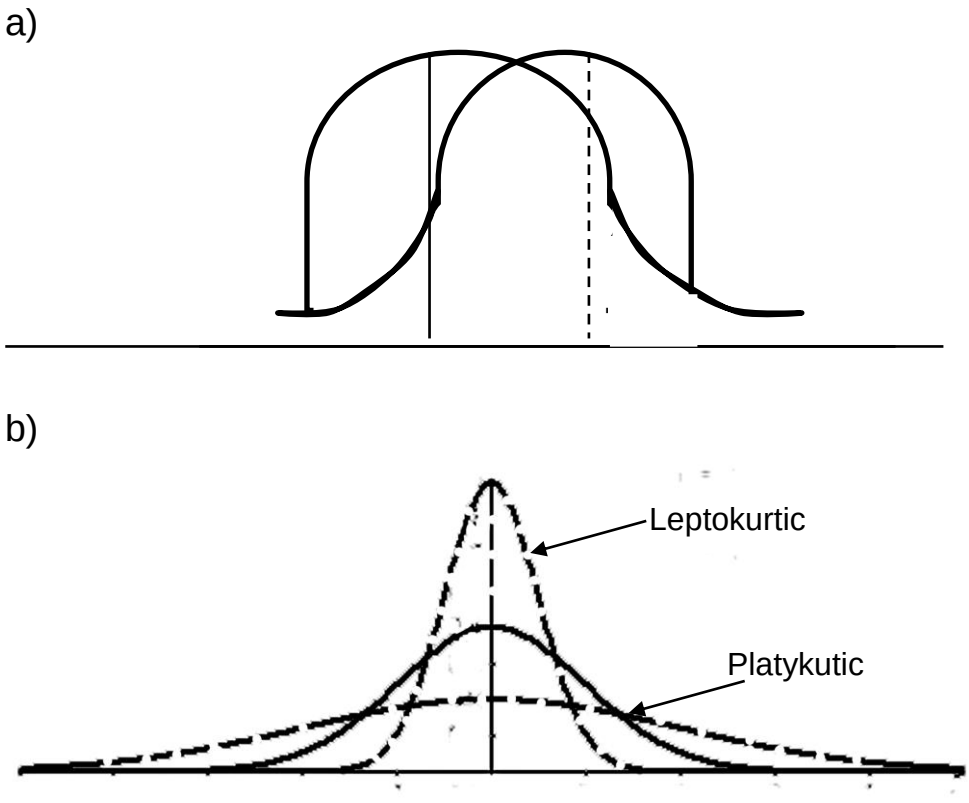



Fig. 5.6. **The asymmetry (a) and excesses (b)**

In symmetrical distribution the variation coefficient must be equal or less than 33 %.

## Applying the concepts

**1.** A car traveled first 320 kilometers of the road at a speed of 97 kilometers per hour. The car traveled the other 180 kilometers of the road at a speed of 105 kilometers per hour.

**Calculate** the average speed of the car during all the road.

**2**. There are 50 employees in a company. 30 of them have 8 working hours per day. 15 of them have 9 working hours per day. 5 of them have 6.5 working hours per day.

**Calculate** the average working hours for on employee in the company.

**3.** The following is a set of data sample of n = 5:

7, 4, 9, 8, 2.

Define the mean, median, and mode.

**4.** The following is a set of data sample of n = 6:

7, 4, 9, 7, 3, 12.

Define the mean, median, and mode.

5. The following is a set of data sample of n = 12:

300, 180, 85, 170, 380, 460.
260, 35, 380, 120 110, 240.

**Compute** the mean, median, first quartile, and third quartile.

6. The marketing director for South East Insurance has been worried about the increasing age of the company's policyholder base. She wants to determine whether a new advertising campaign has had the desired effecting younger customers. She has taken a sample of 10 new policies and has found the following ages:

21, 33, 22, 24, 32, 28, 27, 30, 24, 22.

a) **Compute** the range, interquartile range, and standard deviation for these data.

b) Before the new advertising campaign, the average age of the customers was 36.9. Based on your calculations in part a), has the advertising campaign been effective in reducing the average age of the customers?

7. Suppose that another branch, located in a residential area, is also concerned with the noon-to-1 p.m. lunch hour. The waiting time, in minutes (defined as the time the customer enters the line to when he or she reaches the teller window), of a sample of 14 customers during this hour is recorded over a period of one week. The results are listed below:

| | | | | | | |
|---|---|---|---|---|---|---|
| 7.29, | 10.05, | 5.34, | 9.17, | 6.42, | 4.57, | 5.09. |
| 3.25, | 7.45, | 6.19, | 5.12, | 5.27, | 9.07, | 7.15. |

a) **Calculate** the mean, median, first quartile, and third quartile.

b) **Calculate** the variance, standard deviation, range, interquartile range, and coefficient of variation. Are there any outliers?

c) Are the data skewed? If so, how?

d) As a customer walks into the branch office during the lunch hour, he asks the branch manager how long he can expect to wait. The branch manager replies, "Almost certainly less than 5 minutes". On the basis of the results of a) through c), evaluate the accuracy of this statement.

# Topic 6. Sampling and Sampling Distributions

- The essence of sampling and types the sampling method.
- Evaluating the accuracy of sampled data.
- Defining the necessary sample size.
- Simple regression.

## *6.1. The essence of sampling and types of sampling methods*

The topic "**Sampling and Sampling Distributions**" is considered as a fundamental topic to form inferential statistics.

**Population** and **sample** are two of the most important terms in statistics**.**

A **population** consists of the items or individuals about which you want to draw a conclusion.

A **sample** is the portion of a population selected for analysis.

**Sample observation** is a type of statistical observations. In this observation it is not the entire target population that is investigated. However, only a part of its units, selected in a particular order, is studied.

**Sampling** is a set of mathematical tools and grounds that are used in the sampling observation.

The **theoretical basis** of sampling is the law of large numbers and probability theory.

The scientific principles of the sampling method theory determine the conditions for formation of a sampling population**.**

The **scientific principles** of the sampling method theory are as follows:

- equal opportunities for each unit of the primary (or general) population to fall into the sample,
- a sufficient representation of units in a sample population.

There are three main reasons for selecting the sampling:

- selecting a sample is less time-consuming than selecting every item in the population,
- selecting a sample is less expensive than selecting every item in the population,
- an analysis of a sample is less difficult and more practical than an analysis of the entire population.

The practical use of sampling includes:

- marketing research,
- audit,
- quality control,
- household survey,
- study of public opinion,
- statistical processing of survey results,
- specification of the study,
- checking the results of running the observation.

The sampling process begins by defining the **frame.** A frame is a listing of items that make up the population.

Using different frames to collect data can lead to opposite conclusions.

After selecting a frame, it is necessary to draw a sample from the frame.

*For example*. For studying the population of a country we must know information about their sex, age, place of their residence, occupation etc. This is a **frame.**

If you study the age of the women only, this will be **a sample from the frame.**

There are two **kinds** of samples:

a **nonprobability sample,**

a **probability sample.**

In a nonprobability sample, you select the items or individuals knowing their probabilities of selection. However, their lack of accuracy due to selection bias and the fact, that the result cannot be generalized, reduce this sample's advantages.

Nonprobability samples include:

a judgment sample,

a quota sample,

a chunk sample,

a convenience sample.

Nonprobability samples can have certain advantages, such as convenience, speed, and low cost.

Therefore, you should use **nonprobability sampling** methods only for small-scale studies that precede larger investigations.

In a **probability sample**, you select the items based on the known probabilities. Probability samples allow you to draw an unbiased conclusion about the population of interest. In practice, it is often difficult or impossible to take a **probability sample.**

However, it is necessary to work toward achieving a probability sample and acknowledge any potential biases that might exist.

There are four types of probability samples **which are** most commonly used:

a simple random sample,

a systematic sample,

a stratified sample,

a cluster sample.

These sampling methods vary in their cost, accuracy, and complexity.

In a **simple random sample**, every item from a frame has the same chance of selection as every other item. In addition, every sample of a fixed size has the same chance as every other sample of that size.

Simple random sampling is the most elementary random sampling technique. It forms the base for the other random sampling techniques.

For receiving a sampling method you must use the system of conventional symbols:

Table 6.1

**The system of conventional symbols**

| Indicator | Conventional symbols | |
|---|---|---|
| | **Sample** | **Frame(population)** |
| Size population | N | N |
| Mean | $\tilde{x}$ | $\bar{x}$ |
| Proportion | w | p |
| Alternative proportion | (1 - w) | q |
| Dispersion | w·(1 - w) | p·q |

With a **simple random sample**, it is possible to use *n* to represent the sample size and *N* to represent the frame size. You number every item in the frame from *1* to *N*. The chance that you will select any particular member of the frame on the first selection is 1 / N.

You can select samples with replacement or without replacement.

**Sampling with replacement** means that after you select an item, you return it to the frame, where it has the same probability of being selected again.

However, usually you do not want the same item to be selected again.

**Sampling without replacement** means that once you select an item, you cannot select it again. This process continues until you have selected the desired sample of size *n*.

A **table of random numbers** is a more scientific method of selection.

A table of random numbers consists of a series of digits listed in a randomly generated sequence. Because the numeric system uses 10 digits, the chance that you will randomly generate any particular digit is equal to the probability of generating any other digit.

In a **systematic sample**, you partition the *N* items in the frame into *n* groups of *k* items, where $k = N / n$.

You round *k* to the nearest integer. To select a systematic sample, you choose the first item to be selected at random from the first *k* items in the frame.

If the frame consists of a listing of checks, sales receipts, or invoices, a systematic sample is faster and easier to take than a simple random sample.

To overcome the potential problem of disproportionate representation of specific groups in a sample, you can use either a stratified sample or cluster sampling methods.

Statistical characteristics of sampling are considered as estimates of the characteristics of the frame.

## *6.2. Evaluating the accuracy of the sampled data*

Even when surveys use random probability sampling methods, they are subject to potential errors. A sample is selected because it is simpler, less costly, and more efficient. However, chance dictates which individuals or items will or will not be included in the sample. A sampling error reflects the variation, or "chance differences", from sample to sample, based on the probability of particular individuals or items being selected in the particular samples.

The result obtained from a sample will differ from the results that could be obtained during running observations.

This difference is called a sampling error or an error of representativeness.

The result from the sample with a certain degree of error and with a given probability should be extended to the primary population (or frame).

Point and interval estimates are used in the theory of the sampling method:

*x, σ* are the **point estimate**,

the confidence interval is an interval estimate.

Both point and interval estimates can be calculated for two situations: when *σ* is known and when *σ* is unknown.

**A confidence interval** is such an interval from sample values that if all the possible intervals of a given width were constructed, a percentage of these intervals would include the true population parameter.

The key now is to calculate the upper and lower limits of the interval. The specific method for computing these depends on whether the population standard deviation, *σ* is known or unknown.

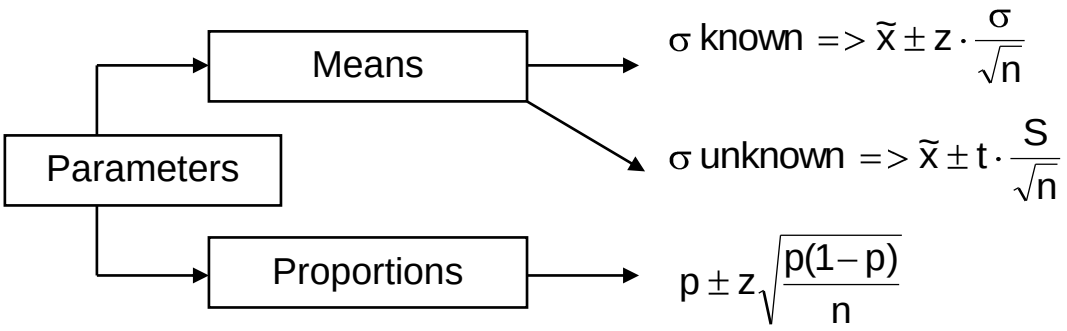



Fig. 6.1. **A flow diagram for confidence interval estimation alternatives**

**A classical scheme** of the confidence interval**:**

**point estimate ± sampling error.**

The new variant of the confidence **interval:**

**point estimate ± (critical value)·(standard error).**

To calculate a **critical value** from a standard normal table for a specified confidence level we must know how to use the standard normal table and understand the properties of the area under a normal curve.

A normal curve is symmetrical. It means that the area under the curve from $\tilde{x}$ to + ∞ ( or - ∞ to $\tilde{x}$**)** makes a half of the area under the whole curve.

This graph demonstrates the part of the areas under the normal curve of the segments.

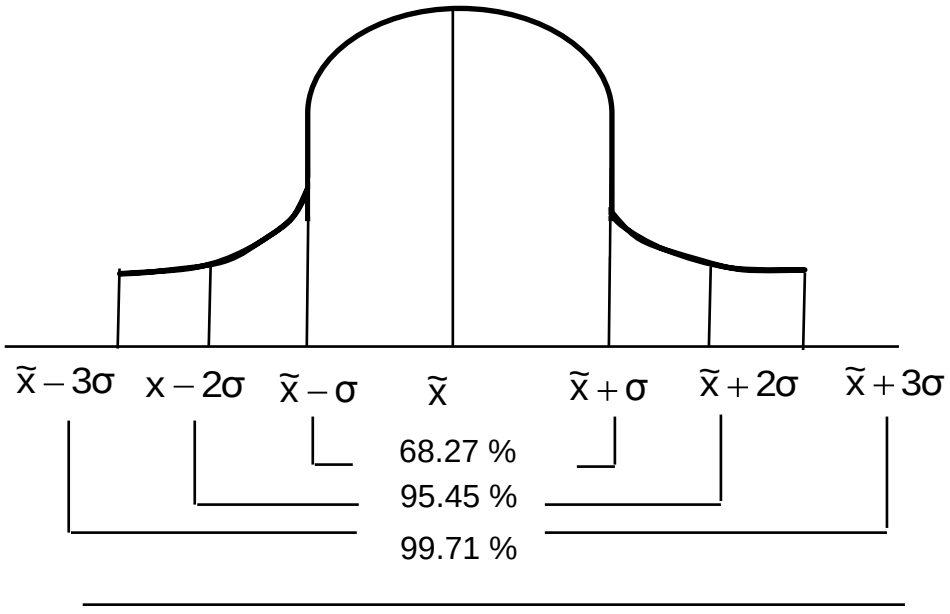



Fig. 6.2. **Approximate areas under the normal curve**

The most common used limits are the following:

| | |
|---|---|
| $\tilde{x} \pm$ 1.645 σ | restricts 90 % of the area under the curve. |
| $\tilde{x} \pm$ 1.96 σ | restricts 95 % of the area under the curve. |
| $\tilde{x} \pm$ 2.575 σ | restricts 99 % of the area under the curve. |

The **critical value** can be found in the distribution table by **Student's *t*-distributions.**

The correspondence between the probability and *t*-value is the most common:

| | | | | | |
|---|---|---|---|---|---|
| **P (t)** | 0.683 | 0.950 | 0.954 | 0.990 | 0.997 |
| **t** | 1.00 | 1.96 | 2.00 | 2.58 | 3.00 |

**Point and confidence interval estimates for a population mean**

**A point estimate** is a single statistic, determined from a sample that is used to estimate a corresponding population parameter.

**A sampling error** is the difference between a measure (a statistic) computed from a sample and a corresponding measure (a parameter) computed from the population.

To overcome this problem with point estimates, the most common procedure is to calculate an interval estimate known as a confidence interval.

**Confidence interval estimates for the population mean if *σ* is known**

A standard error is a value which measures the spread of the sample means around the population mean. The standard error is reduced when the sample size is increased.

The formula for a standard error if we have a selection **with replacement** is the following:

$$\mu = \sqrt{\frac{\sigma^2}{n}} \tag{6.1}$$

The formula for a standard error if we have the selection **without replacement** is the following:

$$\mu = \sqrt{\frac{\sigma^2}{n} \cdot (1 - \frac{n}{N})} \tag{6.2}$$

The first step in developing a confidence interval estimate is to specify the confidence level that is needed to calculate the critical value.

The confidence level is the percentage of all the possible confidence intervals that will contain the true population parameter.

Let us use the following steps to compute a confidence interval estimate for the population mean when the population standard deviation is assumed known and either the population distribution is normal or the size $n$ is > 30:

1. Calculate the population of interest and select a simple random sample of size $n$.
2. Specify the confidence level.
3. Compute the sample mean.
4. Calculate the standard error of the sampling distribution.
5. Calculate the critical value from the standard normal table.
6. Compute the confidence interval estimate.

When the population standard deviation is known, the sampling distribution of the mean has only one unknown parameter: its mean. This is estimated by $X$. However, when the population standard deviation is unknown, there are two unknown parameters: mean and $\sigma$, which can be estimated by $X$ and $s$, respectively.

However, not knowing the population standard deviation does affect the critical value. Recall that when $\sigma$ is known and the population is normally distributed or the Central limit Theorem applies, the critical value is a $z$-value taken from the standard normal table. But when $\sigma$ is not known, the critical value is a $t$-value taken from a family of distributions called the **Student's *t*-distributions.**

**Student's *t*-distributions** are a family of distributions that is bell-shaped and symmetric like the standard normal distribution but with greater area in the tails. Each distribution in the $t$-family is calculated by its degrees of freedom. As the degrees of freedom increase, the $t$-distributions approach the normal distribution.

**Degrees of freedom** are the number of independent data values available to estimate the population's standard deviation.

If $k$ parameters must be estimated before the population's standard deviation can be calculated from a sample of size $n$, the degrees of freedom are equal to $n$-$k$.

**Confidence interval estimates for the population mean if $\sigma$ is unknown.**

For modelling a **confidence interval** we must use the system of conventional symbols:

$\tilde{x}$ is a sample mean,

*t* is a critical value from the *t*-distribution with *n*-1 degrees of freedom for area of $\alpha / 2$ in the (left tail) upper tail,

*s* is the sample standard deviation,

*n* is the sample size.

If in international statistics the *t-value* is named a critical *t-value*, in our statistical tradition this value is named a confidence coefficient.

A confidence interval estimates for the population mean can be developed using the following steps:

1. Calculate the population of interest and select a simple random sample of size *n* from the population.
2. Specify the confidence level.
3. Compute the sample mean and the sample standard deviation.
4. Determine the standard error of the sampling distribution.
5. Determine the critical value for the desired level of confidence.
6. Compute the confidence interval estimate.

## Estimating a population proportion

The previous sections have illustrated the methods for developing c**onfidence interval estimates** when the population value of interest is mean. However, you will encounter many situations in which the value of interest is the proportion of items in the population that possesses a particular attribute.

### Confidence interval estimations for a proportion

The unknown population proportion is represented by the letter *p*. The point estimate for *p* is the sample proportion, $w = m/n$, where *n* is the sample size and *m* is the number of items in the sample having the characteristic of interest.

For modelling a **confidence interval** we must use the system of conventional symbols:

*W* is a sample proportion;

*p* is a population proportion;

*z* is a critical value from the standardized normal distribution;

*n* is the sample size.

The **standard** sampling **error:**

- for a sample proportion for samples **with replacement**:

$$\mu = \sqrt{\frac{p \cdot q}{n}} = \sqrt{\frac{W(1-W)}{n}}; \tag{6.3}$$

- for a sample proportion for samples **without replacement**:

$$\mu = \sqrt{\frac{p \cdot q}{n} \cdot (1-\frac{n}{N})} = \sqrt{\frac{W(1-W)}{n} \cdot (1-\frac{n}{N})}. \tag{6.4}$$

The **sampling error:** $\Delta_x$ **= t·μ.** (6.5)

Some statisticians refer to the sampling error $\Delta_x$ as the "**margin of error**".
This formula has the following variant for:

- the **sampling with replacement:**

$$t \cdot \sqrt{\frac{\sigma^2}{n}};$$

- **the sampling without replacement:**

$$t \cdot \sqrt{\frac{\sigma^2}{n} \cdot (1-\frac{n}{N})}.$$

Sometimes we must compute the **sample estimate for a proportion.**
The needed formula looks as follows $\Delta_p$ **= t·μ**
The **sampling error for a proportion:**

- for the **sampling with replacement:**

$$t \cdot \sqrt{\frac{W(1-W)}{n}}; \tag{6.6}$$

- for **the sampling without replacement:**

$$t \cdot \sqrt{\frac{W(1-W)}{n} \cdot (1-\frac{n}{N})}. \tag{6.7}$$

The range of values of the population, calculated by the sample data with a certain probability has the following formula:

- for the mean:

$$\tilde{x} - \Delta_x \le \bar{x} \le \tilde{x} + \Delta_x; \qquad (6.8)$$

- for the sample proportion:

$$w - \Delta_p \le P \le w + \Delta_p. \qquad (6.9)$$

*For example:* There is a survey conducted in order to monitor a company employees' qualification. 10 % of the employees were subjected to the survey.

Table 6.2

**Primary data**

| Skill-category | I | II | III | IY | Y | YI |
|---|---|---|---|---|---|---|
| Number of employees | 10 | 36 | 88 | 95 | 50 | 21 |

With the probability 0.997 calculate the possible variation margin of the average skill-category of the company's employees.

In order to receive the correct answer, we must calculate the margin of the error.

If there is an opportunity to receive a number value of the frame according to the available data we must use the sampling without replacement.

Then, for calculating the sample margin of the error we must use the following formula:

$$t \cdot \sqrt{\frac{\sigma^2}{n} \cdot (1 - \frac{n}{N})}.$$

In this formula *n* is the sample size. In our example, the total number of employees is the sample size. The total number of employees is equal to the following:

*n* is 10 % – this is 300 people,

*N* is 100 % – this is 3 000 people.

In this situation the frame is equal to 3 000 people.

The sample mean is equal to 3.67 skill-category.

The population variance (dispersion) is equal to 1.41.

In this situation we use the critical *t*-value, because we must calculate the $\sigma$.

Then $\Delta_x = t \cdot \mu = 0.2$ skill-category.

The confidence interval will be equal to $3.67 - 0.2 \leq \bar{X} \geq 3.67 + 0.2$.

***Conclusion***. In 997 cases out of a 1 000 (or with the probability of 99.7 %) we can claim that the average skill-category for all the employees of the company will be within the margin of 3.47 to 3.87.

With the use of the data received before, let us calculate possible variation margins for the proportion of employees who have the $5^{th}$ and $6^{th}$ skill-category with the probability of 0.954.

The proportion of employees in our example will be calculated as $w = m / n = 0.237$ or 23.7 %.

Thus, 23.7 % of the selected employees have the highest skill-categories.

Then, the sample estimate for the proportion is equal to 0.05 or 5 %.

$$2 \cdot \sqrt{\frac{0.237(1-0.237)}{300} \cdot (1-\frac{300}{3\,000})}.$$

***Conclusion***. With the probability of 0.954 we can claim that the proportion of employees of the highest skill-category will be within the margin of 18.7 % to 28.7 %.

Thus, a **sampling distribution** is the distribution of the results if you actually selected all the possible samples. The **sampling distribution of the mean** is the distribution of all the possible sample means if you select all the possible samples of a certain size. If you select all the possible samples of a certain size, the distribution of all the possible sample proportions is referred to as **sampling distribution of the proportion**.

## *6.3. Determining the sample size*

In the business world, a sample size is determined prior to data collection to ensure that the confidence interval is narrow enough to be useful in making decisions.

**Sampling size determination for the mean:**

• for the simple random sample it is equal to:

$$n = \frac{t^2\sigma^2}{\Delta x^2}; \tag{6.10}$$

• for the systematic sample:

$$n = \frac{t^2\sigma^2 N}{\Delta x N + t^2\sigma^2}; \tag{6.11}$$

• for the cluster samples:

$$n = \frac{t^2\delta_c{}^2 N_c}{\Delta^2 x N_c + t\cdot\delta_c^2}; \tag{6.12}$$

• for the stratified sample:

$$n = \frac{t^2\bar{\sigma}^2 N}{\Delta_x^2\cdot N + t\cdot\overline{\sigma^2}}. \tag{6.13}$$

S**ampling size determination for the sample proportion**:

• for the simple random sample it is equal to:

$$n = \frac{t^2 W(1-W)}{\Delta_p^2}; \tag{6.14}$$

• for the systematic sample:

$$n = \frac{t^2 W(1-W)}{\Delta_p^2\cdot N + t^2 W(1-W)}; \tag{6.15}$$

• for the cluster samples:

$$n = \frac{t^2 W_c(1-W_c)\cdot N_c}{\Delta_p^2\cdot N_c + t^2 W_c(1-W_c)}; \tag{6.16}$$

• for the stratified sample:

$$n = \frac{t^2 \overline{W \cdot (1-W)} \cdot N}{\Delta_p^2 \cdot N + t^2 \overline{W(1-W)}}. \qquad (6.17)$$

For a given sample size, a high confidence level will tend to generate a large margin of error. For a given confidence level, a small sample size will result in an increased margin of error. Reducing the margin of error requires either reducing the confidence level or increasing the sample size or both.

## 6.4. *Simple regression*

In regression analysis, the variable you wish to predict is called the dependent variable.

The variables used to make the prediction are called independent variables. In addition to predicting values of the dependent variable, regression analysis also allows you to identify the type of mathematical relationship that exists between a dependent and an independent variable, to quantify the effect that changes in the independent variable have on the dependent variable, and to identify unusual observations.

*For example,* relationship between the stores size in square feet and its annual sales.

The nature of the relationship between two variables can take many forms, ranging from simple to extremely complicated mathematical functions. The simplest relationship consists of a straight-line or **linear relationship** (Fig. 6.3).

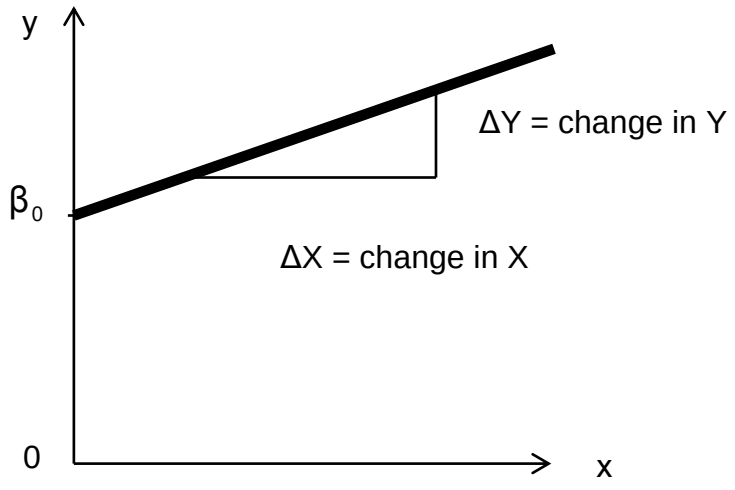



Fig. 6.3. **A positive straight-line relationship**

The straight-line (linear) model can be represented as:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad (6.18)$$

where $\beta_0$ is $Y$ intercept for the population,

$\beta_1$ is the slope for the population,

$\varepsilon_i$ is a random error in $Y$ for observation $i$,

$Y_i$ is a dependent variable (sometimes referred to as the response variable) for observation $i$,

$X_i$ is a independent variable (sometimes referred to as the explanatory variable) for observation $i$.

In this model the portion $Y_i = \beta_0 + \beta_1 X_i$ is a straight line. The slope of the line $(\beta_1)$ represents the unit change in $Y$ per unit change in $X$. It represents the mean amount that $Y$ changes (either positively or negatively) for a one-unit change in $X$. The $Y$ intercept, $\beta_0$, represents the mean value of $Y$ when $X$ equals 0. The $\varepsilon_i$ represents the random error in $Y$ for each observation $i$ that occurs or $\varepsilon_i$ is the vertical distance of the actual value of $Y_i$ above or below the predicted value of $Y_i$ on the line. This term is included because the statistical model is only an approximation of the exact relationship between the two variables.

The selection of the proper mathematical model depends on the distribution of the $X$ and $Y$ values on the scatter plot.

Since we do not have access to the entire population, we cannot compute the parameters $\beta_0$ and $\beta_1$ and obtain the population regression model. The objective then becomes one of obtaining estimates $b_0$ (for $\beta_0$) and $b_1$ (for $\beta_1$) from the sample.

The sample regression equation for representing a straight-line (linear) model is as follows

$$\hat{Y}_i = b_0 + b_1 X_i, \quad (6.19)$$

where $\hat{Y}_i$ is the predicted value of $Y$ for observation $i$,

$X_i$ is the value of $X$ for observation $i$,

$b_0$ is the sample $Y$ intercept,

$b_1$ is the sample slope.

To obtain the statistics $b_0$ and $b_1$ which do minimize the sum of the squared errors in a sample, we use a technique known as the method of least squares. Derive using the sample slope $b_1$ and the sample intercept $b_0$ as follows:

$$b_1 = \frac{\sum(x - \bar{x}) \cdot (y - \bar{y})}{\sum(x - \bar{x})^2}, \qquad \text{or}$$

$$b_1 = \frac{\sum x \cdot y - \frac{\sum x \cdot \sum y}{n}}{x^2 - \frac{(\sum x)^2}{n}}, \tag{6.20}$$

$$b_0 = \bar{Y} - b_1 \bar{X}, \tag{6.21}$$

where $\bar{Y} = \sum Y / n$ and $\bar{X} = \sum X / n$.

The denominator in (6.20) should look familiar: it was used in computing the sum of squares in the analysis of the variance. The numerator is similar, except that both *X* and *Y* values are used at the same time; it is called the sum of cross products. If we abbreviate these quantities as $SS_x$ and $SS_{xy}$, respectively, we have a simplified formula

$$b_1 = SS_{xy} / SS_{x.} \tag{6.22}$$

*For example.* Suppose the placement office at a university wishes to investigate the relationship between the achieved grade-point index and the starting salary of recent graduates majoring in business so that when advising students one may build a model to predict a starting salary of business majors based on the grade-point index (GPI). A random sample of 30 recent graduates is drawn, and the data pertaining to the GPI and the starting salary are recorded for each individual as in Table 6.3.

Calculate a model that best represents the underlying form of the relationship between the starting salary and the grade-point index.

In a simple linear function where regression analysis is used we attempt to develop a linear model from which the values of a dependent variable can be predicted based on particular values of a single independent pair of observations $(X_1, Y_1), (X_2, Y_2) \ldots (X_n, Y_n)$, where $X_i$ represents the $i^{th}$ value of the independent or predictor variable *X* and where $Y_i$ represents the corresponding response – that is, the $i^{th}$ value of the independent variable *Y*.

Table 6.3

**The starting salary and the grade-point index for a random sample of 30 recent graduates majoring in business**

| Individual No. | GPI | Starting salary, £ | Individual No. | GPI | Starting salary, £ |
|---|---|---|---|---|---|
| 1 | 2.7 | 17.0 | 16 | 3.0 | 17.4 |
| 2 | 3.1 | 17.7 | 17 | 2.6 | 17.3 |
| 3 | 3.0 | 18.6 | 18 | 3.3 | 18.1 |
| 4 | 3.3 | 20.5 | 19 | 2.9 | 18.0 |
| 5 | 3.1 | 19.1 | 20 | 2.4 | 16.2 |
| 6 | 2.4 | 16.4 | 21 | 2.8 | 17.5 |
| 7 | 2.9 | 19.3 | 22 | 3.7 | 21.3 |
| 8 | 2.1 | 14.5 | 23 | 3.1 | 17.2 |
| 9 | 2.6 | 15.7 | 24 | 2.8 | 17.0 |
| 10 | 3.2 | 18.6 | 25 | 3.5 | 19.6 |
| 11 | 3.0 | 19.5 | 26 | 2.7 | 16.6 |
| 12 | 2.2 | 15.0 | 27 | 2.6 | 15.0 |
| 13 | 2.8 | 18.0 | 28 | 3.2 | 18.4 |
| 14 | 3.2 | 20.0 | 29 | 2.9 | 17.3 |
| 15 | 2.9 | 19.0 | 30 | 3.0 | 18.5 |

For this example:

the grade-point index is *X*,

the starting salary is *Y*.

Summary computations needed for simple regression and correlation analysis:

$n = 30$ $\quad\quad$ $\sum X = 87.0$

$\bar{X} = 2.90$ $\quad\quad$ $\sum Y = 534.3$

$\bar{Y} = 17.81$ $\quad\quad$ $\sum X^2 = 256.06$

$\sum X Y = 1{,}564.24$ $\quad\quad$ $\sum Y^2 = 9{,}593.41$

Then,

$b_1 = + 3.92819$,

$b_0 = + 6.41825$.

Thus, the equation we want for these data is as follows:

$\hat{Y}_i = b_0 + b_1 X_{i,} = 6.42 + 3.93 X_{i.}$

***Conclusion.*** The slope $b_1$ = 3.93 indicates that for each full unit (point) increase in the grade-point index, the starting salary is predicted to increase by approximately £3.93. The *Y* intercept, $b_0$ = 6.42, represents the predicted value if *X* = 0. As we can see from this case, this predicted value is not necessarily meaningful, since no one could really graduate with a 0.0 grade-point index.

*For example.* An industrial psychologist wishes to predict the time it takes to complete a task based on the level of alcohol consumed during the previous hour. The following data are obtained for a sample of n = 5 students.

Calculate the model of the relationship between the level of alcohol consumed (ounces) and the time to complete the task (minutes).

Predict the time needed to complete the task if alcoholic consumption is as follows: 1.5 ounces and 3.5 ounces.

Table 6.4

**Primary data**

| Number of students | Level of alcohol consumed | Time to complete the task |
|---|---|---|
| 1 | 1 | 2 |
| 2 | 2 | 3 |
| 3 | 3 | 5 |
| 4 | 4 | 7 |
| 5 | 5 | 8 |

The level of alcohol consumed is *X*.

The time to complete the task is *Y*.

Table 6.5

**Calculation data**

| Number of students | X | Y | XY | $X^2$ | $Y^2$ |
|---|---|---|---|---|---|
| 1 | 1 | 2 | 2 | 1 | 4 |
| 2 | 2 | 3 | 6 | 4 | 9 |
| 3 | 3 | 5 | 15 | 9 | 25 |
| 4 | 4 | 7 | 28 | 16 | 49 |
| 5 | 5 | 8 | 40 | 25 | 64 |
| **Total** | **15** | **25** | **91** | **55** | **151** |

Then, $b_1 = 1.5$.

For each additional ounce of alcohol consumed, the time to complete the task increases by 1.5 minutes.

$b_0 = 0.5$.

The constant $b_0 = 0.5$ minutes is the predicted time needed to complete the task when no alcohol is consumed.

Thus, the equation we want for these data is as follows:

$\hat{Y}_i = b_0 + b_1 X_{i,} = 0.5 + 1.5 X_{i.}$

Predicting *Y* when *X* is 1.5 ounces:

$\hat{Y}_i = 0.5 + 1.5 X_i = 0.5 + 1.5 (1.5) = 2.75$ minutes.

Predicting *Y* when *X* is 3.5 ounces:

$\hat{Y}_i = 0.5 + 1.5 X_i = 0.5 + 1.5 (3.5) = 5.75$ minutes.

## Applying the concepts

1. A company wants to select a sample of 32 full-time workers from a population of 800 full-time employees in order to collect information on expenditures concerning a company-sponsored dental plan. How do you select a simple random sample?

A company wants to select a sample of 40 full-time workers from a population of 800 full-time employees. How do you select a systematic sample?

2. Given a population of N = 93, starting in row 29 of the table of random numbers, and reading across the row, select a sample of N = 15

a. with replacement,

b. without replacement.

3. Construct a 95 % confidence interval estimate for the population mean, based on each of the following sets of data, assuming that the population is normally distributed:

Set 1. 1, 1, 1, 1, 8, 8, 8, 8.

Set 2. 1, 2, 3, 4, 5, 6, 7, 8.

Explain why these data sets have different confidence interval even though they have the same mean and range.

4. According to a survey of 100 entrepreneurs (19 %th sample), 20 of them estimate the economic conditions in the region as adverse. With probability of 0.954 **Calculate** the sampling error for the proportion of respondents who are not satisfied with the economic conditions in the region.

5. There is some data on the distribution of companies in terms of production output per year.

| Production output, UAH mln | Up to 8 | 8 – 11 | 11 – 14 | 14 – 17 | 17 and more |
|---|---|---|---|---|---|
| Number of companies | 12 | 15 | 20 | 37 | 1 |

**Check** the degree of homogeneity of the data collected.

**Calculate** the limits of fluctuation (variation) of a factor sign in a frame (primary population) if the data was received with the help of 12 % sample. It must be Calculated with the probability of 0.997.

With the probability of 0.954, Calculate the limits of fluctuations for the proportion of production output which are included into the last group of the interval distribution series. Explain the received results.

6. You plan to conduct a marketing experiment in which students are to taste one of two different brands of soft drink. Their task is to correctly identify the brand tasted. You select a random sample of 200 students and assume that the students have no ability to distinguish between the two brands. (Hint: If an individual has no ability to distinguish between the two soft drinks, then each brand is equally likely to be selected.).

What is the probability that the sample will have between 50 % and 60 % of the identifications correct?

The probability is 90 % that the sample percentage is contained within what symmetrical limits of the population percentage?

What is the probability that the sample percentage of correct identifications is greater than 65 %?

Which is more likely to occur – more than 60 % correct identifications in the sample of 200 or more than 55 % correct identifications in a sample of 1 000? Explain the result.

7. A telephone company wants to estimate the proportion of households that would purchase an additional telephone line if it were made available at a substantially reduced installation cost. A random sample of 500 households would purchase the additional telephone line at a reduced installation cost.

Construct a 99 % confidence interval estimate of the population proportion of households that would purchase the additional telephone line.

How would the manager in charge of promotional programs concerning residential customers use the results in 1st question?

8. You are given the following sample data for variables *X* and *Y*:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| X: | 1, | 7, | 3, | 8, | 11, | 5, | 4. |
| Y: | 16, | 50, | 22, | 59, | 63, | 46, | 43. |

Construct a scatter plot for these data and describe what, if any, relationship appears to exist.

Compute the regression equation based on these sample data and interpret the regression coefficients.

# Topic 7. Time Series Analysis

• Characteristics of the dynamics intensity
• Analysis of development tendencies

## *7.1. Characteristics of the dynamics intensity*

**Time series** is a sequence of values of statistic indicators that characterize the time variation of social-economic phenomena or processes.

A specific value of a dynamic series is named the **series level.** It is marked *(Yt)*, where *t* is time.

The type of time series can be:

moment type,
interval type.

Values of moment and interval types can be represented by absolute values, relative values and average values.

**Time** series study involves the following research:

- tendency of development or trend,
- seasonal variations,
- random variations.

According to the **series level** you can evaluate:

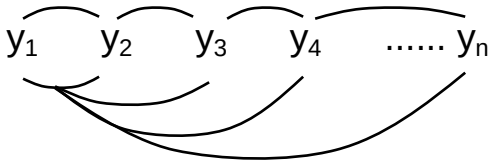
- a trend of dynamics,
- a rate of phenomena change,
- a nature of dynamics.

The comparison of series levels is the methodological basis for evaluating the dynamics intensity.

**Basis** and **chain indicators** are the generalizing characteristics of time series intensity.

The **chain** interconnection of a time series level takes place when each of the following levels is compared to the previous one.

The **basis** interconnection of a time series level takes place when each of the following levels is compared to a definite level, which is referred to as a base for comparison.




**chain** interconnection

**basis** interconnection

The **time** series indicators

1. ***The average series level ($\bar{Y}$)***:

This indicator is calculated for both **moment** time series and **interval** time series.

**for a moment** dynamic series:

- ***with equal intervals of time between dates***:

$$\bar{y} = \frac{\frac{1}{2}y_1 + y_2 + ... + \frac{1}{2}y_n}{n-1}. \qquad (7.1)$$

This formula is named a **chronological average** value formula.

*For example*, there is some information about the circulating assets balance within a company by the 1 date of each month.

| **Date** | **The sum of the balance** |
|---|---|
| 1.01 | 2 € |
| 1.02 | 7 € |
| 1.03 | 3 € |
| 1.04 | 4 € |

Using this data you must calculate the average company's circulating assets balance for the first quarter.

We are offered a moment time series with equal intervals of time between the dates, therefore, to calculate the average circulating assets balance for the first quarter, we must use the chronological average.

***Conclusion****.* The average company's circulating assets balance for the first quarter will be equal to 4.33 €.

Any account balance and remainder can be **transformed** into different contents.

*For example*, today you have some remainder of goods, but tomorrow they can be transformed into the sold products.

The formula of an average series level for a **moment** time series.

- ***with unequal intervals between dates***:

$$\bar{y} = \frac{\sum y \cdot t}{\sum t}, \qquad (7.2)$$

*where* *y* is a series level,

*t* is a time interval between dates.

*For example*, there is some information on the goods delivered to a shop for the current month.

| **The dates of the delivery** | **The cost of the goods** |
|---|---|
| 1.04 | 2 € |
| 11.04 | 7 € |
| 13.04 | 3 € |
| 21.04 | 4 € |
| 29.04 | 4 € |
| 1.05 | 8 € |

Calculate the average cost of goods delivered to the shop for the current month.

You are offered a moment time series with unequal intervals of time between dates. So to calculate the average cost of goods delivered during a month, we have to use a **weighted arithmetic average**.

In this case, there are 6 levels of a time series, but there are 5 time intervals between the dates. We have a problem. The number of time series levels and time intervals between the dates must be equal.

To do this, we calculate the average value between two adjacent series levels.

***Conclusion.*** The average cost of goods delivered to the shop for the current month will be equal to 4.23€.

**for the interval *time*** series**:**

$$\bar{y} = \frac{\sum_{1}^{n} y_i}{n}, \qquad (7.3)$$

where *n* is a number of time series levels.

*For example***.** There is some information on production output in a company per month.

The output is 5 000 € in January, 7 200 € in February, 6 370 € in March.

Calculate the average cost of output in the first quarter.

You are offered an interval time series; therefore, you have to use a simple arithmetic average to calculate the average cost of the output in the first quarter.

$\bar{Y}$ = (5 000 + 7 200 + 6 370) : 3 = 6 190 €.

***Conclusion.*** The average cost of output in the first quarter will be equal to 6 190€.

2. **The absolute increase or decrease** indicator **(Δy):**

- for the **chain** interconnection (**ch.):** $\Delta y_{ch} = y_i - y_{i-1}$, (7.4)
- for the **basis** interconnection(**b.):** $\Delta y_b = y_i - y_0$. (7.5)

***Attention!!! Growth*** in statistics is *ROST* in Russian and ***increase*** is *PRIROST* in Russian.

3. **The growth rate (or decrease rate)** indicator**:**

- for the **chain** interconnection (**ch.):** $T_{ch} = \frac{y_i}{y_{i-1}}$, (7.6)

where $y_i$ is the current ( or under review ) time series level,

$y_{i-1}$ is the previous time series level.

- for the **basis** interconnection (**b.):** $T_b = \frac{y_i}{y_0}$, (7.7)

where $y_0$ is the basis dynamic series level.

4. **The increase rate** has the same formula for both a **chain** interconnection and a **basis** interconnection (ΔT).

$$\Delta T = T_{(ch.\ or\ b.)} - 1\ (100). \qquad (7.8)$$

The formula of increase rate has the following view:

$\Delta T = T_{(ch.\ or\ b.)} - 1$. This formula is used if you calculate the growth rate in coefficients;

$\Delta T = T_{(ch.\ or\ b.)} - 100$. This formula is used if you calculate the growth rate in percent.

5. The absolute value of 1 % increase rate has the following formula:

$$A = 0.01 \cdot Y_{i-1}. \qquad (7.9)$$

The **average absolute** and **relative dynamics** rate is characterized by the following indicators:

- **the average absolute increase (decrease**) indicator :

$$\overline{\Delta} = \frac{y_n - y_0}{n} = \frac{\sum_{i=1}^{n} \Delta y_{ch}}{n}, \qquad (7.10)$$

where $\Delta y_{ch.}$ is the chain absolute increases,

$n$ is the number of chain absolute increases.

- **the average annual growth rate** (decrease rate):

A ***universal formula*** for calculating the average annual growth rate is as follows:

$$\overline{T} = \sqrt[n-1]{\frac{y_n}{y_0}}. \qquad (7.11)$$

This formula is used when:

there are only initial and final values of the time series levels available,

within a dynamic series there is a time gap between the time series levels.

The g**eometric average** value is one of the average value types.

The g**eometric average** value for calculating the average annual growth rate:

$$\overline{T} = \sqrt[n]{T_1 \cdot T_2 \cdot ... \cdot T_n}, \qquad (7.12)$$

where $Ti$ is the chain growth rates,

$n$ the number of chain growth rates.

The expression under the root is a product of chain growth rates.

The power *(n)* reflects the number of the chain growth rates.

The formula for calculating $\overline{T} = \sqrt[n]{T_1 \cdot T_2 \cdot ... \cdot T_n}$ is the most frequently used when there is no time gap between the levels of a dynamic series.

- **the average annual increase rate**:

$$\Delta\overline{T} = \overline{T}_{(ch.\ or\ b.)} - 1\ (100). \qquad (7.13)$$

*For example*. There is some information about the dynamics of a company's goods export to neighboring countries, UAH thou.

Table 7.1

**Primary data**

| **Years** | 2007 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 |
|---|---|---|---|---|---|---|---|
| **Export, UAH thou** | 10 | 100 | 105 | 85 | 90 | 102 | 87 |

The solution to this task is represented as a table. It is recommended to use EXCEL.

Table 7.2

**Calculation data**

| **Years** | **Export, UAH th.** | **Absolute increase, Δy** | | **Growth rate, T** | | **Increase rate, ΔT** | | **Absolute value of 1 % increase rate A, UAH thou** |
|---|---|---|---|---|---|---|---|---|
| | | **ch.** | **b.** | **ch.** | **b.** | **ch.** | **b.** | |
| **2007** | **10** | | | | | | | |
| **2009** | **100** | | | | | | | |
| **2010** | **105** | 5 | 5 | 1.05 | 1.05 | 0.05 | 0.05 | 1.0 |
| **2011** | **85** | -20 | -15 | 0.81 | 0.85 | -0.19 | -0.15 | 1.05 |
| **2012** | **90** | 5 | -10 | 1.06 | 0.90 | 0.06 | -0.10 | 0.85 |
| **2013** | **102** | 12 | 2 | 1.13 | 1.02 | 0.13 | 0.02 | 0.9 |
| **2014** | **87** | -15 | -13 | 0.85 | 0.87 | -0.15 | -0.13 | 1.02 |

Which year must be taken as a base for comparison?

It is not correct to take 2007 as a base for comparison, because the information per two years will be reflected in 2009, and the following years will reflect the change of information per single year.

2009 must be taken as the base for comparison, since this is the first year in the dynamic series, which has no time gap.

The **chain** interconnection of a time series level takes place when each of the following levels is compared to the previous one.

The **basis** interconnection of a time series level takes place when each of the following levels is compared to a definite level, which is referred to as a base for comparison.

**There is a rule:**

When calculating time series indicators it is impossible to compare the past to the present, it is necessary to compare the present to the past.

How can we understand the contents of these indicators?

***Conclusion.*** In 2011 the export cost of the products decreased by UAH 20 thou compared to 2010. And in 2011 the export cost of the products decreased by UAH 15 thou compared to 2009.

In 2013 the export of the products increased 1.13 times (or by 13 % if the growth rate is calculated as percent) compared to 2012. And compared to 2009 the cost of the export products increased 1.02 times or by 2 %.

If in 2010 the company export increased compared to 2009, the increase in export by 1 per cent corresponds to the sum of UAH 1 thousand. If in 2011 the company export decreased compared to 2010, the export reduction by 1 percent corresponds to UAH 1.05 thousand.

**There i**s **a rule:**

- the sum of the chain absolute increase is equal to the last basic absolute increase,
- the product of chain growth rate is equal to the last basis growth rate.

Let us calculate the average annual growth rate for the period from 2007 to 2009. In this situation we use a ***universal formula*** for calculating the average annual growth rate.

***Conclusion***: the average annual growth rate was 3.16 or 316 %, and the increase rate was 216 % over the period under study.

Let us calculate the average annual growth rate for the period from 2007 to 2014. In this situation we use a ***universal formula*** for calculating the average annual growth rate too.

Let us calculate this indicator for the period from 2009 to 2014. In this situation we use a formula of the g**eometric average** value, because there is no time gap between the levels of a time series.

**Estimation of changes in the dynamics rates** can be:

- **Absolute acceleration** (deceleration):

$$A_t = \Delta_{y_i} - \Delta_{y_{i-1}}. \quad (7.14)$$

If $A_t > 0$ is larger then 0, you have its acceleration, if $A_t$ is less then 0, $A_t < 0$ we have its deceleration.

- **Relative acceleration** (deceleration):

$$K_t = \frac{\Delta y_i}{\Delta y_{i-1}}. \quad (7.15)$$

If $K_t > 0$ is larger then 0, you have its acceleration, if it is less then 0, $K_t < 0$ we have its deceleration.

The **absolute indicators** are used to research the dynamics of the same signs that have the same units of measurement.

**Relative indicators** (growth rate, increase rate) are used to compare the dynamics of different signs.

Comparative analysis of the dynamics of interrelated indicators *x* and *y* reflects the **empirical coefficient of elasticity:**

$$K_{el} = \frac{\Delta T_x}{\Delta T_y}. \quad (7.16)$$

## *7.2. Analysis of development tendencies*

There is a **rule** in the ***analytical practice:*** it is impossible to begin analyzing time series, if you have not calculated a general trend of these time series, during the period under study.

Defining the trend (or a general tendency) of a time series is an obligatory condition of its analysis.

**Methods** of time series transformation are as follows:

- moving average,
- analytical equalization,
- seasonality index,
- reduction of time series to a single base,
- joining time series,
- interpolation and extrapolation.

The moving average is the easiest way to transform time series.

*For example.* There is some information on energy consumption by a company for the current 9 months of the year, thousand kW per hour.

Table 7.3

**Primary and calculation data**

| Month | Energy consumption | Calculating moving average |
|---|---|---|
| 1 | 47 | – |
| 2 | 42 | (47 + 42 + 45) : 3 = 44.67 |
| 3 | 45 | (42 + 45 + 48) : 3 = 45.00 |
| 4 | 48 | … |
| 5 | 47 | … |
| 6 | 50 | … |
| 7 | 49 | … |
| 8 | 51 | (49 + 51 + 50) : 3 = 50.00 |
| 9 | 50 | – |

For calculating the ***moving average*** the month data are transformed into quarter data and they gradually calculate the quarterly average value.

This process is named a three-month moving average.

The annual data are mostly transformed into data for 5 years.

This process is named a 5 year moving average.

Thus, you get some of the time series of the transformed values, which should show the trend (decrease or increase).

The disadvantage of this method is the absence of the first and last values of the dynamic series level.

The following method of analysis for development tendencies is the ***method of analytical equalization****.*

Since your school years you have known how to write an equation of a straight line:

$$y = a \cdot x + b. \qquad (7.17)$$

Let us introduce the time parameter into this equation.

To estimate the parameters of the straight line we will use the least squares method.

For the linear trend, the system of normal equations has the following shape:

$$\overline{Y}_t = a_0 + a_1 \cdot t, \qquad (7.18)$$

where $a_0$ and $a_1$ are the parameters of the straight line:

$a_0 = \sum y/n$ is the initial level of a trend,

$a_1 = \sum y \cdot t / \sum t^2$ is the average absolute increase per time unit.

*For example*. There is some information about the income tax value from the salary of a company employee.

Observation of the initial values of dynamic series levels allows us to notice linear dependence.

So that we could calculate the trend and rate of the change for initial values of dynamic series, let us equalize the values.

The solution for this task is represented as a table. It is recommended to use EXCEL.

Table 7.4

**Primary and calculation data**

| **Months** | **Income tax, y** | **t** | **$t^2$** | **y·t** | **$y_t = a_0 + a_1 \cdot t$** |
|---|---|---|---|---|---|
| 1 | 5 | -7 | 49 | - 35 | $Y_{(-7)} = 7.33 + 0.24 \cdot (-7) = 5.65$ |
| 2 | 8 | -5 | 25 | - 40 | 6.13 |
| 3 | 7 | -3 | 9 | - 21 | 6.61 |
| 4 | 6 | -1 | 1 | - 6 | 7.09 |
| 5 | 8 | 0 | 0 | 0 | 7.33 |
| 6 | 5 | 1 | 1 | 5 | 7.57 |
| 7 | 10 | 3 | 9 | 30 | 8.05 |
| 8 | 11 | 5 | 25 | 55 | 8.53 |
| 9 | 6 | 7 | 49 | 42 | 9.01 |
| Total | **66** | **0** | **168** | **40** | **65.97 ≈ 66** |

Since the parameter *t* is conditional, then the sum of *t* must be equal to 0 ($\Sigma t = 0$).

To get $\Sigma t = 0$ it is necessary to use the rule of the calculated zero reading.

**There is a rule:**

If the number of dynamic series levels is odd, we must find its middle and mark it as *0*. And the other values on both sides are marked as follows:

the first dynamic series level down is marked as *1*, the second dynamic series level down is marked as *3*, the third dynamic series level down is marked as *5* and so on;

the first dynamic series level up is marked as *-1*, the second dynamic series level up is marked as *-3*, the third dynamic series level up is marked as *-5* and so on.

If the number of dynamic series levels is even, we don't use "0" but the distribution of marking is similar to the previous sample.

Let us calculate the parameters of the straight line for our situation:

$a_0$ is 66 / 9 = 7.33,

$a_1$ is 40 / 168 = 0.24.

Thus, the equation of a linear trend is the following: $y_t = a_0 + a_1 \cdot t = 7.33 + 0.24 \cdot t$.

Time series levels change at the same rate with a linear trend.

In our example, the value of the income tax increase is UAH 0.24 thou monthly.

Using the equation of a straight line we can calculate the forecast value of the given indicator for October:

$$Y_{10} = 7.33 + 0.24 \cdot 9 = 7.33 + 2.16 = \text{UAH } 9.49 \text{ thou.}$$

If there is some information on activities that depend on seasonal works (agriculture, transport, trade and so on) it is necessary to take into consideration **seasonal influence** on the indicator, which is studied.

*For example.* The rail freight turnover of the South railway is characterized by the following data, thousands of ton-kilometers (Table 7.5).

Table 7.5

**Primary and calculation data**

| The value rail freight turnover per quarter | Years | | | $\overline{y_i}$ | $I_s$ |
|---|---|---|---|---|---|
| | 2012 | 2013 | 2014 | | |
| 1 | 5 | 8 | 7 | 6.7 | 0.421 |
| 2 | 17 | 15 | 18 | 16.7 | 1.049 |
| 3 | 21 | 40 | 30 | 30.3 | 1.903 |
| 4 | 11 | 7 | 12 | 10.0 | 0.628 |
| **Total** | 54 | 70 | 67 | 63.7 | |

Calculate the seasonal change of the rail freight turnover for the South railway.

The seasonal change must be calculated at least for three periods.

The average freight turnover in the first quarter for three years is 6.7 thou ton/km (thousand ton-kilometers).

The overall for the period under study was 15.92 ≈ 16 thou ton/km:

$\bar{y}$ = 54 + 70 + 67/3·4 = 6.7 + 16.7 + 30.3 + 10.0/4=15.92 ≈ 16.

In this situation, the seasonality index was calculated as follows:

$$I_s = \overline{y_i} \; / \; \bar{y}. \qquad (7.19)$$

***Conclusion.*** Without taking into account the seasonality, the freight turnover of the railway would be approximately 16 thou ton/km quarterly.

The seasonality index is used for receiving forecast values.

The following method of analysis of a tendency is a method of ***reduction of time series to a single base.***

*For example.* There is some information about the sales volume by a company's employees and their average salary (Table 7.6).

Table 7.6

**Primary and calculating data**

| Months | Average salary,$ | Sales volume, UAH thou | The basis growth rate | | Leading coefficient |
|---|---|---|---|---|---|
| | | | average salary | sales volume | |
| 1 | 554.8 | 23.25 | | | |
| 2 | 584.4 | 23.70 | 1.053 | 1.019 | 1.033 |
| 3 | 598.1 | 25.00 | 1.078 | 1.077 | 1.001 |
| 4 | 600.7 | 25.20 | 1.083 | 1.087 | 1.004 |
| 5 | 601.0 | 25.99 | 1.083 | 1.118 | 1.032 |
| 6 | 610.5 | 28.06 | 1.100 | 1.2207 | 1.097 |

You must control savings or overruns for the salary fund.

Calculate how well the company's employees work if their global task is sales.

You must calculate the basis growth rates for comparison.

In this situation, January is taken as a basis for comparison.

To make the time series comparable it is necessary to calculate a **leading coefficient**.

The leading coefficient = a larger growth rate / a smaller growth rate.

***Conclusion.*** During the given period, there is no clear trend showing that the growth rate of sales volume outpaces the average salary growth rate.

The following method is **a joining time series method.**

*For example.* There is some information about the income tax within a company (Table 7.7).

You must analyze a change of the income tax within the company for all the period.

To join dynamic series it is necessary to calculate the basis growth rates. The change in tax rates took place in 2009.

Table 7.7

**Primary and calculation data**

| **Amount of income taxes (UAH thou) per year** | **2007** | **2008** | **2009** | **2010** | **2011** | **2012** | **2013** | **2014** |
|---|---|---|---|---|---|---|---|---|
| **Before the change in the tax rate** | 100.00 | 82.00 | 75.00 | | | | | |
| **After the change in the tax rate** | | | 180.00 | 200.0 | 197.00 | 186.0 | 195.00 | 202.0 |
| **The basis growth rate** | **135.3** | **109.3** | **100** | | | | | |
| | | | **100** | **111.1** | **109.4** | **103.3** | **105.3** | **112.2** |

To obtain the absolute values of the amount of the income tax for the entire period under study, you can use 2 variants of calculation:

- the first variant does not take into account the change in the tax rate for the entire period under study (correcting coefficient = 0.417);
- the second variant is based on the fact that the new tax rate has functioned for the entire period under study (correcting coefficient = 2.4).

***Conclusion***. For this period the amount of the income tax has had a multiple-valued trend to decrease from 135.3 % to 112.2 %.

***Interpolation*** is defining an unknown dynamic series level within this dynamic series.

***Extrapolation*** is defining an unknown dynamic series level outside this dynamic series (a short-term forecast).

*For example.* There is some information about the income tax within a company (Table 7.8).

Table 7.8

**Primary and calculation data**

| Years | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | Forecast for 2015 |
|---|---|---|---|---|---|---|---|---|---|
| **Amount of income taxes, UAH thou** | 100.00 | 82.00 | | 83.40 | 82.15 | 77.56 | 81.32 | 84.23 | |

Interpolation is used in case of need to find one missing value, and numerical characteristic of this value may not be very accurate.

The amount of the income tax in the company can be calculated as: 82 + 83.4 = 82.7 UAH thou.

Using the extrapolation method, you can model a forecast for the amount of the income tax in the company in 2015.

To do this, you need to know the average annual growth rate of the income tax in the company for the previous period.

The amount of the income tax in the company in 2015 will be equal to the product of the level in 2014 and the average annual growth rate for the previous period.

## Applying the concepts

1. There is some information about labour hours losses, days:

| Years | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 |
|---|---|---|---|---|---|---|
| **Labour hours losses, days** | 0.75 | 1.07 | 0.64 | 0.47 | 0.78 | 0.51 |

**Analyze** the dynamics of labour hours losses.

2. For two years, the employees' salary at a company increased by 25.7 %. The salary increase rate in 2012 was 12 %.

**Calculate** the salary increase rate in 2013 compared to 2012.

3. **Calculate** all the original and calculated indicators that are missing in the table below.

| Years | Production output, tons | Basis characteristics of the dynamics | | |
|---|---|---|---|---|
| | | Absolute increase, tons | Growth rate,% | Increase rate,% |
| **2010** | 600 | | | |
| **2011** | | | | -2 |
| **2012** | | -28 | | |
| **2013** | | | 97 | |
| **2014** | | | | -6 |

4. In 2009 the product output of the company "OOO" was 2 % higher than in 2008.

In 2010 the production output was 105 % compared to 2011.

In 2011 the production output was 1.2 times as high as in 2008.

In 2012 the company's manufactured products were worth $25 thou.

This sum is 10 % higher than the production output in 2011.

In 2013 the production output was worth $26.9 thou, but in 2014 the production output was worth $24.6 thou.

**Calculate** all the missing indicators of the dynamic series levels and indicators characterizing this dynamic of series: the average annual rate of change in the output for the entire period under study.

5. Manuel Gutierrez correctly predicted the increasing need for home health care services due to the country's aging population.

Five year years ago, he started a company offering meal delivery, physical therapy, and minor housekeeping services in the Galveston area. Since that time he has opened offices in seven additional Gulf State cities. Manuel is presently analyzing the revenue data from his first location for the first five years of operation.

a) Plot these data. Based on your visual observations, what time-series components are present in the data?

b) **Determine** the seasonal index for each month.

c) **Comment** on the linear trend model.

d) Use the seasonal index values computed in part b) to provide seasonal adjusted forecasts for each month of the year 2015.

| Months | Revenue ($ 10,000) | | | | |
|---|---|---|---|---|---|
| | 2010 | 2011 | 2012 | 2013 | 2014 |
| 1 | 29 | 43 | 60 | 63 | 78 |
| 2 | 35 | 54 | 60 | 67 | 58 |
| 3 | 45 | 41 | 57 | 68 | 70 |
| 4 | 46 | 45 | 40 | 60 | 63 |
| 5 | 48 | 40 | 53 | 57 | 64 |
| 6 | 57 | 60 | 56 | 49 | 47 |
| 7 | 63 | 53 | 58 | 64 | 80 |
| 8 | 65 | 72 | 70 | 74 | 69 |
| 9 | 57 | 65 | 70 | 53 | 65 |
| 10 | 60 | 62 | 68 | 73 | 79 |
| 11 | 72 | 82 | 80 | 83 | 87 |
| 12 | 75 | 65 | 69 | 72 | 76 |

# Topic 8. The Index Method

• The kinds of indices and the system of composite indices
• Index method of analysis

## *8.1. Kinds of indices and the system of composite indices*

**An index** can be understood as*:*

- a symbol or number that indicates the individual array elements (classification),
- an indicator to compare two states of the same phenomenon,
- a relative value that characterizes the change of a phenomenon in time, space and so on.

For modelling an index we must use the system of conventional symbols:

*q* is the volumetric rate in physical terms (or quantity in units),

*p* is the price per unit,

*pq* or *(Q)* is the cost (turnover)

*z* is the cost per unit,

*zq* is the total cost of production,

*T* is the number of employees (time spent),

*w* is the level of labor productivity.

## Kinds of indices

- According to t**he nature of comparison** indices may be:
  dynamic
  territorial,
  compared to a certain standard.
- According to t**he degree of generalization of population units** we consider:
  an individual (simple) index,
  a composite index which can be:
    general,
    group.
- According to t**he methods of modeling** one can distinguish:
  an aggregate index,
  a weighted average of an individual index,
  an index of the average level dynamic.

## The functions of composite indices

- **A synthetic function** is a generalizing characteristic of a phenomenon change.
- **An analytical function** is evaluation of an effect of separate factors on a received result.

**Individual** (simple) **indices** (i) characterize the ratio of indicator levels for separate population units or homogeneous groups.

**Composite indices** (I) reflect the generalized values change within the entire population.

**Rules** for modeling composite indices

**Index modeling** implies the presence of the **indexed value** and **weight** in its structure**:**

the indexed value is a sign, whose dynamics is studied,

weight is a value which allows us to generalize heterogeneous elements of population.

The choice of the calculation formula for an index depends on:

the goals of the study,

the economic content of the indicator,

the information that is available.

Statistics studies cause-and-effect relations between phenomena and processes.

In the economy, any relationship between phenomena or indicators is both quantitative and qualitative at the same time.

**Quantitative factors** are extensive factors.

*For example:* 12 computers perform more work than 3 ones, 25 people must do more work than 10 people.

Thus, the number of workers, their time spent, the number of pieces of equipment, output in physical units is **quantitative factors**.

**Qualitative factors** are intensity factors (or efficiency).

*For example:* profit, cost per unit, price per unit.

The content of any studied factors determines relations between their attributions to quantitative or qualitative ones.

If the production unit cost is multiplied by the number of output you can get the total cost of production output.

In this situation, the production unit cost is a quality factor, and the amount of production is a quantitative factor because it reflects the structure of its output.

If the production profitability is calculated as the ratio of profit to the production unit cost, the profit indicator can be represented as the product of profitability and production unit costs.

In this situation, profitability is a quality factor, and the unit cost is a quantitative factor.

**The aggregate index is the main form of general indices.**

The **total index of the actual volume** as an aggregate one:

$$I_q = \frac{\sum q_1 p_0}{\sum q_0 p_0}. \qquad (8.1)$$

In this index a volume indicator in physical units changes but a unit price is weight (or a constant value).

**Rule 1.** If the quantitative factor changes, the corresponding quality factor (being a constant value or weight) must be calculated for the base period.

The general **price index** as an aggregative one.

In this index a unit price is indexed changes, but a volume indicator in physical units is weight (or a constant value):

$$I_p = \frac{\sum p_1 q_1}{\sum p_0 q_1}. \quad (8.2)$$

**Rule 2**. If the quality factor changes, the corresponding quantitative factor (being a constant value or weight) must be calculated for the current period.

The general **cost index** (or turnover index):

$$I_{qp} = \frac{\sum p_1 q_1}{\sum p_0 q_0}. \quad (8.3)$$

The **interconnection of indices**:

multiplicative: $I_{qp} = I_q \cdot I_p$ (8.4)

additive: $\Delta qp = \sum q_1 p_1 - \sum q_0 p_0 = \Delta_q + \Delta_p$ (8.5)

The general index implies summing elements it consists of.

It is not enough to get only the calculated values of the indices.

It is necessary to understand the content of the numeric values for the indices.

According to the analytical functions of general indices you have the following interpretation for these indices:

**the index of the actual volume** characterizes the change in the value of products manufactured by a furniture factory during the current period compared to the base period.

**the price index** characterizes the change in the value of products manufactured by the furniture factory during the current period.

**the cost index** characterizes the change in the value of products manufactured by the furniture factory during the current period compared to the base period both due to the change of the number of manufactured units and due to the unit price change.

Thus, the cost index takes into account the influence of both factors.

The **additive interconnection** allows us to obtain the absolute values of changes of the indexed value.

The difference between the numerator and denominator in the aggregative index of the actual volume indicates the absolute change in the cost of the manufactured product caused only by the change of the amount of this product:

$$\Delta q = \sum q_1 p_0 - \sum q_0 p_0 \ . \qquad (8.6)$$

The difference between the numerator and denominator in the aggregative price index indicates the absolute change in the cost of the manufactured product only due to the change in prices per unit of this product:

$$\Delta_p = \sum q_1 p_1 - \sum q1 p_0 . \qquad (8.7)$$

The difference between the numerator and denominator in the aggregative cost index indicates the absolute change in the cost of the manufactured product both due to changes in the amount of the manufactured product and due to the change of price per unit.

Two basic systems of indices are used in international practice.

These systems are named:

The Paasche systems,

The Laspeyres systems.

**The Paasche** system is used if:

there is a structural change in a country economy,

social issues are almost not taken into account.

**Rule 1.** Weight indicators in modeling Paasche system are taken into account during the current period.

*For example*:

$$I_p = \frac{\sum p_1 q_1}{\sum p_0 q_1} .$$

**The Laspeyres** system is used if:

there are no significant structural changes in a country's economy,

social issues are taken into account together with economic ones.

**Rule 2.** Weight indicators in modeling the Laspeyres system are taken into account during the base period.

*For example*:

$$I_p = \frac{\sum p_1 q_0}{\sum p_0 q_0}.$$

For international comparisons it is necessary to use a unified system of indices.

If different countries use different systems of indices it is necessary to make them comparable.

Fisher system is used for this purpose.

The essence of Fischer system is the following:

$$I_p = \sqrt{\frac{\sum p_1 q_1}{\sum p_0 q_1} \cdot \frac{\sum p_1 q_0}{\sum p P_0 q_0}}. \quad (8.8)$$

***An average weighted index***

Let us consider this question in particular examples.

*For example:* A company sells products.

Calculate in both absolute and relative values how the change only in the quantity of the sold goods influences the company's turnover (Table 8.1).

Table 8.1

**Primary data**

| Goods | Turnover in April, thou € | The change in the quantity of the sold goods in May compared to April, % |
|---|---|---|
| A | 50 | -2.7 |
| B | 17 | +15.0 |
| C | 4.5 | -0.6 |

It is necessary to use the symbols to solve this task.

In this task the information for May and April is given. Then, May can be a current period and April can be a base period.

The time change in the quantity of the sold goods of the given range is the individual index of the actual volume.

Thus, if there is a decrease in the sold goods of 2.7 %, the individual index is equal to 0.973. If there is an increase in the number of sold goods by 15.0 %, the individual index is equal to 1.15.

In the task you must calculate how only the change in the quantity of sold goods influences the amount of the company's turnover.

Thus, it is necessary to calculate the total index of the actual volume:

$$I_q = \frac{\sum q_1 p_0}{\sum q_0 p_0}.$$

However, from the task we are aware of only the denominator (turnover in the base period) but the numerator (turnover in the current period according to the base prices) is still unknown.

To calculate the numerator it is necessary to make the following transformations.

Then $q_1$ is calculated as: $i_q \cdot q_0$.

This index has the following formula:

$$I_q = \frac{\sum i_q \cdot p_0 q_0}{\sum p_0 q_0} = 1.016. \qquad (8.9)$$

**Rule.** Mathematical transformations cannot be used with indices.

This index is identical with the aggregate index.

***Conclusion.*** The company's turnover in May increased by 1.6 % only due to the change in the quantity of sold goods.

*For example:* A company sells products.

Calculate in both absolute and relative values how the change only in prices per unit influences the company's turnover (Table 8.2).

Table 8.2

**Primary data**

| Goods | Turnover in May, thou € | The change in price per unit in May compared to April,% |
|---|---|---|
| A | 50 | -2.7 |
| B | 17 | +115.0 |
| C | 4.5 | +0.6 |

The turnover in May is the value for the current period.

This indicator can be marked as $q_1 p_1$.

The change of price per unit in May compared to April is the individual price index.

To calculate how only the change in prices per unit affected the value of the company's turnover, we must calculate the total price index.

The aggregative formula of this index is known to us:

$$I_p = \frac{\sum p_1 q_1}{\sum p_0 q_1}.$$

However, from the task you are aware of only the turnover for the current period.

To calculate this value let us make the following transformations.

$p_0$ is calculated as: $p_1 / i_p$

$$I_p = \frac{\sum p_1 q_1}{\sum \frac{p_1 q_1}{i_q}}. \qquad (8.10)$$

You will have the average harmonic price index which is identical with the aggregate price index.

***Conclusion.*** The company's turnover in May increased by 12.1 % only due to the change of prices per unit.

### *The average level dynamics index*

The average level dynamics is characterized by:

the **index of variable composition**,

the **index of constant composition**,

the **index of the shift of proportions**.

*For example*. You have some information about the sales of product A in the markets of Kharkov (Table 8.3).

Table 8.3

**Primary data**

| Market | 2012 | | 2013 | |
|---|---|---|---|---|
| | The price for unit, € | The sales amount, ton | The price for unit, € | The sales amount, ton |
| **Alekseevskiy** | 3.8 | 2.4 | 6.0 | 2.3 |
| **Konnyy** | 3.4 | 4.1 | 6.2 | 3.8 |
| **Saltovskiy** | 3.5 | 3.9 | 5.8 | 5.0 |

**Calculate** how the average price changed for product A in Kharkov, if you know that this product was only sold in Alekseevskiy, Konnyy and Saltovskiy markets.

The average price will be calculated as the total cost of the product in the three markets divided by the quantity of the sold product in these markets.

Calculate how only the price per unit in each market influences the average price per unit in Kharkov as a whole.

This influence characterizes the **index of constant composition:**

$$I = \frac{\sum x_1 \cdot f_1}{\sum f_1} : \frac{\sum x_0 f_1}{\sum f_1} = \frac{\sum x_1 \cdot d_1}{\sum x_0 \cdot d_1} = 1.695. \quad (8.11)$$

***Conclusion.*** The average price of product A in Kharkov increased by 69.5 % only due to changes in prices per unit in each city market.

Mathematical transformations with indices cannot be used because they can distort the indices' contents.

Calculate how only the quantity of the product in every market influences the average price per unit in Kharkov.

This influence characterizes the **index of shift of proportions**:

$$I = \frac{\sum x_0 f_1}{\sum f_1} : \frac{\sum x_0 \cdot f_0}{\sum f_0} = 0.994. \quad (8.12)$$

***Conclusion.*** The average price of product A in Kharkov decreased by 0.06 % only due to the change of quantity of the product in the city markets.

Calculate how both the price per unit and quantity of the product in each market influences the average price per unit in the city.

This influence characterizes the index of variable composition:

$$I = \frac{\sum x_1 \cdot f_1}{\sum f_1} : \frac{\sum x_0 \cdot f_0}{\sum f_0} = 1.694. \quad (8.13)$$

***Conclusion****.* The average price per unit for product A in Kharkov increased by 69.4 %. This growth takes place both due to increase of price per unit in each market by 69.5 %, and due to the change of quantity of the product sold in the city markets by 0.06 %.

The **interconnection** of indices has the following formula:

The index of variable composition is equal to the product of index of constant composition and the index of the shift proportions.

## *8.2. The index method of analysis*

Labour productivity is equal to the quantity of productis divided by the number of employees.

Then, the amount of output is equal to the quantity of product of labour productivity and the number of employees.

The quantity of productis is the result of the company's performance.

Employees' labour productivity and their number are the factors that influence the quantity of the manufactured products.

If the quantity of the output is equal to labour productivity multiplied by the number of employees, the index of the quantity of output is equal to the labour productivity index multiplied by the number of employees index.

Let us write the index of quantity of output as the interconnection of labour productivity and the number of employees:

$$q = W \cdot T, \tag{8.14}$$

$$I_q = I_w \cdot I_T. \tag{8.15}$$

The index of quantity of output is equal to

$$I_q = \frac{\sum W_1 T_1}{\sum W_0 T_0}. \tag{8.16}$$

***Conclusion****.* The quantity of output increased (or decreased) both due to the change in labour productivity, and due to the changes in the number of employees.

The labor productivity index is equal to

$$I_w = \frac{\sum W_1 T_1}{\sum W_0 T_1}. \tag{8.17}$$

***Conclusion****.* The quantity of output increased (or decreased) only due to the changes in labour productivity of the employees.

**The labor productivity index is equal to**

The index of the number of employees is equal to

$$I_T = \frac{\sum W_1 T_1}{\sum W_0 T_0}. \qquad (8.18)$$

***Conclusion.*** The quantity of output increased (or decreased) only due to the changes of the number of employees.

## Applying the concepts

1. In January sale proceeds of meat were twice as high as sale proceeds of fish in a city market. In February the quantity of the sold meat increased by 15 %, and the quantity of the sold fish increased by 20 %.

**Calculate** the average percent of change in the volume of sales for both kinds of goods.

**Calculate** the absolute change in the turnover due to the change in the quantity of the sold products if we know that in January 24 000 UAH were received for the sold meat.

2. During the current period, the prices of nonfood commodities did not change, but the prices of food commodities grew by 15.7 %.

**Calculate** how the sale proceeds of the goods changed due to the price changes. It is known that during the current period the share of sale proceeds from food commodities was 60 %.

3. During the base period the working day was 7.5 hours and the labour productivity per employee per hour was 10 units. During the current period these indicators increased 1.03 and 1.12 times respectively.

**Calculate** the absolute change in the labour productivity per day only due to the changes in working hours.

4. Problems 16 – 10 through 16 – 13 refer to U.S. Homes, a major developer of housing communities in the New England area. The company

has kept a record of the relative cost of labor and materials in its market areas for the last 11 years. These data are as follows:

**Primary data**

| Year | Hourly wages | Average material cost |
|---|---|---|
| 2004 | 30.20 | 66.500 |
| 2005 | 30.80 | 68.900 |
| 2006 | 31.70 | 70.600 |
| 2007 | 32.50 | 70.900 |
| 2008 | 34.00 | 71.200 |
| 2009 | 34.50 | 71.700 |
| 2010 | 35.10 | 72.500 |
| 2011 | 35.00 | 73.700 |
| 2012 | 34.80 | 73.400 |
| 2013 | 33.80 | 74.100 |
| 2014 | 34.20 | 74.000 |

a) using 2004 as the base year, construct a separate index for each component in the construction of a house,

b) plot both series of data and comment on the trend you see in both plots,

c) construct a Paashe index for 2011 using the data. Use 2004 as the base year and assume that in 2011 sixty percent of the cost of a house was in materials,

d) construct a Laspeyres index using the data, assuming that in 2004, 40 % of the cost of a house was labor.

# Recommended Literature

## Main


Бабешко Л. О. Основы эконометрического моделирования : учеб. пособ. / Л. О. Бабешко. – Изд. 3-е, стереотипное. – М. : КомКнига, 2007. – 432 с.

Бек В. Л. Теорія статистики : навч. посіб. / В. Л. Бек – К. : Центр навчальної літератури, 2002. – 288 с.

Венецкий И. Г. Основные математико-статистические понятия и формулы в экономическом анализе : справочник / И. Г. Венецкий, В. И. Венецкая. – М. : Статистика, 1979. – 447 с.

Гусаров В. М. Теория статистики / В. М. Гусаров. – М. : ЮНИТИ-ДАНА, 2003. – 464 с.

Джессен Р. Методы статистических обследований / Р. Джессен ; пер. с англ. Ю. П. Лукашина, Я. Ш. Паппэ ; под ред. Е. М. Четыркина. – М. : Финансы и статистика, 1985. – 318 с.

Єріна А. М. Статистика : навч.-метод. посібник для самост. вивч. дисц. / А. М.Єріна, Р. М. Моторін, А. В. Головач ; за заг. ред. А. М. Єриної, Р. В.Моторіна. – К. : КНЕУ, 2001. – 448 с.

Єріна А. М. Статистичне моделювання та прогнозування : навч. посіб. / А. М. Єріна. – К. : КНЕУ, 2001. – 170 с.

Єріна А. М. Теорія статистики : практикум / А. М. Єріна, З. О. Пальян. – 5-те вид., стереотип. – К. : Знання, 2005. – 256 с.

Єріна А. М. Економічна статистика : практикум / А. М. Єріна, О. К. Мазуренко, З. О. Кальян. – К. : ТОВ "УВПУ "Екс Об", 2002. – 232 с.

Ефимова М. Р. Общая теория статистики : учебник / М. Р. Ефимова, Е. В. Петрова, В. Н. Румянцев. – 2-е изд., испр. и доп. – М. : ИНФРА-М, 2007. – 416 с.

Кендэл М. Временные ряды / М. Кендэл ; пер. с англ. и предисл. Ю. П. Лукашина. – М. : Финансы и статистика, 1981. – 358 с.

Ковтун Н. В. Загальна теорія статистики : курс лекцій / Н. В. Ковтун, В. С. Столяров. – К. : Четверта хвиля, 1996. – 144 с.

Про державну статистику : Закон України // Голос України. – 1992. – № 43. – С. 1–4.

Berenson Mark L. Basic Business Statistics: Concepts and Applications / Mark L. Berenson, David M. Levine. – 10th ed. – NJ : Prentice Hall, 2006. – 662 p.

Fowler Floyd J. Survey Research Methods / Floyd J. Fowler. – 3rd ed. – Thousand Oaks, CA : Sage Publications, 2001. – P. 26–107.

## Additional

Кормен Т. Алгоритмы. Построение и анализ / Т. Кормен, Ч. Лейзерсон, Р. Ривест. – М. : МЦНМО, 2000. – 960 с.

Лугінін О. Є. Статистика : підручник / О. Є. Лугінін, С. В. Білоусова. – К. : Центр навчальної літератури, 2006. – 580 с.

Лугінін О. Є. Статистика : підручник / О. Є. Лугінін. – 2-е вид., переробл. та доп. – К. : Центр учбової літератури, 2007. – 608 с.

Макарова Н. В. Статистика в Excel : учеб. пособ. / Н. В. Макарова, В. Я. Трофимец. – М. : Финансы и статистика, 2002. – 368 с.

Мармоза А. Т. Практикум з теорії статистики / А. Т. Мармоза. – К. : Ельга, Ніка-Центр, 2003. – 344 с.

Мармоза А. Т. Теорія статистики / А. Т. Мармоза. – К. : Ельга, Ніка-Центр, 2003. – 392 с.

Общая теория статистики: Статистическая методология в изучении коммерческой деятельности : учебник / под. ред. О. Э. Башиной, А. А. Спирина. – 5-е изд., доп. и перераб. – М. : Финансы и статистика, 2000. – 440 с.

Опря А. Т. Статистика (з програмованою формою контролю знань). Математична статистика. Теорія статистики / А. Т. Опря. – К. : Центр навчальної літератури, 2005. – 472 с.

Пасхавер И. С. Средние величины в статистике / И. С. Пасхавер. – М. : Статистика, 1979. – 172 с.

Практикум по статистике / под ред. В. М. Симчеры. – М. : ЗАО "Финстатинформ", 1999. – 260 с.

Практикум по теории статистики / под ред. Р. А. Шмойловой. – М. : Финансы и статистика, 2000. – 416 с.

Статистика / за ред. С. С. Герасименко. – К. : КНЕУ, 1998. – 468 с.

Статистическое моделирование и прогнозирование : учеб. пособ. / под общ. ред. Ю. Г. Королева. – М. : МЭСИ, 2005. – 103 с.

Статистическое моделирование и прогнозирование : учеб. пособ. / под ред. А. Г. Гранберга. – М. : Финансы и статистика, 2000. – 384 с.

Теория статистики / под ред. Р. А. Шмойловой. – 3-е изд., перераб. – М. : Финансы и статистика, 1999. – 464 с.

Фещур Р. В. Статистика: теоретичні засади і прикладні аспекти. / Р. В. Фещур, А. Ф. Барвінський, В. П. Качур. : за заг. ред. Р.В.Фещура. – Львів : Інтелект-Захід, 2003. – 576 с.

Эконометрика : учеб. пособ. / С. А. Бородичю. – Мн. : Новое знание, 2001. – 408 с.

Юзбашев М. М. Статистический анализ тенденций и колеблемости / М. М. Юзбашев, А. М. Манелл. – М. : Финансы и статистика, 1998. – 207 с.

Keller G. Statistics for Management and Economics / G. Keller, B. Warrack. – Fifth edition. – Duxbury, 2000. – 1000 p.

Microsoft Excel 2007. – Redmond, WA : Microsoft Corp., 2007. –376 p.

## Internet resources

Сайт Укрстат. – Режим доступа : www.ukrstat.gov.ua.

Сайт Евростат. – Режим доступа : // www. europa.eu.int/comm/eurostat/.

Statistics Canada copyright statement for teachers [Electronic resource]. – Access mode : http://www.statcan.ca/english/edu/copy.htm.

Сайт Организации экономического сотрудничества и развития (OECD). – Режим доступа : www.oecd.org/std.

# Contents

EDUCATIONAL EDITION

I. **Serova**

# STATISTICS

**Summary of Lectures**
**for full-time students of training directions 6.140103 "Tourism", 6.030601 "Management"**
**of specialization "Business Administration"**

Editorial director **O. Rayevnyeva**

Editor-in-chief **L. Syedova**

Editor **Z. Zobova**
Proof-reader **Z. Zobova**

НАВЧАЛЬНЕ ВИДАННЯ

**Сєрова** Ірина Анатоліївна

# СТАТИСТИКА

**Конспект лекцій**
**для студентів напрямів підготовки 6.140103 "Туризм", 6.030601 "Менеджмент"**
**спеціалізації "Бізнес-адміністрування"**
**денної форми навчання**

**(англ. мовою)**

Відповідальний за випуск **Раєвнєва О. В.**

Відповідальний редактор **Сєдова Л. М.**

Редактор **Зобова З. В.**
Коректор **Зобова З. В.**

План 2014 р. Поз. № 62-К.
Підп. до друку 03.12.2014 р. Формат 60 х 90 1/16. Папір MultiCopy. Друк Riso.
Ум.-друк. арк. 6,25. Обл.-вид. арк. 7,81. Тираж 50 прим. Зам. № 318.

Видавець і виготівник – видавництво ХНЕУ ім. С. Кузнеця, 61166, м. Харків, пр. Леніна, 9-А

*Свідоцтво про внесення до Державного реєстру суб'єктів видавничої справи*
***Дк № 481 від 13.06.2001 р.***