Methods of statistical description

An analysis of socio-economic phenomena in European countries

Anna Gdakowicz • Marta Hozer-Koćmiel
Iwona Markowicz • Urszula Ala-Karvia

Methods of statistical description. An analysis of socio-economic phenomena in European countries

UNIWERSYTET SZCZECIŃSKI
ROZPRAWY I STUDIA T. (MCDXXXVII) 1363

Anna Gdakowicz
Marta Hozer-Koćmiel
Iwona Markowicz
Urszula Ala-Karvia

Methods of statistical description. An analysis of socio-economic phenomena in European countries

Szczecin 2025

Anna Gdakowicz, University of Szczecin, Institute of Economics and Finance
ORCID 0000-0002-4360-3755

Marta Hozer-Koćmiel, University of Szczecin, Institute of Economics and Finance
ORCID 0000-0003-0443-516X

Iwona Markowicz, University of Szczecin, Institute of Economics and Finance
ORCID 0000-0003-1119-0789

Urszula Ala-Karvia, University of Helsinki, Ruralia Institute
ORCID 0000-0002-8327-1310

DOI 10.18276/978-83-7972-916-6
ISBN 978-83-7972-916-6 (online)
ISBN 978-83-7972-915-9 (print)
ISSN 0860-2751

WYDAWNICTWO NAUKOWE UNIWERSYTETU SZCZECIŃSKIEGO
Wydanie I | Ark. wyd. 9,0 | Ark. druk. 13,4 | Format 170/240

Spis treści

Introduction

In modern times, it is impossible to imagine any society, field of science, or economy without access to data on socio-economic phenomena. IT development facilitates both the collection of data and their availability (Paradysz 2012). However, conclusions can only be drawn through analysing these data with the use of quantitative methods. Statistical methods are the backbone of data analysis. Statistics is recognised as a method of extracting information from data. Statistical expertise is a valuable asset in all professions (Domański et al. 2014). As Tadeusiewicz emphasises, people constantly use statistics, especially politicians (Stanisz 2006). Knowledge of statistical methods, a well-functioning system of public statistics, and widespread access to statistical data are important elements of modern democracy (Szreder 2009).

Statistics can be defined as the science of collecting and analysing data (King, Eckersley 2019). The purpose of data analysis is to provide information that reduces risk when making economic decisions (Gatnar, Walesiak 2004). This information is essential for both investors and local authorities wishing to attract investors (Dehnel 2010). Although we now have access to a variety of data (often in huge databases, called big data) and computer programmes that facilitate the investigation of these data, to draw the right conclusions we still need specific knowledge.

When analysing diverse phenomena, it is necessary to choose the right research methodology (Pociecha 2009; Dudek, Krawiec, Landmesser 2011; Kończak, Trzpiot 2018). The relevant literature points to inappropriate applications of quantitative methods in the analysis of specific data (Sokołowski 2004; Wiśniewski 2013). It also highlights both the transfer of methods from other fields of science (Bieszk-Stolorz, Markowicz 2019) and the emergence of new methods (Dudek 2013) used in economic research.

Hence, the methodology for studying socio-economic phenomena has become an evolving and extensive set of research methods, with numerous conditions for their application. The authors' intention in this book is to facilitate the structuring and expanding of the knowledge of methods of statistical description and their use in socio-economic analysis. Statistical data analysis means searching for statistical regularities in the structure of phenomena, their correlation and changes over time (Hozer 2004). The layout of the book is consistent with the known types of statistical regularities and the research methodology used in their identification.

The main authors of the book are scientists working at the Department of Econometrics and Statistics of the Institute of Economics and Finance at the University of Szczecin. They have been researching socio-economic phenomena using quantitative methods for many years (Gdakowicz, Hozer-Koćmiel, Markowicz 2022). So far, the authors' individual research interests have focused on such issues as the real estate market, demography, household time budgeting, corporate sustainability, unemployment, and intra-EU trade in goods.

This book provides an overview of the methods applied in the search for statistical regularities in socio-economic phenomena in European countries. The authors refer to methods for analysing one- and two-dimensional statistical data. They explain how survey methods should be chosen depending on the type of statistical series, their characteristics, and the purpose of the analysis. All descriptive parameters and functions are determined for the sample source data. Moreover, the manner of their interpretation is specified.

The book is the result of cooperation between authors representing the University of Szczecin (Institute of Economics and Finance; Poland) and the University of Helsinki (Ruralia Institute; Finland).

Anna Gdakowicz
Marta Hozer-Koćmiel
Iwona Markowicz
(University of Szczecin)

Urszula Ala-Karvia
(University of Helsinki)

Chapter 1

Basic concepts

1.1. Statistics – definition and outline

The very term "statistics" is used quite widely and has a variety of meanings. It can be equated with the presentation of data on a particular issue,[1] or it can be used to describe activities such as collecting and assembling statistical characteristics. It is also broadly understood as a science, and it is this interpretation of statistics that is adopted in this book and defined below.

Statistics is the science that provides theories and methods for studying mass phenomena (processes). Mass phenomena recur frequently (e.g., births, studies, work, production, purchases) and, when studied on a large scale, show inherent regularities. These regularities cannot be found in a single case. The purpose of statistical research is to detect statistical regularities (Section 1.3), with statistical populations being the object of study (Section 1.2).

It is generally emphasised that statistics has an interdisciplinary character, as it is employed in a wide variety of scientific fields (economics, management, medicine, etc.). Hence, the application of statistical methods requires expertise in both statistics and the relevant field of scientific research.

Statistics is generally divided into descriptive and mathematical statistics. **Descriptive statistics** focuses on studying statistical populations and describing them by means of descriptive measures (e.g., numerical characteristics). The observed

[1] The literature points out that there is a growing demand for statistical analyses, and that the traditional table-graphic communication of 'dry' numerical data is becoming less and less satisfactory for recipients (Młodak 2006). The information provided to the audience needs to be precise, which requires the use of quantitative analysis methods (Panek 2007).

regularities relate to the examined population. The population is described by specific numerical characteristics. This is this part of statistics that this book focuses on.

Mathematical statistics, on the other hand, concentrates on a fraction of a population and, based on probability theory, transfers the results to the whole population. This process is referred to as statistical inference.

1.2. Statistical population, statistical characteristics, and measurement scales

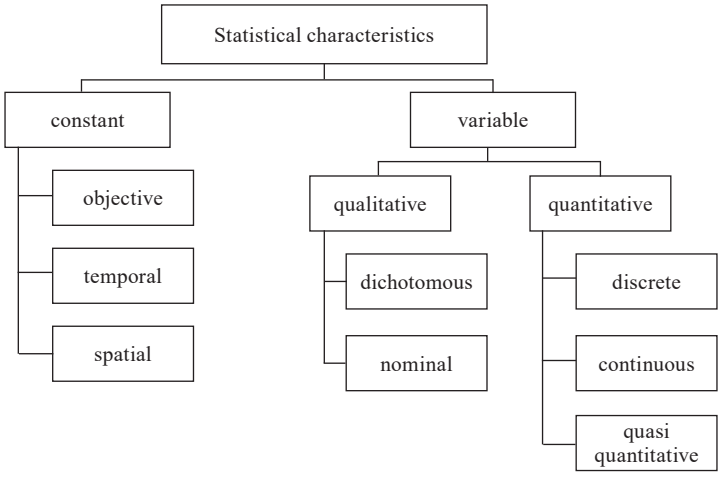
As already mentioned, the objects of statistical surveys are statistical populations. **A statistical population** is a collection of distinct units (such as persons, objects, or phenomena) that share characteristics relevant to the purpose of the study. A single element of the population is **a statistical unit**. The population should be defined in factual (who or what is being examined), temporal (which period/moment is being surveyed), and spatial (where the survey is being conducted) terms. The population could include, for instance, the inhabitants of Rome in 2022. In this case, it has been defined who, where, and when we are surveying. Here, the statistical units vary in several characteristics, such as gender, age, household size, and number of siblings. If the population were to be studied, for example, in terms of labour market participation, then the age of the individuals under study would need to be further defined to narrow down the population. Another example of a statistical population could be ships entering seaports in the European Union between 2018 and 2022. Distinguishing characteristics of the units could be ship type, port, flag, capacity, or deadweight.

The breakdown of the statistical characteristics, i.e. the properties of the statistical units mentioned above, is presented in Figure 1.1.

In general, we divide statistical characteristics into constant and variable ones. **Constant variables** take the same variants or values for all individuals in the population. They identify the population and determine an individual's affiliation with this population (hence the term "constant"). The different types of constant characteristics, i.e. objective, temporal, and spatial, define the population (who or what is being studied, when, and where?). **Variable characteristics** (also called variables) differentiate individuals in the population. They are divided into qualitative and quantitative variables. **Qualitative (categorical) variables** are usually expressed in words, although they can also be given in numbers. In this case, numbers are treated

as signs/symbols, e.g., a postal code, an individual identification number, a flat number, or a school number. If a characteristic has only two variants, it is called dichotomous.[2] **Quantitative (numeric) characteristics** are expressed numerically in units of measurement. Among them, we distinguish between **discrete characteristics,** which take values from the set of integers. Examples include the number of children in a household, the number of company employees, and the number of cars registered in each country. Other quantitative characteristics are **continuous**, taking values from the set of real numbers, for example, a child's age, length of service, floor area of dwellings, or production costs in firms. The last type of quantitative characteristics belongs to a rather special group of **quasi-quantitative characteristics**. These are characteristics initially expressed numerically, but ultimately presented in subgroups labelled with words. Examples of such characteristics are height: low, medium, high (where values expressed in, for example, centimetres are grouped and named accordingly); size of companies: small, medium, large (where the number of employees is usually taken into consideration); and size of shops by floor area: small, large.

Figure 1.1. Breakdown of statistical characteristics



Source: own study.

[2] Among qualitative characteristics, the literature distinguishes between nominal and ordinal characteristics (Jajuga 1993). In the present discussion, what matters in the study of correlations (Chapter 3) is whether a qualitative characteristics is ordinal or not.

Statistical characteristics for individual units in a population are measured on specific measurement scales. Generally, there are four **measurement scales** proposed by Stevens (1950), two non-metric: nominal and ordinal (rank scale) and two metric scales: interval and ratio scales (Siegel, Castellan 1988; Gatnar, Walesiak 2004). The measurement scales are ordered from weakest to strongest, based on possible algebraic operations that can be performed with the numbers. The following are examples of characteristics that can be measured on a given scale:

- nominal scale (identifying relations of equality and dissimilarity between the categories of characteristics): code of company activity (e.g., Central Register and Information on Economic Activity CEIDG), region of residence, country of product origin, postcode, individual identification number, satisfaction with the product (yes/no); qualitative characteristics,
- ordinal scale (allowing ordering of the characteristic's variants): education (lower secondary, basic vocational, secondary, tertiary); qualitative (ordinal) and quasi-quantitative characteristics,
- interval scale (where difference between the characteristic values can be determined, but quotient cannot be determined due to absence of 'absolute zero' as the zero is conventional): temperature (°C), IQ measurement scale, pH scale, financial score; quantitative characteristics,
- rank scale (enabling determination of differences and quotients for the characteristic's values; the start of the scale is zero): price (something can be twice as expensive), mass, number of employees, profit, age; quantitative characteristics.

At this point, so-called *dummy variables* should be mentioned (Bieszk-Stolorz, Markowicz 2019). These are qualitative or quantitative characteristics that have been assigned numerical values, for example:

- education: lower secondary – 1, basic vocational – 2, secondary – 3, tertiary – 4,
- age (in years): <20, 30) – 1, <30, 40) – 2, <40, 50) – 3.

Regarding education, the characteristic remains qualitative despite the assigned numerical values and is measured on an ordinal scale. In contrast, age, which is a quantitative characteristic expressed on a ratio scale, becomes a characteristic on an ordinal scale once the numbers are assigned to the intervals.

There is a relationship between the measurement scale of the characteristics under study and the ability to determine the relevant measures of statistical description (Table 1.1).

Table 1.1. Possible measures of statistical description depending on the scale of measurement of the statistical characteristics under study

Scale	Measures of		
	central tendency	variation	interdependency (two characteristics)
Nominal	mode	dispersion with respect to classification	relation strength coefficient
Ordinal	mode median quartiles	positional measures of variation	rank correlation coefficient
Interval and ratio	mean mode median quartiles	positional and classical measures of variation	linear and curvilinear correlation coefficients

Source: own study.

The groups of measures summarised above are discussed later in the book. Measures of central tendency are presented in Chapter 2.1, measures of variation in Chapter 2.2 and measures of interdependency in Chapter 3.

1.3. Statistical regularities

Socio-economic phenomena constitute a complex mechanism of interconnections and interdependencies. Not only economists but also any curious observer of the surrounding reality may find it interesting how wages are distributed in the economy, what the level of inflation is expected next year, or whether additional professional training can boost individual salaries. By seeking patterns or regularities in the behaviour of individual variables or processes, one aims to identify statistical regularities. The term **statistical regularity** refers to the repeatability, causality, and orderliness seen in mass phenomena and processes. Such regularity reveals objective interdependence among phenomena, causes, and effects (Główny Urząd Statystyczny – GUS). Regularities can be seen only within large masses of individuals. A single observation does not allow the use of statistical methods, and no general conclusions about the entire population can be drawn from it. Only in mass processes, i.e. in sufficiently large sets of observations, can statistical regularities be sought (Hare 1994; Ivanenko, Labkovsky 2015). A collected set of observations can provide the basis for hypotheses, conclusions, and their verification.

Mass phenomena arise from two types of causes: primary and secondary. **Primary causes** are intrinsic and result from systematic factors (systematic component), while **secondary causes** are incidental and result from random factors (random component). Statistical methods make it possible to separate these two components and quantify their values.

By observing objects, events, and processes both spatially and temporally, four types of statistical regularities can be identified (Hozer 1993):

- regularity of the distribution of values of statistical characteristics (structural regularity),
- regularities in the relationships between phenomena in space (regularity of spatial interdependence),
- regularities in the relationships between phenomena over time (regularity of temporal interdependence),
- regularities in the dynamics of a phenomenon (regularity of dynamics).

Figure 1.2 shows a graphical representation of statistical regularities. For example, the analysis of the age distribution of the unemployed in France in 2022 illustrates the regularity of the **distribution of the statistical characteristic values** (Figure 1.2a). Hence, the regularity refers to a single variable observed at a specific place and time (methods for studying this regularity are set out in Chapter 2).

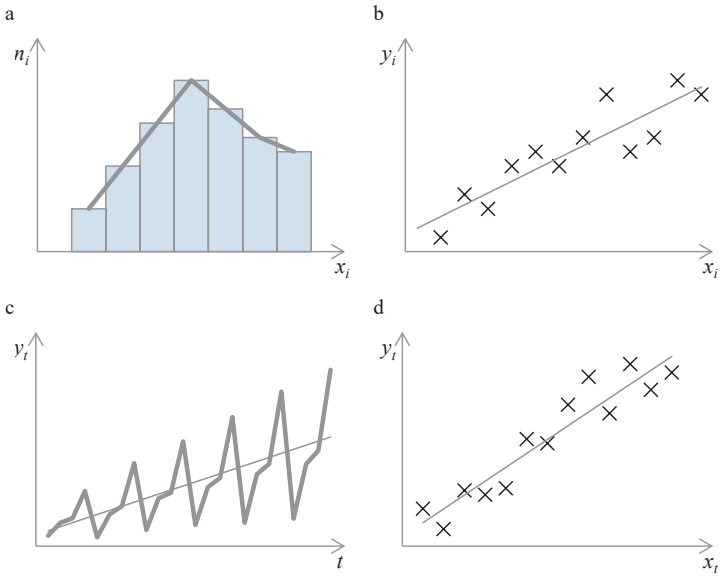
The regularity of spatial interdependence of phenomena (Figure 1.2b) refers to the observation of at least two variables that interact with each other within a certain unit of time. For instance, the relationship between the price of one litre of fuel and the price of one loaf of bread in selected European cities in 2022, or the relationship between the money spent on advertising and the profits generated by outlets selling a popular clothing brand in individual countries in 2020, can be analysed.

The regularity of temporal interdependence of phenomena (Figure 1.2d) also concerns the relationship between at least two variables, but unlike spatial regularity, this interdependence is observed over a certain period of time and within a certain space. For example, managers of a certain enterprise might find it interesting to examine relationship there is between the company's costs and revenues during the successive months of the year under review. Another example of regularity is temporal interdependence is the relationship between money spent on advertising and profits in a selected single store selling a popular fashion brand from 2000 to 2020. When studying regularities in interdependencies among variables (either over time or space), it is essential to use the methods presented in Chapter 3.

A regularity in phenomenon dynamics (Figure 1.2c) refers to a single variable and its changes over a certain period in a defined space. For example, live births in Spain between 1950 and 2020, ice cream sales by company X in Italy by month from 2015 to 2020, or the average price of one square metre of a flat on the secondary market in Berlin between 2010 and 2020 (see Chapter 4 for research methods).

In every socio-economic phenomenon or economic system, all four types of regularities **exist,** along with their interrelationships. The awareness of this fact allows researchers to adopt a holistic-structural approach. It is important to note that the causes of statistical regularities exist regardless of whether they have been explained or not.

Figure 1.2. Four types of statistical regularity



a – structural regularity, b – spatial regularity, c – dynamic regularity, d – temporal regularity.

Source: own study.

1.4. Forms of characteristics presentation

The collected characteristics to be examined are presented in a formalised manner to facilitate the recognition of statistical regularities. According to the adopted ordering criteria, **statistical clustering** (statistical classification) is carried out, meaning the sample population is divided into homogeneous or relatively homogeneous groups (Krzysztofiak, Luszniewicz 1981). Statistical clustering can be as follows (Zając 1994):

- typological,
- variance-based,
- analytical.

Typological (qualitative) **clustering** involves separating homogeneous groups from a statistical population according to variants of a given qualitative or quantitative characteristic measured on weaker measurement scales (nominal and ordinal). An example of this clustering is the clustering of residents by gender, occupation, or education. **Variance-based clustering** involves separating homogeneous groups from a statistical population based on a given quantitative characteristic (measured on interval or ratio scales). Units are grouped into classes or ranges of numerical values. Examples of such clustering include the division of cities by population, the division of the economically active by age, or the classification of divorces by the number of children in a marriage. **Analytical clustering** is used when the research objective is to establish correlations between the characteristics under study, e.g. the work experience and salary, or the flat floor area and the price of one square metre.

A good classification should meet several conditions (Zając 1994) and must be performed in a manner that is:

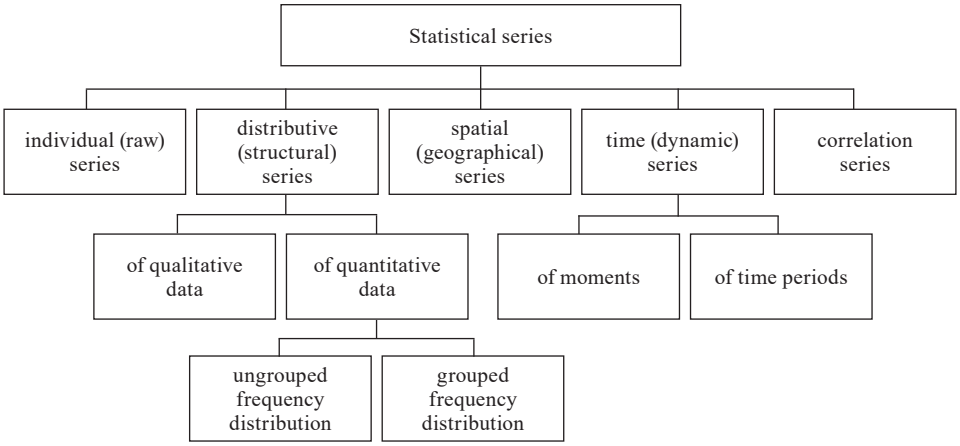
- **disjunctive** – individual units with specific characteristics must be explicitly assigned to classes,
- **complete** – classes (groups) must cover all units present in a given population.

The statistical data compiled according to the adopted statistical clustering should be appropriately presented using statistical **series**, which can be displayed in the form of statistical **tables** and **figures**.

1.4.1. Statistical series

A statistical series is an ordered set of values or variants of a specific characteristic according to the adopted ordering criteria. Figure 1.3 shows the classification of statistical series.

Figure 1.3. Classification of statistical series



Source: own study.

Statistical series can be classified as follows (Hozer 1998; Kukuła 1998; Sobczyk 2007):

- **unorganised** – this includes detailed series, that present information extracted from censuses, surveys, or other primary sources in a random order,
- **organised** – a sequence of increasing or decreasing statistical quantities, ordered by specific characteristics or categories.

An individual (raw) **series** is an unorganised or organised compilation of raw observations extracted directly from the survey. For example, a summary of the age of people surveyed:

22, 22, 21, 20, 20, 21, 20, 20, 21, 25, 22, 22, 20, 20, 20, 22, 24, 21, 21, 22, 23,

or their gender (F – female, M – male):

F, F, F, F, M, F, F, M, M, M, F, F, F, F, F, F, F, F, F, F, M,

as well as any other characteristics that has been derived from the source survey. The general notation of the detailed series will be as follows: $x_1, x_2, x_3, ..., x_n$, where n is the number of units in the survey.

Detailed series are especially useful at the data collection stage, as they cover the entire statistical data set. However, they are not very transparent when the number of cases is large. They are typically used as a form of data presentation when the number of observations is relatively small.

The importance of individual series has significantly grown with the development of computer tools supporting data analysis (i.e. spreadsheets such as MS Excel, dedicated statistical programs as STATISTICA or Gretl, and programming languages e.g. Python, R language). Large or very large databases (consisting of many individual series) can easily be analysed and processed to identify statistical regularities.

The use of other types of statistical series depends on the type of characteristics and statistical regularity being analysed (see Chapters 1.2 and 1.3). A series typically consists of at least two columns: the first column contains the variants of the characteristics or time units (often in increasing order), denoted by the symbol x_i where i stands for the number of classes or variants (for time series, this is a unit of time: t), while the second column shows the size of individual classes or variants (number of units for a given characteristic value), denoted by the symbol n_i (in time series, this is the value of a phenomenon in successive periods y_t).

In distributive series, the first column x_i presents the variants of the characteristics (its values, or intervals, depending on the characteristic presented; see Tables 1.2–1.4), while the second column contains the **sample size** n_i corresponding to the successive intervals, i.e., the number of units with a given variant (value of the characteristic or a value within the interval). In additional columns of the resolution series, the following measures can be presented:

– **relative frequency** $\frac{n_i}{n}$, i.e., the proportion of the number of units in the sample population with a given variant,

– **percentage** $\frac{n_i}{n} \cdot 100$ – indicating the percentage of units in the sample population with a given variant.[3]

[3] The percentage of the variable under study is applicable only when the number of observations in the sample population is at least 100 ($n \geq 100$).

Table 1.2. Distribution series for qualitative characteristics

Characteristic's variants x_i	Number of units n_i	Relative frequency $\frac{n_i}{n}$	Percentage (%) $\frac{n_i}{n} \cdot 100$
x_1	n_1	$\frac{n_1}{n}$	$\frac{n_1}{n} \cdot 100$
x_2	n_2	$\frac{n_2}{n}$	$\frac{n_2}{n} \cdot 100$
⋮	⋮	⋮	⋮
x_k	n_k	$\frac{n_k}{n}$	$\frac{n_k}{n} \cdot 100$
$\sum$	n	1	100

Source: own elaboration based on Hozer 1994.

Table 1.3. Series for quantitative characteristics – discrete series

Characteristic's values x_i	Number of units n_i	Relative frequency $\frac{n_i}{n}$	Percentage (%) $\frac{n_i}{n} \cdot 100$
x_1	n_1	$\frac{n_1}{n}$	$\frac{n_1}{n} \cdot 100$
x_2	n_2	$\frac{n_2}{n}$	$\frac{n_2}{n} \cdot 100$
⋮	⋮	⋮	⋮
x_k	n_k	$\frac{n_k}{n}$	$\frac{n_k}{n} \cdot 100$
$\sum$	n	1	100

Source: own elaboration based on Hozer 1994.

Table 1.4. Series for quantitative characteristics – continuous series

Range values x_{0i}–x_{1i}	Number of units n_i	Relative frequency $\frac{n_i}{n}$	Percentage (%) $\frac{n_i}{n} \cdot 100$
x_{01}–x_{11}	n_1	$\frac{n_1}{n}$	$\frac{n_1}{n} \cdot 100$
x_{02}–x_{12}	n_2	$\frac{n_2}{n}$	$\frac{n_2}{n} \cdot 100$
⋮	⋮	⋮	⋮
x_{0k}–x_{1k}	n_k	$\frac{n_k}{n}$	$\frac{n_k}{n} \cdot 100$
$\sum$	n	1	100

Source: own elaboration based on Hozer 1994.

The **distributive** (structural) **series** is an organised set of information representing the structure of a characteristic under study at a particular place and time. These series are used to detect regularities in the structure of phenomena. The resolution series can be divided into:

- **a series for qualitative characteristics** – this category includes series for characteristics expressed on weak measurement scales: for qualitative characteristics (Table 1.5) and ordinal characteristics (Table 1.6),
- **a series for quantitative characteristics** – depending on the nature of the characteristics being described, this series can be divided into:
  - an ungrouped frequency distribution for a discrete characteristic (Table 1.7)
  - a grouped frequency distribution for:
    - a continuous characteristic (Table 1.8),
    - a discrete characteristic with a large number of variants (Table 1.9).

See Section 2.1 for more on creating intervals for quantitative variables.

Table 1.5. Population of the European Union (EU-27) by gender as of 1 January 2022

Gender x_i	Population (million) n_i
Females	2,283.9
Males	2,183.4
Total	4,467.4

Source: Eurostat (demo_pjangroup), accessed 12/11/2023.

Table 1.6. Unemployment in the EU-27 by educational attainment of citizens aged 15–74 in 2022

Educational attainment x_i	Unemployed (thousand) n_i
Levels 0–2	4,646
Levels 3 and 4	5,660
Levels 5–8	2,958

Source: Eurostat (une_educ_a), accessed 20/11/2023.

Table 1.7. Number of household-dwelling units in Finland by number of rooms in 2022

Number of rooms x_i	Number of dwelling units n_i
1	75,265
2	343,386
3	437,946
4+	869,771
Total	1,726,368

Source: Statistics Finland, dwellings, and housing conditions, https://pxcharacteristics.stat.fi/PxWeb/pxweb/en/StatFin/StatFin__asas/statfin_asas_pxt_115y.px, accessed 11/12/2023.

Table 1.8. Live births by mother's age in EU-27 in 2021

Age (year) x_i	Live births (thousand) n_i
10–19	84.4
20–29	1,447.2
30–39	2,324.7
40–49	230.7
50+	1.4
Total	4,088.3

Source: Eurostat (demo_fasec), accessed 12/12/2023.

Table 1.9. Structure of NUT 3 regions by live births in 2022

Live births (persons) x_i	NUT3 regions n_i
less 200	6
200–499	66
500–999	193
1,000–1,999	361
2,000–4,999	445
5,000–9,999	165
10,000–19,999	69
over 20,000	28
Total	1,333

Source: Eurostat (demo_r_births), accessed 30/06/2024.

The **spatial** (geographical, territorial) **series** represent the distribution of the phenomena under study across space (parts of the world, countries, economic, or geographical regions, administrative units) during a specific period. Examples of such

a series include the daily temperature forecast in European cities or internet users in world regions in 2019 (Table 1.10).

As with the distributive series, in the spatial series, the variants of the characteristic are listed in the first column x_i, while the second column n_i contains the number of units with individual variants in the population under study. The columns may also contain the relative frequency $\frac{n_i}{n}$ and percentage $\frac{n_i}{n} \cdot 100$, analogous to Table 1.2.

Table 1.10. Internet users by geographical region in 2019

Geographical region x_i	Internet users (per 1000 inhabitants) n_i
Europe and Central Asia	664
South Asia	295
East Asia and Pacific countries	510
Middle East and North Africa	509
Sub-Saharan Africa	187
Central and South America	617
Globally	490

Source: Statistical Yearbook of the Republic of Poland (2020), 747.

In contrast to the series presented above, **time** (chronological, dynamic) **series** refer to characteristics that change over time. The characteristic by which the population is grouped is the unit of time (year, quarter, month, week, day). Time series are used to detect patterns of change over time (more on this in Chapter 3). The time series are divided into:

- **moments:** these refer to characteristics observed at a specific point in time, e.g., the water level of the Vistula River on successive days of January 2020, or the population of the European Union from 2013 to 2022 as of 1st January (Table 1.11),
- **time periods:** these show the development of the analysed phenomenon over successive time periods (days, months, quarters, years), e.g., beer sales in the quarters of 2020–2023 in chain shops, or the number of nights spent by tourists in the EU-27 during individual months from January 2022 to June 2023 (Table 1.12).

Time series are also recorded in at least two columns, where the first column contains time units (always in ascending order) denoted by t, while the second column shows the value of the phenomenon in successive periods y_t.

Table 1.11. The population of European Union (EU-27) from 2013 to 2022 (as of 1st of January)

Year T	Population (million) y_t
2013	441.3
2014	442.9
2015	443.7
2016	444.8
2017	445.5
2018	446.2
2019	446.4
2020	447.3
2021	447.1
2022	446.7

Source: Eurostat (demo_pjan), accessed 30/11/2023.

Table 1.12. Nights spent at tourist accommodation establishments in the EU-27 – monthly characteristics from 01/2022 to 06/2023

Unit of time t	Accommodated nights (million) y_t
2022-01	89.6
2022-02	111.4
2022-03	134.2
2022-04	189.2
2022-05	226.7
2022-06	306.0
2022-07	444.8
2022-08	479.9
2022-09	286.4
2022-10	208.0
2022-11	129.9
2022-12	136.2
2023-01	129.6
2023-02	140,1
2023-03	154.8
2023-04	210.6
2023-05	250.2
2023-06	312.8

Source: Eurostat (tour_occ_nim), accessed 30/11/2023.

The **correlation series** are produced by combining at least two characteristics to identify regularities in correlations over time or space (more on this in Chapter 3). Depending on the type of regularity analysed, the designations of columns in the correlation series vary. Regarding regularities in:

– the interdependence of variables in space, where time is strictly defined: The first column presents the first variable – x_i (where i is the number of cases or objects analysed), and the second column contains the second variable – y_i (e.g. the share of people 60+ at risk of poverty or social exclusion versus the share of pension expenditure in GDP in the EU27 in 2021 – Table 1.13), and
– the interdependence of variables over time, where the space (the object of study) is strictly defined: The first column contains the first variable – (x_t where t means subsequent units of time), and the second column shows the second variable ($y_{t)}$. For example, the share of people 60+ at risk of poverty or social exclusion vs. share of pension expenditure in GDP in the EU-27 from 2015 to 2021 (Table 1.14).

An additional column is added, marked with i or t, providing information about the spatial or temporal range of the analysed variables. More variables may be presented in a correlation series, with successive variables denoted by successive letters of the alphabet and the subscript i or t respectively (depending on the type of regularity being analysed).

Table 1.13. Percentage of persons 60+ at risk of poverty or social exclusion and share of expenditure on pensions as percentage of gross domestic product (GDP) in EU-27 countries in 2021

EU countries i	Persons (%) x_i	Share of GDP (%) y_i
1	2	3
Austria	15.2	15.0
Belgium	18.3	12.6
Bulgaria	41.7	7.8
Croatia	31.7	9.8
Cyprus	19.1	8.8
Bohemia	11.3	8.9
Denmark	13.1	11.8
Estonia	37.6	8.1
Finland	13.7	13.4
France	14.8	14.9
Germany	20.9	12.2
Greece	21.8	16.4
Hungary	20.3	7.0
Ireland	21.8	4.5
Italy	19.2	16.3
Latvia	42.5	7.9
Lithuania	35.3	7.1

1	2	3
Luxembourg	11.2	9.6
Malta	29.2	6.4
Netherlands	18.5	12.1
Poland	19.2	10.8
Portugal	25.3	14.2
Romania	38.3	8.8
Slovakia	13.4	8.5
Slovenia	18.4	10.0
Spain	22.8	13.9
Sweden	13.2	10.6

Source: Eurostat (ilc_peps01n and spr_exp_pens), accessed 11/12/2023.

Table 1.14. Percentage of persons 60+ at risk of poverty or social exclusion and share of expenditure on pensions as percentage of gross domestic product (GDP) in EU-27 from 2015 to 2021

Year t	Persons (%) x_t	Share of GDP (%) y_t
2015	19.8	13.1
2016	19.9	13.0
2017	19.8	12.8
2018	20.2	12.7
2019	20.4	12.7
2020	20.8	13.6
2021	20.3	12.9

Source: Eurostat (ilc_peps01n and spr_exp_pens), accessed 11/12/2023.

1.4.2. Statistical tables

Statistical tables are used to present statistical series in a tabular format. Tables offer a breakdown of collected data in a concise and clear manner. Statistical tables are built in a conventional way, but through the proper organisation of data, they allow, among other things, the comparison of specific figures. Statistic tables can be classified according to two criteria:

- chronological: Reference (raw) table and summary (analytical) table,
- mode of construction: Simple table and complex table.

Reference tables are used at the preliminary stage of a statistical survey and their purpose is to collect statistical material (i.e., a compilation of detailed series).

Summary tables are a compilation of the results of a statistical survey, serving as the basis for statistical analyses and suitable for publication.

Simple tables are a **compilation** containing only one statistical series (Tables 1.5–1.12), whereas **complex tables** present several characteristics of a single population or several populations described by a single characteristic (more on this in Chapter 3.1).

A correctly built statistical table should contain several elements that precisely define the characteristics it contains:

- **Title of the table** – this should concisely define the content of the table. The title should identify the population represented (factual, spatial, and temporal definition) and the type of characteristics by which the clustering was performed.
- **Column and row headings, or captions and stubs** – these provide a verbal indication of the content of the rows and columns. Rows represent the statistical units (or groups) or time periods under study, while columns display the characteristics under study.
- **Units of measurement** in which the quantities under consideration are expressed. If all the characteristics in the table are presented in the same unit, including this information in the table title is sufficient.
- **Headnotes** relating to the figures presented, placed directly below the table.
- **Source notes concerning statistical characteristics**, that allow the figures to be checked and lend credibility to the figures presented.
- **Symbols** to mark boxes not filled in the table, values not included, or completed or corrected.

A statistical table that includes all the elements listed above should make the information presented clear and understandable without requiring readers to refer to the comments in the body of the study or report.

All boxes of the statistical table should be filled in. However, sometimes it is not possible to fill in some of the boxes with the accurate figures, so the following **symbols** are used (e.g. *Statistical Yearbook* 2020):

- dash (–) – magnitude zero,
- zero (0) or (0.0) – magnitude greater than zero, but less than 0.5 of a unit,
- dot (.) – data not available, classified data (statistical confidentiality) or when providing data is impossible or purposeless,
- cross (×) – entering an item is not possible or advisable due to the layout of the table,
- "of which" – indicates that not all elements contributing to the total are provided.

For better clarity in statistical tables, the data presented should be properly arranged. Several classification schemes can be used:

- alphabetical (Table 1.5),
- chronological (Tables 1.11 to 1.12),
- geographical (Table 1.10),
- by size (Tables 1.7 to 1.9),
- by generally adopted classification (Table 1.6).

Numbers in the summary table are rounded off, while numbers in the reference table are always presented in their raw form. In addition, data can be presented as measures (the numerical state of a specific phenomenon) or indices (the ratio of two economic or other categories – Tables 1.2–1.4). Indices are particularly useful when comparing the structure of the analysed variables.

1.4.3. Statistical figures

Statistical figures are graphical representations of statistical series. They help visualise the information contained in statistical tables, thus adding depth to the analysis but also disseminating knowledge by presenting the data in an interpretable way. The principal advantages of charts for presenting statistical results are as follows (*Graphical Presentation of Statistical Characteristics* 2014):

- graphic presentation is more appealing and convincing than the numbers in the table,
- graph gives an idea of the general character of the phenomenon,
- when carefully designed, they are attractive to the user,
- eye-catching way of presenting data grabs the user's attention,
- graphs are more effective because visual perception is faster than reading many figures and tables,
- they provide an opportunity to convey complex information in a simple way,
- they are helpful for comparing and interpreting phenomena,
- they help to remember information about phenomena,
- they are universal tools, adaptable to various fields and languages,
- they are easy to understand, even for non-specialists.

Just like statistical tables, each figure must have several essential characteristics:

- **Appropriate title** – it should concisely define the image presented in the figure. The title should identify the population depicted (in objective, spatial, and temporal terms) and the type of characteristic (or characteristics) presented.

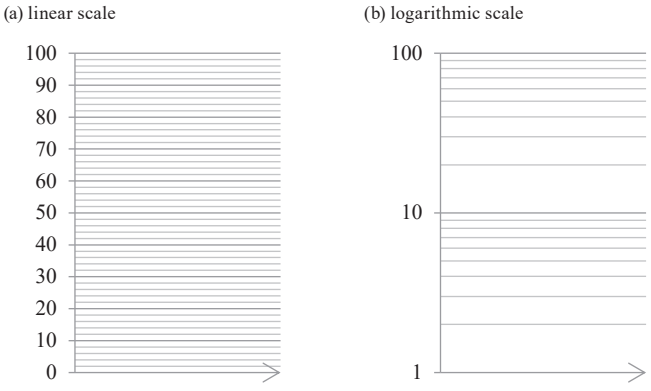
– Adequately selected **unit** and **scale** on the axes in the coordinate system.
– **Graphic presentation** – it should be carefully chosen regarding the analysed data and statistical regularity.
– **Labelled axes** in the coordinate system.
– **Legend** – if more than one characteristic is presented in the graph, it is necessary to provide a legend which explains the use of colours, symbols, lines, etc.
– **Source note** – it should indicate place where information used to create the diagram was obtained.

Figures are mostly drawn in a Cartesian coordinate system, where the *x*- and *y*-axes are placed perpendicular to each other, and each point on the plane corresponds to a pair of numbers – the coordinates of that point. The **scale** on each axis can be:

– **linear**, where equal graphic intervals correspond to equal numerical intervals (Figure 4a),
– **non-linear**, where unequal graphical intervals correspond to equal numerical intervals, or equal graphical intervals correspond to unequal numerical intervals, e.g., logarithmic scale (Figure 4b) or a square root scale.

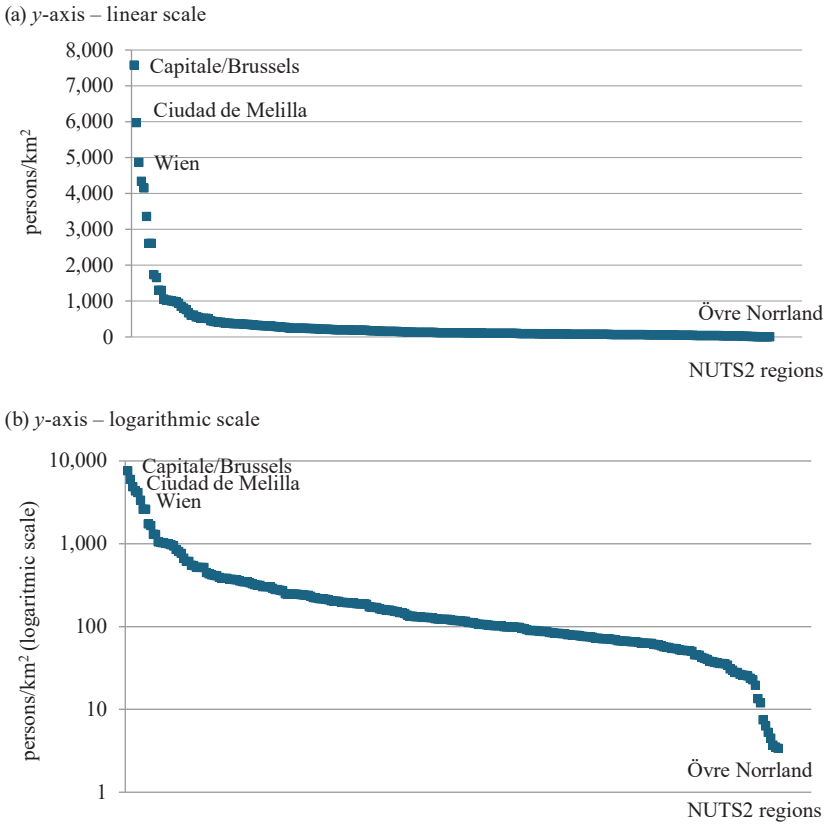
The logarithmic scale is the most common non-linear scale, used to represent changes in ratios (coefficients – Figure 1.5) or variables with a wide range of characteristics.

Figure 1.4. Linear scale and logarithmic scale



Source: own study.

Figure 1.5. Population density in NUTS2 regions in Europe in 2022 – linear versus logarithmic scale



Source: own compilation based on Eurostat (tgs00024) accessed 19.02.2024.

Both axes can be represented using a linear scale or a logarithmic scale, but it may also be the case that only one axis (usually the *y*-axis) is logarithmic. In this case, it is referred to as a semi-logarithmic scale.

Charts are commonly used to visualise characteristics in statistical research including:

- pie charts,
- bar charts,
- dot plots,
- heat maps,
- line charts,

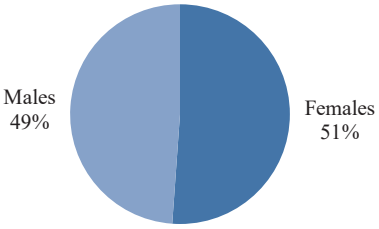
– slope graphs, and
– statistical maps.

Different types of graphs are used depending on the type of characteristic being analysed and the statistical series in which the characteristic is presented.

In a **pie chart,** a circle is divided into sections, where the arc of each section is proportional to the intensity or size of the observed phenomenon. While commonly used, the pie chart is often criticised (Wilkinson 2005). It is recommended that this type of chart be used with caution because of the difficulty of comparing different sections of a single chart or comparing characteristics across different charts (Few 2007).

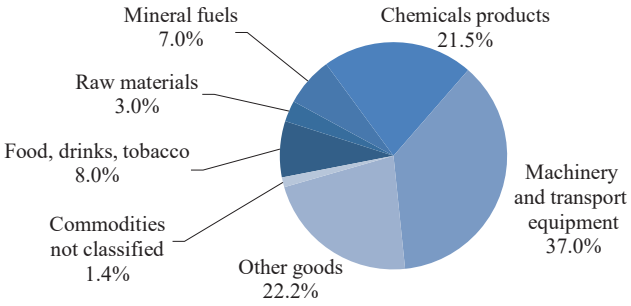
A well-designed pie chart should be used primarily to visualise qualitative features presented on a nominal scale, with not too many sections (Figures 1.6–1.7). 3D charts should be avoided, as they make it even more difficult to assess the size of individual sections.

Figure 1.6. Population of the European Union (EU-27) by gender 01 Jan 2022



Source: Table 1.5.

Figure 1.7. Structure of product exports by EU27 countries in 2020



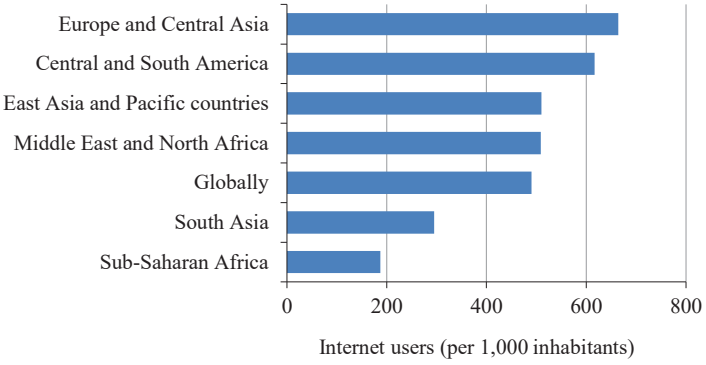
Source: Eurostat (tet00013), accessed on 19/02/2024.

Bar charts are graphs used to represent the level of a phenomenon using columns whose heights or lengths are proportional to the size of the phenomenon. One axis shows the variants of the characteristic, while the other axis shows the volume of individual variants. Variants can be presented on either the *x*-axis or the *y*-axis, especially when variant labels are long (Wilke 2020).

Bar charts are used to present characteristics in geographical (Figure 1.8) and resolution series that relate to characteristics such as: qualitative (Figure 1.9); ordinal (Figure 1.10); quantitative discrete (Figure 1.11); or quantitative continuous (Figure 1.12). The bars should be arranged according to the natural (arbitrary) order of the characteristic's variants (in Figure 1.10 – from the lowest level of education to the highest, in Figures 1.11–1.12 – successive numbers). When the chart depicts a characteristic presented on a nominal scale or in a geographical series, and the variants cannot be logically ordered (except alphabetically), the bars should be ordered by the magnitude of the phenomenon (Figures 1.8–1.9).

A bar chart made for a continuous characteristic (Figure 1.12) is called a **histogram**. This type of graph helps visualise the distribution of a variable and is the most commonly used graph for examining the correctness of analysis of structure (cf. section 2.1).

Figure 1.8. Internet users by geographical region in 2019

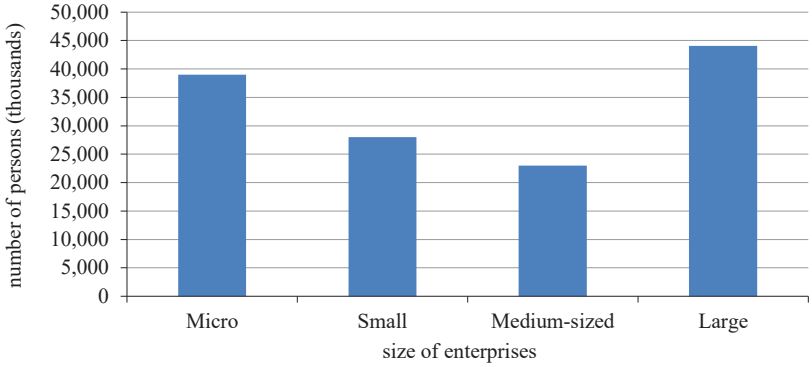


Source: Table 1.10.

Figure 1.9. Number of enterprises in EU-27 countries in 2021 by statistical classification of economic activities (NACE Rev.2)

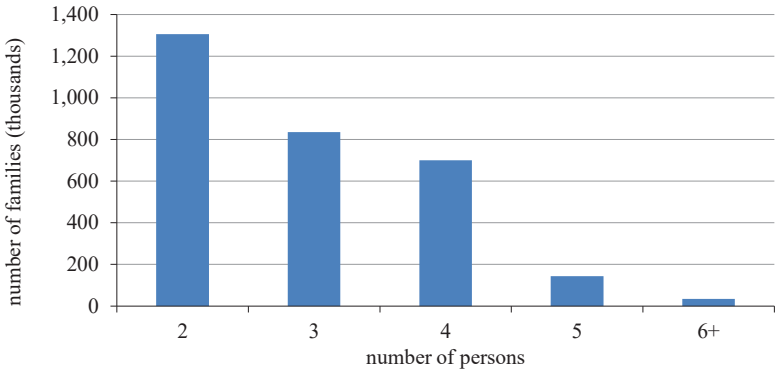
Source: own compilation based on Eurostat (bd_l_form), accessed 19/02/2024.

Figure 1.10. Number of persons employed (thousands) in the non-financial business economy in EU-28 in 2012 – by size of enterprise



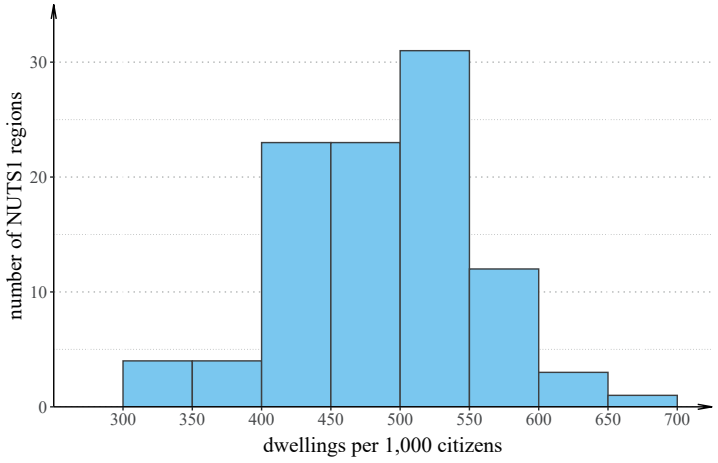
Source: *Archive: Business economy – size class analysis*, https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Archive:Business_economy_-_size_class_analysis&direction=next&oldid=75115, accessed 19/02/2024.

Figure 1.11. Families in Greece by number of persons in 2011



Source: Eurostat (cens_11fts_r3), accessed 19/02/2024.

Figure 1.12. Distribution of dwellings per 1,000 inhabitants in the NUTS1 regions of Europe in 2011



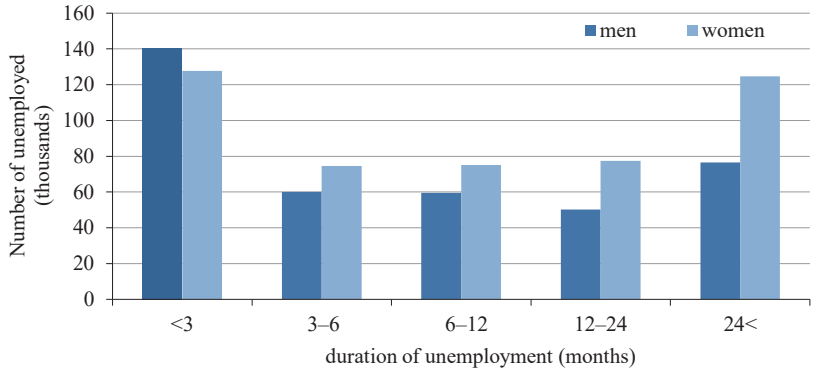
Source: own calculations based on Eurostat (cens_11ag_r3; cens_11dwob_r3), accessed 19/02/2024.

As for bar charts, we distinguish between **grouped charts** (a group of bars is created for one variant – Figure 1.13) and **cumulative charts** (bars are drawn one on top of the other, which is particularly useful when information about the sum of values is relevant – Figure 1.14). When the structure of a characteristic is more

interesting than its summed value, a cumulative bar chart can be used, where the sum of values for individual objects (voivodeships) equals 100% (Figure 1.15).

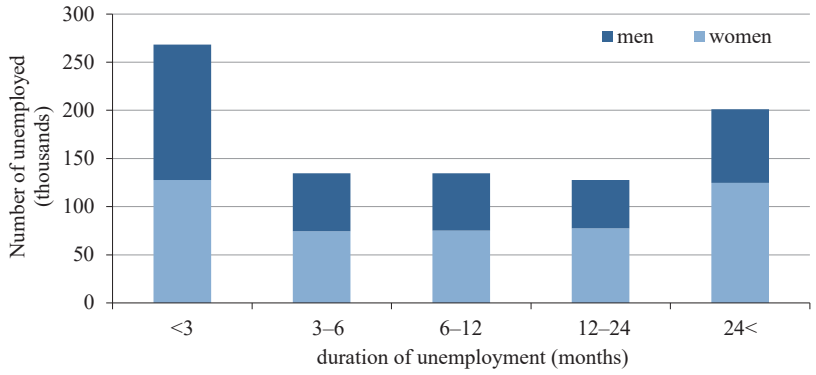
In **dot plots** (or dot charts), a characteristics can be presented in qualitative characteristic series, quantitative discrete characteristic series, and in geographic and correlation series.

Figure 1.13. Registered unemployed by duration of unemployment and gender in Poland in 2019 (as of 31.12) – a grouped bar chart



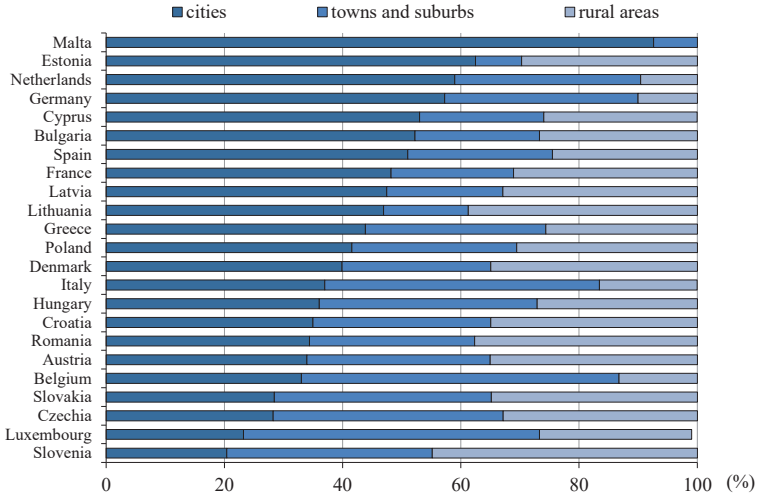
Source: CSO, Local Characteristics Bank, accessed 26/06/2021.

Figure 1.14. Registered unemployed by duration of unemployment and gender in Poland in 2019 (as of 31.12) – a cumulative bar chart



Source: CSO, Local Characteristics Bank, accessed 26/06/2021.

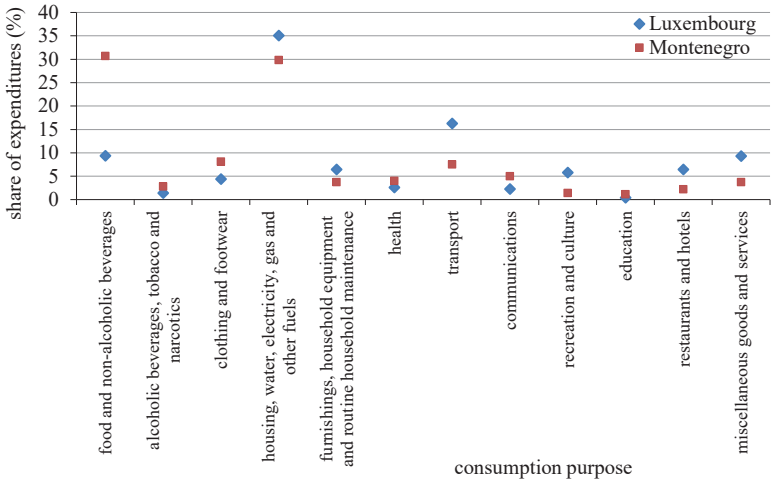
Figure 1.15. Household characteristics by degree of urbanization in selected countries of Europe in 2020



Source: Eurostat (hbs_car_t315), accessed 19/02/2024.

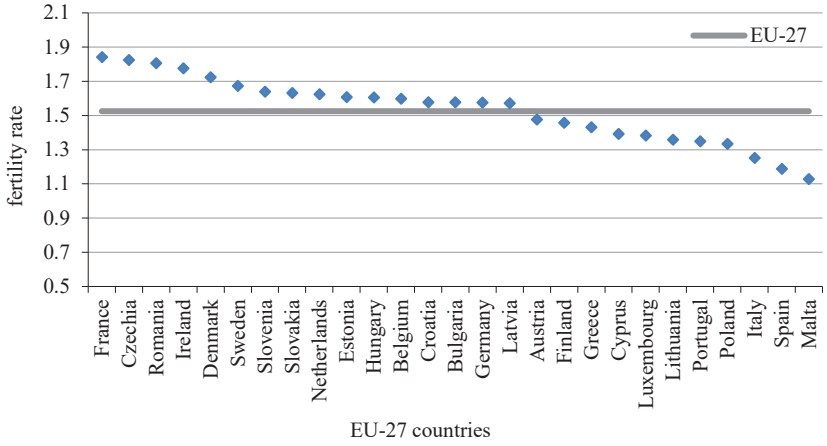
For the first three series, the diagrams look similar to Figures 1.8–1.11, except that different dot shapes are used instead of rectangles. In this case they are a rhombus and a square (Figures 1.16–1.17).

Figure 1.16. Structure of expenditure by consumption purpose (%) in 2020 of Luxemburg and Montenegro inhabitants



Source: own study Eurostat (hbs_str_t211), accessed 19/02/2024.

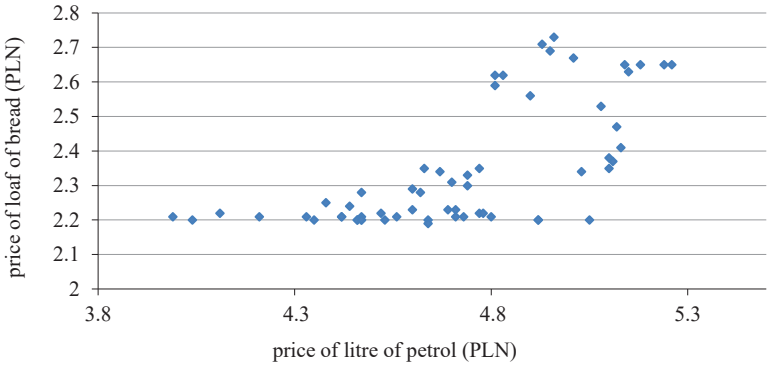
Figure 1.17. Level of fertility rates in EU-27 countries against mean EU-27 in 2021



Source: Eurostat (demo_frate), accessed 20/02/2024.

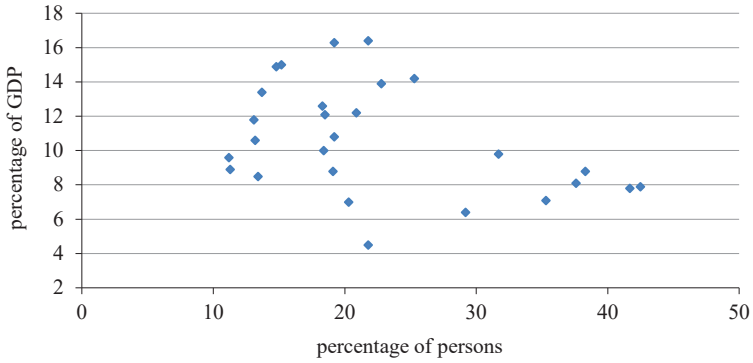
Correlation diagrams (also known as scatterplots for empirical expectations) are a variation of a dot plot. These diagrams are used to present correlation series. The values of one variable are plotted on the *x*-axis, and the values of the other variable on the *y*-axis. The scatterplot provides an easy way to see the nature of the relationship (or the lack thereof) between the variables. Correlation diagrams are used to illustrate patterns of interdependence, both in time (Figure 1.18) and in space (Figure 1.19).

Figure 1.18. Price of litre of petrol and a loaf of wheat-rye bread (0.5 kg) in Poland in the following months from 2015 to 2019



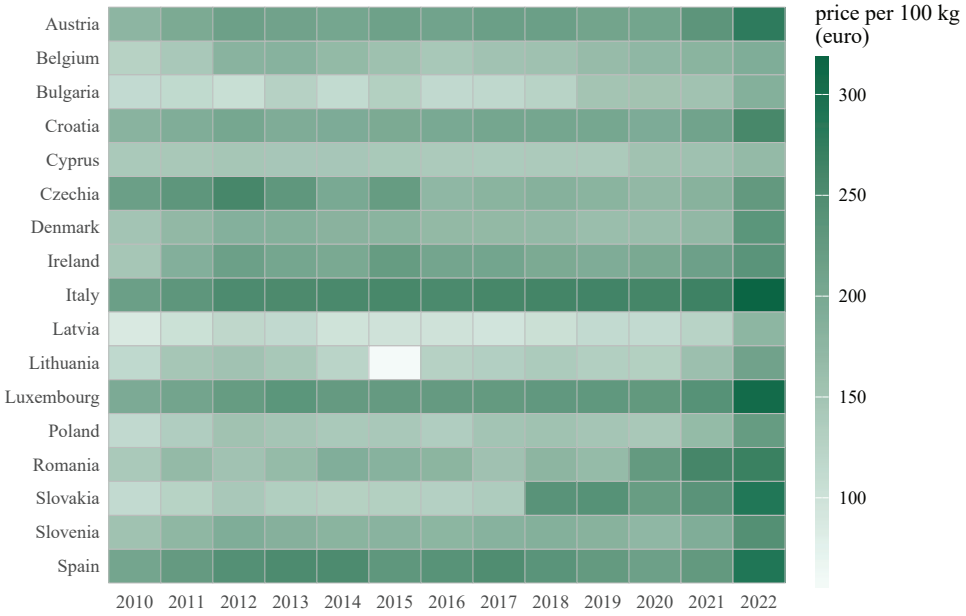
Source: CSO, Local Characteristics Bank, accessed 25/06/2021.

Figure 1.19. Percentage of persons 60+ at risk of poverty or social exclusion and share of expenditure on pensions as percentage of gross domestic product (GDP) in EU-27 countries in 2021



Source: Table 1.13.

Figure 1.20. Prices per 100 kg live weight young cattle in chosen countries of Europe in years 2010–2022



Source: Eurostat (apri_ap_anouta), accessed 20/02/2024.

A heat map is another way of graphically representing the intensity of a phenomenon, this time by using colours (Wilkinson, Friendly 2009). The map may use one colour, with its shade indicating the intensity of the phenomenon, or a fixed colour scale. Although, the graph cannot directly display the magnitude of the data presented, it clearly shows more general trends.

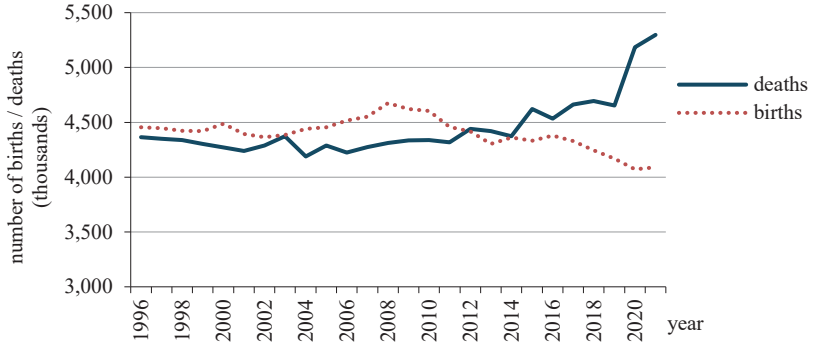
There are two categories of heat maps: clustered and spatial. In a **clustered heat map**, the values of a variable are arranged in a matrix of fixed cell size (Figure 1.20), while a **spatial heat map** reflects the position of a value in space, where the phenomenon constantly changes.

A line graph is designed to represent continuous characteristics and is therefore commonly used when visualising changes over time. It is very simple to prepare and easy for the audience to understand. The *x*-axis shows the units of time during which the variable is studied (these can be years, quarters, months, or days). On the *y*-axis, the level of the given phenomenon is marked. This type of chart is popular because it helps to simultaneously compare multiple variables, reveals trends, and provides the opportunity to produce forecasts, and identifies a range of the variable's evolution (Figure 1.23). It also permits the identification of atypical observations.

A well-designed line graph should allow for an easy identification of variables, e.g., through the use of different line styles (Figure 1.22), colours (Figure 1.21), or markers (when long time series are presented on the graph, the use of markers would make the graphical image difficult to read). The order of the characteristics in the graph legend should correspond to the order of the feature values in the most recent analysis period. The *x*-axes of Figures 1.21–1.23 are labelled only for selected periods (to avoid cluttering the chart), but all the characteristics are shown at the same time intervals (annual, quarterly, and daily respectively).

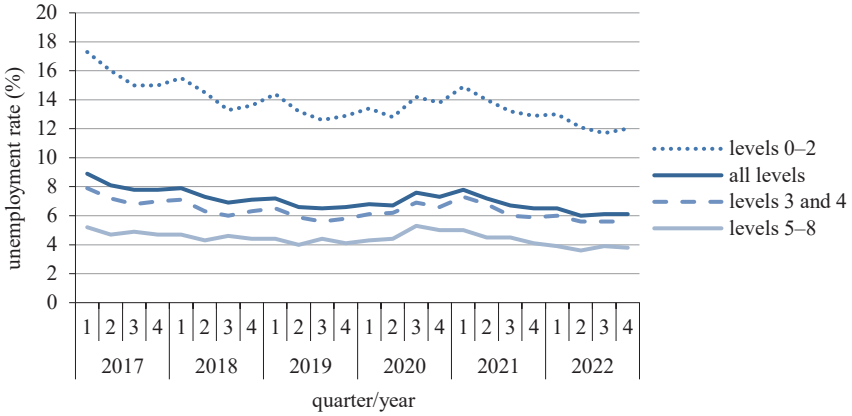
Particular attention should be paid to adjustment made to the range of the *y*--axis. The *y*-axis does not always have to start at 0 (especially if the characteristic's variability is only observable within a certain range, as shown in Figure 1.23). However, it is essential to ensure that the nature of the relationship between the presented variables remains intact when the axis is rescaled. In addition, recipients of the graph by default expect the *x*- and *y*-axes to intersect at 0, so the y-axis should only be rescaled in justified cases.

Figure 1.21. Births and deaths in EU-27 in years 1996–2021



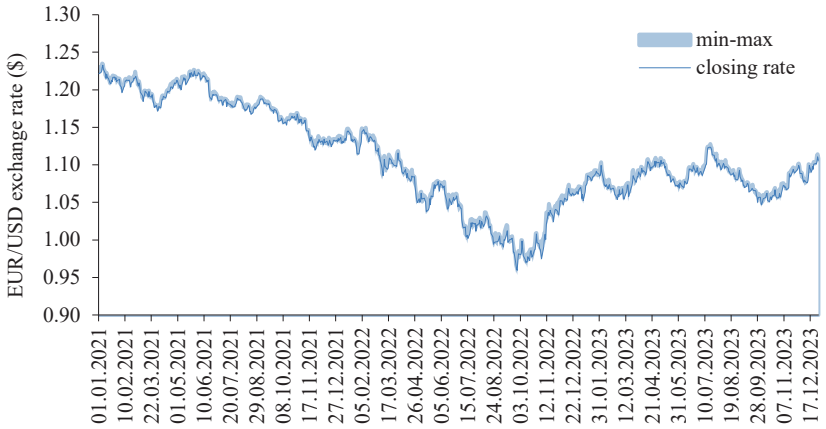
Source: Eurostat (demo_fmonth, demo_mmonth), accessed 20/02/2024.

Figure 1.22. Unemployment rate for EU-27 by educational attainment level of unemployed (by educational attainment level) in quarters 2017–2022



Source: Eurostat (lfsq_urgaed), accessed 21/02/2024.

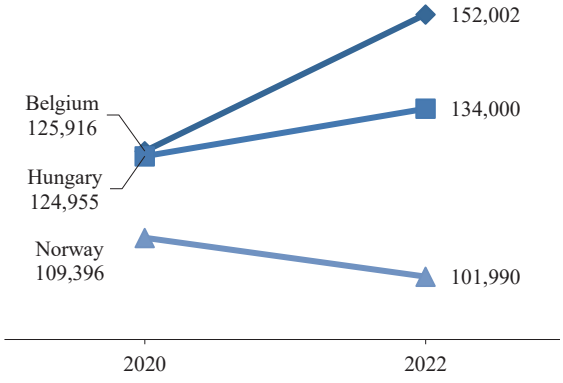
Figure 1.23. Daily EUR/USD ($) exchange rate from 1.01.2021 to 31.12.2023



Source: own compilation based on https://pl.investing.com, accessed 21.02.2024.

The slope graph is a specific type of a line graph that is primarily recommended for visualising short-term changes. This method of presenting data allows absolute values to be rearranged and helps users quickly assess which changes were major and which were minor during the periods under study (Figure 1.24). The slope of each line represents the magnitude and direction of change. However, this type of graph is not recommended when presenting data with multiple intersecting lines.

Figure 1.24. Changes in the number of dwellings bought on the secondary market in selected European countries in 2020 and 2022



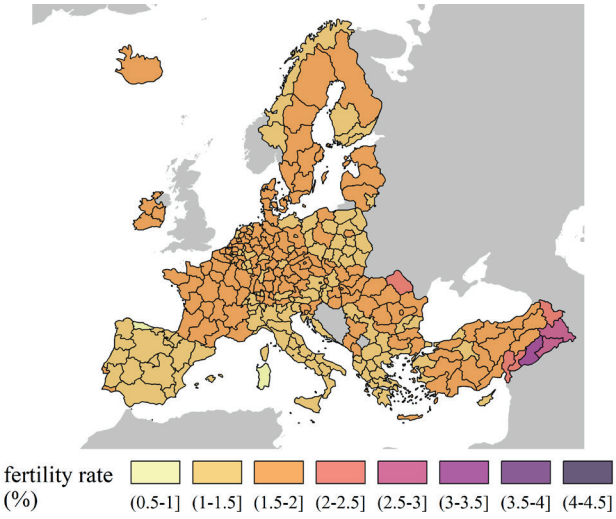
Source: Eurostat (prc_hpi_hsna), accessed 21/02/2024.

For characteristics presented in the form of geographical series, maps are the best way to present them graphically. When the basis for constructing a map is a numerical characteristic (e.g., economic or social variables) referring to a specific point, line, or surface objects, the map is defined as a **statistical map** (Kocimowski, Kwiatek 1977). Statistical maps allow for an intuitive understanding of the distribution of the presented variables. A distinction is made between methods of presenting characteristics on maps (*Graphical presentation...*, 2014):

– qualitative methods: signature method, range method, chorochromatic method,
– quantitative methods: cartodiagrams, dot method, cartograms, isoline method.

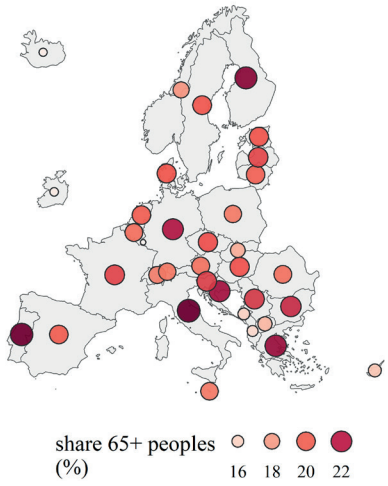
The cartogram method is most frequently used for visualising statistical characteristics and provides a quantitative presentation of the intensity of a given phenomenon on a map within the boundaries of adopted reference areas (e.g., country, statistical region, province, county, or commune). It is assumed to occurs with the same intensity in the entire reference area (Figure 1.25). Cartograms may present both relative characteristics (e.g., population density, unemployment rate, average area of a new dwelling, or density of the railway network) and absolute characteristics (population, number of unemployed, number of dwellings completed, or length of railway line). Cartograms can be drawn using a single colour, where the intensity of the phenomenon is reflected by the intensity of the colour, or a graphical scale.

Figure 1.25. Fertility rates (%) at chosen NUTS 2 regions in 2020



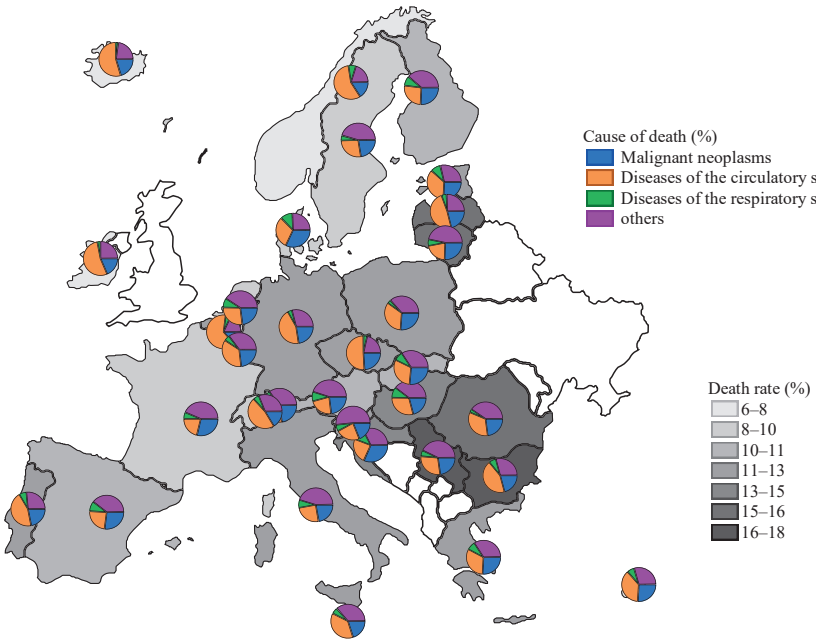
Source: Eurostat (demo_r_frate2), accessed 21/02/2024.

Figure 1.26. Share of people aged 65+ in the population of selected European countries in 2022



Source: Eurostat (demo_r_pjanaggr3), accessed 22/02/2024.

Figure 1.27. Death rate and main causes of death in selected European countries in 2020



Source: own study Eurostat (tps00029, hlth_cd_asdr2), accessed 23/02/2024.

The cartodiagram method is the second most common quantitative method for representing spatial characteristics. Here, data are displayed on a physical map using diagrams, such as bubbles (Figure 1.26), bars, and other geometric figures, as well as pie charts (Figure 1.27). Diagrams can represent a point (e.g., the size of a natural reserve), a line (e.g., the number of airline connections) or a surface (e.g., entities newly registered in provinces).

It is common to combine different methods of presenting geographical characteristics, particularly the cartogram method and the cartodiagram method (Figure 1.27). Characteristics presented this way should be thematically related so that relationships between them to be immediately identified.

1.5. Statistical study

All the activities that a statistician performs during their work are called a **statistical study**. These activities include planning, data collection, analysis, formulating conclusions, and the publication and sharing of the results. A statistical study is carried out to detect statistical regularities concerning the phenomena of interest (Weinbach, Grinnell 2007; Hozer 1998). In such a study, different types of regularities may be detected (Section 1.3): regularities in the structure of the phenomenon; regularities in the relationships between the given phenomenon and other phenomena in time or space and, finally, regularities in the dynamics of the phenomenon.

Planning a study involves defining the study's objective and formulating a research hypothesis. Researchers may formulate one or more objectives as well as one or more hypotheses. Often, there is a main objective and several specific objectives. A similar situation applies to hypotheses; the researcher may formulate a main hypothesis and several sub-hypotheses that comprise it. At the planning stage of the study, the factual, spatial, and temporal scope of the statistical population are defined, i.e., determining who or what, where, and when will be studied. It is also decided what sources of information will be used, how data will be collected, and what the costs will be. Although this is an initial stage, it is also determined how and where the results from the statistical study will be applied.

Data collection involves gathering of statistical data that describe the items that make up the studied statistical population. There are two types of statistical data: primary and secondary. **Primary data** refers to data collected specifically for a particular statistical study, whereas **secondary data** is data collected for other purposes

but used by the researcher in the research process. An example of primary data is information from a census conducted by the Statistical Offices of European countries, while an example of secondary data is information provided by a selected City Hall about local inhabitants' registered place of residence, which was used in a statistical study. An interesting combination of data sources occurring during the 2011 census in Poland. It was the first census to combine administrative data sources with a survey sampling method (Gołata, Dehnel 2021).

Regarding the scope of observation, statistical studies can be either complete (full) or partial. As the name implies, a **full study** covers all statistical units within the population, whereas a **sample study** considers selected units from the whole population. A sample study is conducted when the population is very large or difficult to count, when the cost of the study is too high, when the study would take too long, or when it would be destructive. An example of a full study is a census of the population of a selected country, while an example of a sample survey would be an analysis of the relationship between the length of employment and leave-of-absence among 500 randomly selected employees from a construction company.

Statistical studies, whether full or sample, are divided by duration into continuous, periodic, and ad hoc. In **continuous studies,** phenomena are observed continuously, such as the records of births, deaths, and marriages in the civil status register. **Periodic studies** are conducted systematically at specific intervals. An example would be a census conducted periodically every 10 years. **Ad hoc studies,** on the other hand, are carried out on a one-off basis to obtain the information needed. Many studies are ad hoc; for example, an analysis of the value of women's work in a selected year of observation. An ad hoc study could also include an assessment of the effectiveness of an investment, such as the dredging of the fairway at the Szczecin and Świnoujście Ports.

Full studies, i.e., those covering all items in a population, are carried out using a variety of methods. These include census, inventory, current registration, and statistical reporting (Hozer 1998). The **census** covers all units within the surveyed population, most often humans. However, other units may also be censused, e.g., farms, dwellings, or farm animals. The second method, **inventory**, is a process that aims to establish the condition and structure of the assets of the surveyed unit. This could be a company operating in the market or an administrative unit, e.g., a housing association. Another method, **current registration**, involves the continuous recording of specific facts, for example, births, deaths, and marriages. The last method, **reporting**, concerns the mandatory presentation of data by businesses to national

statistical offices or other relevant institutions. This information is submitted in the form of reports and declarations.

Sample studies are subdivided according to the method employed to select the sample. There are three methods: representative, monographic, and survey. In the **representative method,** a sample is selected at random to represent the whole population. Thus, a random sample is obtained by drawing lots, i.e., only the probability decides which element of the total population will be included in the sample, and which will not (Domański et al. 2014). The statistician examines the sample and, based on it, draws conclusions about the whole population. This is referred to as statistical inference. Inference involves two procedures: estimation and verification. Estimation refers to assessing the values of a parameter for the whole population based on results from a random sample. Verification involves formulating hypotheses about the whole population and testing their validity based on results from a random sample.

In the **monographic method,** one or more units are deliberately selected for further observation. The selected unit should be representative, i.e., one that occurs most frequently in the population. It can also be a leading unit that indicates the direction of the phenomenon's progress and is therefore worth investigating. An example of a study using the monographic method is the Report on the State of the City of Warsaw, published in 2022.

The survey method involves collecting information through a questionnaire. By filling in the questionnaire, the respondent provides answers to specific questions. Answers may be given orally, in writing, or electronically, depending on the investigator's decision. There are two main types of questions: open questions and closed questions. A study based on the survey method is voluntary, meaning respondents are not obliged to answer. An example of a survey conducted by the Central Statistical Office of Poland is the 2013 Time-Use Survey, which aimed to analyse how individuals and households allocate their time.

Using one of these three methods, the researcher obtains statistical data. The data may be subject to errors and therefore needs to be controlled. **Statistical errors** are divided into systematic and random errors. The former, **systematic** errors, significantly distort the results of the study. They are unidirectional, causing over- or underestimation of the data collected. This type of error may be caused by the interviewer unintentionally influencing the respondents' answers. **Random** errors cancel out when a large number of observations is made, since they cause distortions that vary

in direction. The cause of random errors may be careless collection the data, resulting in an inappropriate measurement.

1.6. Statistical indicators

A statistical study is carried out based on quantitative variables, i.e., those that are expressed numerically, and qualitative variables, expressed in words. As regards quantitative variables, individual values may be absolute measures (metrics) or relative measures (indicators) (Nowak 2001; Markowicz 2012). **An absolute** measure is a **numerical representation** of a phenomenon expressed in a specific unit, for example, a salary in EUR or the floor area of a flat in square metres. A **relative measure** examines the studied phenomenon by describing the relationship between two related categories.

Relative measures in statistics are referred to as **indicators**. There are three types of statistical indicators: structure index, dynamic index, and intensity index. The **structure index** denotes the relationship of a part of a population to its whole. It is written:

$$w_s = \frac{n_i}{n} \cdot 100\% \tag{1.1}$$

where: w_S – structure index; n_i – count of a part of the population; n – total count of the population.

An example of a structure indicator is the share of women in the population; e.g., in 2022, it was 52% in Poland and 50% in Norway. This means that Polish men are slightly outnumbered by women, while in Norway they are not. Other examples of structure indicators are the proportions of people in pre-working age, working age, and post-working age in a selected country.

The dynamics indicator is used to study changes in a phenomenon over time. It describes the change in the magnitude of the phenomenon between two periods or moments under study. It is written as follows:

$$w_d = \frac{y_2}{y_1} \cdot 100\% \tag{1.2}$$

where: w_d – dynamic index; y_1 – value of the variable in the reference period (used as the basis for comparison); y_2 – value of the variable in the period under review.

An example of a dynamics indicator is a 20% increase in the production volume of a food manufacturing company in 2023 compared to 2015. In this case, the indicator would be 120%.

The third type of a statistical indicator is the **intensity indicator**. It represents the ratio of two quantities that have a logical relationship with each other. It is written:

$$w_n = \frac{n_1}{n_2} \cdot c \tag{1.3}$$

where: w_n – intensity indicator; n_1 – count of the first population; n_2 – count of the second population; c – 1, 10, 100, 1000 (a constant).

An example of an intensity indicator is the rate of new marriages. According to Eurostat, in 2021, Poland recorded 4.5 marriages per 1,000 people, which meant a higher intensity of the phenomenon than in the European Union as a whole. For all EU-27 countries, the rate was 3.9. A lower rate was seen in France at 3.2 marriages per 1,000 people. Relatively high values were reported in the Baltic states of Lithuania and Latvia at 6 marriages per 1,000 people (Eurostat data).

Other examples of intensity indicators include: Population density in persons per km^2 (122 persons per km^2 in Poland and as many as 26,507 per km^2 in Monaco in 2021); Business start-up rate (the ratio of the number of businesses started in a given period to the number of businesses operating at the end of the previous period; 8.19 in 2008 in Szczecin; Markowicz 2012); Work time distribution indicator (the ratio of housework to professional work time; 1.17 in 2013 in Poland; Hozer-Koćmiel 2013); Data quality indicators of intra-commonwealth trade (Markowicz, Baran 2021).

Example 1.1

Based on selected demographic variables describing the situation in the Czech Republic contained in Table 1.15, determine:

a) population structure indicators by gender in 2022,

b) population growth rate in 2022 compared to 2013,

c) birth rate per 1,000 population in 2021.

Table 1.15. Selected demographic variables describing the situation in the Czech Republic in 2013, 2021 and 2022

Year	Variable	Value
2013	total population	10,512,419
2021	total population	10,516,707
	number of births	101,299
2022	number of women	5,519,006
	number of men	5,308,523
	total population	10,827,529

Source: own compilation based on Eurostat data (variables demo_pjan and demo_fmonth, accessed 4/03/2024).

d) The structure indicators for women and men in 2022 were determined based on Formula (1.1):

for women: $w_{s(k)} = \frac{5,519,006}{10,827,529} \cdot 100\% = 51\%$,

for men: $w_{s(m)} = \frac{5,308,523}{10,827,529} \cdot 100\% = 49\%$.

This means that women in the Czech Republic accounted for 51% of the population in 2022, while men made up 49%.

The example is based on Eurostat data, which provides population data as of January 1 of a given year, corresponding to the population as of December 31 of the previous year. This assumption has been used in this example.

e) The growth rate was calculated by comparing the population in 2022 to that in 2013. For this purpose, Formula (1.2) was used:

$$w_d = \frac{10,827,529}{10,512,419} \cdot 100\% = 103.00\%$$

A growth rate of 103% means that the population increased by 3% in 2022 compared to 2013. To interpret the indicator in_d, subtract 100 from the indicator value: $w_d - 100$. If the difference is positive, as in this example, the studied phenomenon has increased, whereas a negative value would indicate a decrease.

f) The intensity indicator, specifically the rate of births per 1,000 people in the Czech Republic in 2021, was calculated using Formula (1.3):

$$w_n = \frac{101,299}{10,516,707} \cdot 1000 = 9.63$$

The birth rate in the Czech Republic in 2021 accounted for 9.63 births per 1,000 inhabitants. Other indicators of intensity, e.g. the mortality rate, marriage rate, or number of dwellings per 1,000 inhabitants, are calculated and interpreted in a similar way.

Chapter 2

Structure analysis

Finding regularities in the distribution of statistical variables requires a statistical description of the population via:

- measures of central tendency,
- measures of dispersion,
- measures of asymmetry (skewness),
- measures of kurtosis and concentration.

Structure analysis is conducted for quantitative variables presented in a detailed series, a point series, or an interval series. The detailed series is often regarded as a primary data source, as it contains raw data collected from a survey. The point and interval series are based on data from the detailed series – they are ordered versions of the detailed series (see more on statistical series in chapter 1.4).

Any structure analysis must be preceded by an analysis of the empirical distribution of the studied variable. This is done using a distribution series (point or interval) and an appropriate graph, such as a histogram.

2.1. Building an interval series and a histogram

Quantitative (discrete and continuous) variables collected in a detailed series are not easily readable – especially when the number of observations is large. For instance, it is difficult to find the minimum and maximum value of a variable or determine its distribution. In order to present a continuous variable in an interval series, it is necessary to order the values of the variable into a certain number of separate intervals (i.e., to discretise the variable). The advantage of the discretisation process is that it makes computation shorter, simplifies the notation of the classification rules and

permits the use of a dataset containing outliers and missing data (Gatnar, Walesiak 2011). The following discretisation methods have been presented in the literature:

- local and global methods (Chmielewski, Grzymala-Busse 1996; Dougherty, Kohavi, Sahami 1995),
- static and dynamic methods (Catlett 1991; Dougherty, Kohavi, Sahami 1995),
- supervised and unsupervised methods (Gatnar 2000),
- disjoint and non-disjoint methods (Yang, Webb 2002).

In further calculations, we will refer to unsupervised methods, which do not take into consideration the membership of objects to classes of another variable (context methods, however, take this information into account). Context-free methods allow a continuous variable to be divided into **intervals with equal length** (one of the simplest methods) and into intervals of equal counts of units.

The construction of an interval series requires several problems to be solved that are related to:

1. The number of class intervals (classes).
2. The span of class intervals.
3. The method by which the limits of class intervals are determined.

The number of class intervals can be determined by any of the formulae, where the number of classes k depends only on the size of the population – n (Heiler, Michels 1994; Hozer 1994):

$$k = \left[\sqrt{n}\right] \tag{2.1}$$

$$k = [1 + 3{,}322 \cdot \log n] \tag{2.2}$$

$$k = [5 \cdot \log n] \tag{2.3}$$

or it also considers a range of variable's values:

$$k = \left[\frac{x_{\max} - x_{\min}}{h}\right] \tag{2.4}$$

where (Scott, 1979):

$$h = 3.49 \cdot S(x) \cdot \sqrt[3]{n} \tag{2.5}$$

where: $S(x)$ is the standard deviation of the variable,
or (Freedman, Diaconis, 1981):

$$h = 2 \cdot (Q_{3.4} - Q_{1.4}) \cdot \sqrt[3]{n} \tag{2.6}$$

where $Q_{1.4}$ and $Q_{3.4}$ are the first and third quartiles of the variable, respectively.

Formulae (2.1)–(2.3) allow for the number of classes to be determined for all the continuous variables in the dataset simultaneously (the number of classes will be the same), while Formula (2.4) allows for obtaining a different number of classes for individual variables.

Table 2.1 summarises the number of classes for a population with a given number of units *n as* determined by Formulae (2.1)–(2.3).

Table 2.1. Number of class intervals determined by Formulae (2.1)–(2.3)

n	$\lceil\sqrt{n}\rceil$	$[1 + 3{,}322 \cdot \log n]$	$[5 \cdot \log n]$
30	5	5	7
40	6	6	8
50	7	6	8
100	10	7	10
150	12	8	10
200	14	8	11
300	17	9	12
500	22	9	13
1,000	31	10	15
1,500	38	11	15

Source: own study.

The span of the class interval (length) is determined by:

$$h = \frac{x_{max} - x_{min}}{k - 1} = \frac{R}{k - 1} \tag{2.7}$$

or

$$h = \frac{R}{1 + 3{,}322 \cdot \log n} \tag{2.8}$$

where: x_{max} – the highest value of the given variable; x_{min} – the lowest value of the given variable; $R = x_{max} - x_{min}$ – the span; k – the number of classes; n – the size of the population.

The span of the class intervals can also be determined based on the relationship:

$$h \approx \frac{R}{k} \tag{2.9}$$

whereby h is taken as an approximation rounded up so that $hk \geq R$ and with the accuracy with which the measurements were taken.

Class intervals should have equal spans, especially in the case of symmetric distributions. However, different spans of class intervals are acceptable when:

– classes are established for units that are reasonably similar in qualitative terms (e.g., age distribution of the population),
– most of the units have lower or higher values of the variable (e.g. distribution of villages by population).

The first and the last class intervals may be open-left and open-right, respectively, to avoid the intervals to which no unit will be assigned, the so-called empty intervals. The first interval will contain all units with the values of the variable below the upper limit of the open-left first interval (class label: "below x_{11}"), and the last interval will contain all units with values of the variable greater than the lower limit of the last, open-right interval (class label: "above x_{0i}") (cf. Tables 1.7–1.9 in Chapter 1.4.1).

In the next step, the **lower** x_{01} and **upper** x_{11} **endpoints of the first interval** must be determined:

$$x_{01} = x_{\min} - \frac{1}{2}h \tag{2.10}$$

$$x_{11} = x_{01} + h \tag{2.11}$$

The upper endpoint of the first interval is usually the lower endpoint of the second interval $x_{11} = x_{02}$, but the lower endpoints of the subsequent intervals can also be determined by Formula (2.12). The upper endpoint of the next interval is determined by adding the span of the interval h to the lower endpoint (Formula 2.13).

$$x_{0i} = x_{01} + (i-1) \cdot h \tag{2.12}$$

$$x_{1i} = x_{0i} + h \tag{2.13}$$

for $i = 1, 2, 3, ..., k$.

There are diverse ways of determining and presenting the endpoints of class intervals, depending on the type of quantitative variable:

1. For a continuous variable:
   – the upper endpoint of the preceding interval overlaps with the lower endpoint of the next interval. There can be no doubt about the assignment of units with values of the variable at upper endpoints, so it is necessary to determine the disjointness of classes and establish in which interval the endpoint value

belongs: to the preceding interval (right-closed intervals) or the following one (left-closed intervals) (see Example 2.1).

A child of 2 years of age will be classified in interval 2 if the interval is left--closed, and in interval 1 if it is right-closed. Left-closed intervals are used by default.

Example 2.1

Table 2.2. Interval series for a continuous variable; interval bounds overlap

Age of children in years x_{0i}–x_{1i}	Closure of interval bounds		Class interval midpoint $\dot{x}_i$	Interval span h_i
	left $<x_{0i}, x_{1i})$	right $(x_{0i}, x_{1i}>$		
0–2	$0 \leq x_1 < 2$	$0 < x_1 \leq 2$	$(0 + 2)/2 = 1$	$2 - 0 = 2$
2–4	$2 \leq x_1 < 4$	$2 < x_1 \leq 4$	$(2 + 4)/2 = 3$	$4 - 2 = 2$
4–6	$4 \leq x_1 < 6$	$4 < x_1 \leq 6$	$(4 + 6)/2 = 5$	$6 - 4 = 2$
6–8	$6 \leq x_1 < 8$	$6 < x_1 \leq 8$	$(6 + 8)/2 = 7$	$8 - 6 = 2$

Source: exemplary data own study.

– the upper endpoint of the preceding interval does not overlap with the lower endpoint of the next interval (see Example 2.2). Children who are under 2 years of age (e.g., aged 1 year 11 months and 20 days) will be classified in the first-class interval, while children from their second birthday onwards will be classified in the second interval.

Example 2.2

Table 2.3. Interval series for a continuous variable; interval bounds do not overlap

Age of children in years x_{0i}–x_{1i}	Interval bounds $<x_{0i}, x_{1i})$	Class interval midpoint $\dot{x}_i$	Interval span h_i
0–1	$0 \leq x_1 < 2$	$(0 + 2)/2 = 1$	$2 - 0 = 2$
2–3	$2 \leq x_1 < 4$	$(2 + 4)/2 = 3$	$4 - 2 = 2$
4–5	$4 \leq x_1 < 6$	$(4 + 6)/2 = 5$	$6 - 4 = 2$
6–7	$6 \leq x_1 < 8$	$(6 + 8)/2 = 7$	$8 - 6 = 2$

Source: exemplary data, own study.

2. For a discrete variable:
   – the upper endpoint of the preceding interval does not overlap with the lower endpoint of the following interval, to emphasise the discrete nature of the variable (see Example 2.3). One- or two-room flats will be assigned to the first-class interval, three- and four-room dwellings to the second, and so on.

Example 2.3

Table 2.4. Interval series for a discrete variable; interval bounds do not overlap

Number of rooms x_{0i}–x_{1i}	Interval bounds $<x_{0i}, x_{1i}>$	Class interval midpoint $\dot{x}_i$	Interval span h_i
1–2	$1 \le x_1 \le 2$	(1 + 2)/2 = 1.5	2 – 1 = 1
3–4	$3 \le x_1 \le 4$	(3 + 4)/2 = 3.5	4 – 3 = 1
5–6	$5 \le x_1 \le 6$	(5 + 6)/2 = 5.5	6 – 5 = 1

Source: exemplary data, own study.

For data collected in an interval series, it is necessary to determine the means and the spreads of the individual intervals, which are determined as follows:

1. For a continuous variable:
   – when the upper endpoint of the preceding interval overlaps with the lower endpoint of the next interval, the midpoint of the interval is equal to half of the sum of the lower and upper endpoints:

$$\dot{x}_i = \frac{x_{0i} + x_{1i}}{2} \tag{2.14}$$

   the span is the difference between the interval upper and lower endpoints (see Example 2.1):

$$h = x_{1i} - x_{0i} \tag{2.15}$$

   – when the upper endpoint of the preceding interval does not overlap with the lower endpoint of the following interval, the interval midpoint is equal to half of the sum of the two lower endpoints of the adjacent classes:

$$\dot{x}_i = \frac{x_{0i} + x_{0(i+1)}}{2} \tag{2.16}$$

   The span is determined as the difference between the lower endpoints of adjacent classes (see Example 2.2):

$$h = x_{0(i+1)} - x_{0i} \tag{2.17}$$

2. For the discrete variable:
   – the midpoint of the interval is half of the sum of the lower and upper class endpoints – Formula (2.14), and the span of the interval is calculated as the difference of the upper and lower endpoints – Formula (2.15) (see Example 2.3).

An open-ended first or last interval causes complications in determining the midpoint of the interval for these cases. An open-ended interval can be conventionally regarded as closed if the size of the interval does not exceed 5% of the size of the total population. To determine the lower endpoint of the first interval (or the upper endpoint of the last interval), the span of these intervals must first be determined. The span of the first interval is assumed to be equal to the span of the second interval, and the span of the last interval is assumed to be equal to the span of the preceding interval. For the interval bounds thus determined, the midpoint of the interval is calculated by Formula (2.14) or (2.16), depending on whether the class bounds overlap or not.

If the smallest and the largest value of the studied are known, then the lower endpoint of the open-ended first interval is equal to the smallest value $x_{01} = x_{\min}$, and the upper endpoint of the open-ended last interval is equal to the largest value $x_{1i} = x_{\max}$.

If the size of an open-ended interval does not exceed 1% of the size of the total population, there is no need to close the interval to determine its midpoint, as the interval can be ignored in further analyses.

The next step in constructing the interval series is to **count the number of units** with the value of the variable belonging to the subsequent intervals (example 2.4). If the first child was 3 years old, they were assigned to the second interval, and this is marked with a vertical line in the Unit Count column. The next child, aged 1, is classified in the first interval, which is then marked with a vertical line. The same procedure is repeated for all studied units. To simplify the counting, when 4 units are assigned to one interval (and there are four vertical lines in the Unit Count column), the fifth unit is marked with a horizontal line. Adding groups of five consecutive observations makes counting easier than using a series of vertical lines, especially when there are many observations. The summed number of units in successive intervals (interval counts) is recorded in the column n_i.

The partial counts n_i can be replaced by the **cumulative counts** $n_{sk,i}$. They are calculated by adding the counts of the subsequent intervals:

$$n_{sk,i} = \sum_{j=1}^{i} n_j \tag{2.18}$$

where: $i = 1, 2,..., k$; k – number of classes.

In the first interval, the cumulative count is equal to the partial count ($n_{sk,1} = n_1$), while in the second interval: the cumulative count is equal to the partial count of the first and second intervals ($n_{sk,2} = n_1 + n_2$). The counting should be continued in the same way for all the intervals. The last interval in the cumulative count should be equal to the total number of observations (n). The values presented in the column of cumulative counts answer the question: For how many units in the population the value of the variable is smaller than the upper endpoint of the class interval. For example: How many children under 4 years old were there? 14.

Example 2.4

Table 2.5. Age distribution of children in the study population

Age of children in years x_{0i}–x_{1i}	Unit counting	Number of children n_i	Cumulative count $n_{sk,i}$
0–2	𝍸	5	5
2–4	𝍸 ||||	9	14
4–6	𝍸 𝍸 |	11	25
6–8	|||	3	28
∑	×	28	×

Source: exemplary data.

The correctness of the grouping can be checked by calculating the sum of the absolute differences of the midpoints of the class intervals $\dot{x}_i$ and the arithmetic means of the values belonging to this interval $\overline{x}_i$. If the calculated sum is less than half of the interval span, the grouping is considered correct:

$$\sum_{i=1}^{k} \left| \dot{x}_i - \overline{x}_i \right| \leq \frac{1}{2} \cdot h \tag{2.19}$$

If the relation (2.19) is not fulfilled, the series would have to be rebuilt by changing either the number of classes or the span of the class intervals, or by correcting the lower endpoint of the first interval and/or the upper endpoint of the last interval.

Example 2.5

Food consumption expenditure of households in 31 European countries in 2022, expressed as a percentage of total expenditures (%) were as follows:

11.3; 18.9; 14.3; 10.3; 10.1; 17.7; 7.0; 15.7; 11.8; 12.1; 15.6; 13.2; 11.7; 17.7; 18.0; 8.2; 14.3; 11.6; 10.7; 8.9; 16.3; 16.4; 23.7; 12.6; 18.0; 10.8; 11.4; 10.1; 28.8; 27.3; 21.5

Source: Eurostat (nama_10_co3_p3), accessed 11/02/2024.

Task: Present the collected information in the form of a distribution series and check the correctness of the grouping. Calculate cumulative counts and interpret selected measurements.

The household consumption expenditure on food, as presented in the detailed series above, is a continuous variable. Presenting the variable in a disaggregated series for a continuous variable requires:

1. Determining the number of classes – Formula (2.1):
   Information on n = 31 countries was collected in the series, so the number of classes is:
$$k = \left[\sqrt{31}\right] = [5.6] \approx 5$$

2. Determining the span of the intervals – Formula (2.7):
$$h = \frac{7.0 - 28.8}{5 - 1} = 5.45 \approx 5$$

3. Determining the lower and upper endpoints of the first interval – Formulae (2.10)–(2.11):
$$x_{01} = 7.0 - \frac{1}{2} \cdot 5 = 4.5 \approx 5, \quad x_{11} = 5 + 5 = 10$$

4. Counts of units (countries) with a share of food expenditure belonging to consecutive intervals, assuming that the intervals are left-closed (Table 2.6). The table additionally calculates the cumulative counts ($n_{sk,i}$).

The correctness of the grouping was checked using equation (2.19): $2.29 \leq \frac{1}{2} \cdot 5$, hence: 2.29 ≤ 2.5. The fulfilment of the inequality means that the classification has been conducted correctly. The supporting calculations are shown in Table 2.7.

Table 2.6. Classification of units in an interval series

x_{0i}–x_{1i}	Units assigned to intervals	n_i	$n_{sk,i}$
5–10	7.0; 8.2; 8.9	3	3
10–15	10.1; 10.7; 10.8; 11.8; 12.1; 11.6; 12.6; 11.7; 13.2; 11.4; 14.3; 10.1; 14.3; 11.3; 10.3	15	18
15–20	17.7; 18.0; 16.4; 16.3; 15.6; 17.7; 15.7; 18.0; 18.9	9	27
20–25	23.7; 21.5	2	29
25–30	28.8; 27.3	2	31
∑	×	31	×

Source: own compilation based on data from Eurostat (nama_10_co3_p3), accessed 11/02/2024.

Table 2.7. Supporting calculations for Example 2.5

Class interval x_{0i}–x_{1i}	Interval midpoint $\dot{x}_i$	Average value of measurements $\overline{x}_i$	$\lvert\dot{x}_i - \overline{x}_i\rvert$
5–10	7.50	8.03	0.53
10–15	12.50	11.75	0.75
15–20	17.50	17.14	0.36
20–25	22.50	22.60	0.10
25–30	27.50	28.05	0.55
∑	×	×	2.29

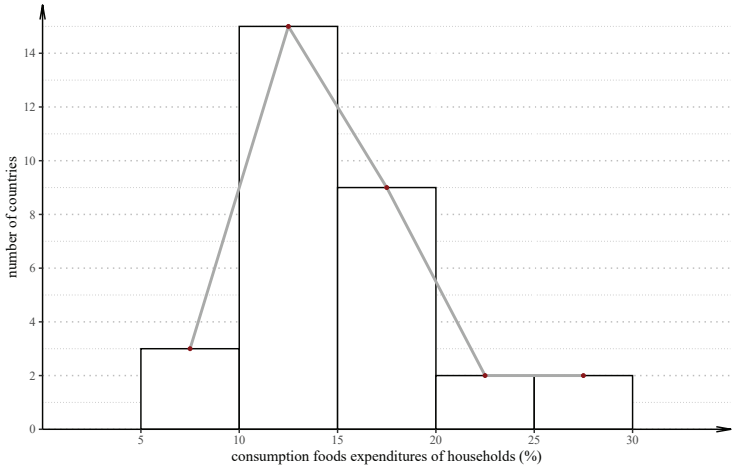
Source: own calculations.

Interpretation of the counts in Table 2.6:

- n_2 = 15 means that in 2022 there were 15 European countries where the share of household expenditure on food was between 10% and 15% of total expenditure,
- $n_{sk,4}$ = 29 means that in 2022 there were 29 European countries where the share of household food expenditure was less than 25% of total expenditure.

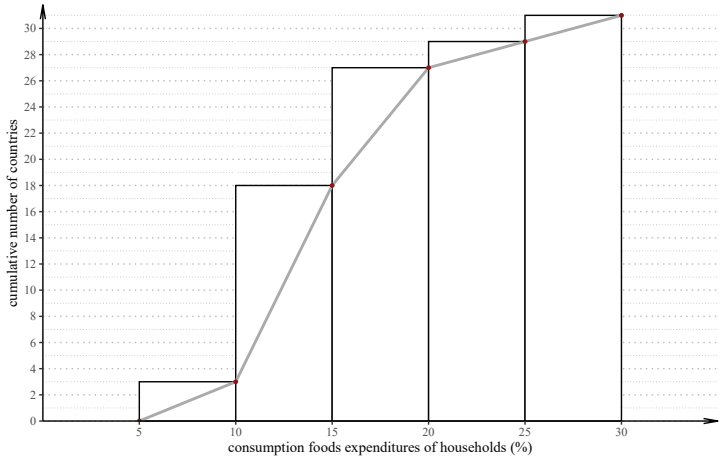
For an interval series, histograms – graphical representations of the distribution series – can be prepared. An **ordinary histogram** is composed of adjacent rectangles (bins), placed with their shorter base on the *x*-axis. The shorter base of the rectangle is equal to the interval span (*h*), and the height of the rectangle is the class count (relative frequency or percentage) (see Figure 2.1). Cumulative counts (cumulative frequencies or cumulative percentages) can also be represented on the graph. A graph constructed in this way is called a **cumulative histogram** (Figure 2.2).

Figure 2.1. Histogram and ordinary frequency polygon



Source: own elaboration based on Table 2.6.

Figure 2.2. Histogram and cumulative frequency polygon



Source: own elaboration based on Table 2.6.

Using an ordinary histogram, we can build a **frequency polygon**. It is obtained by joining the midpoints of the tops of the rectangles in the histogram. A frequency polygon is a polyline connecting the points with coordinates $(\dot{x}_i, n_i)$, i.e. the interval midpoint and the interval count. Frequencies, percentages or a density ratio (when the intervals have different spans) can be taken instead of the interval count.

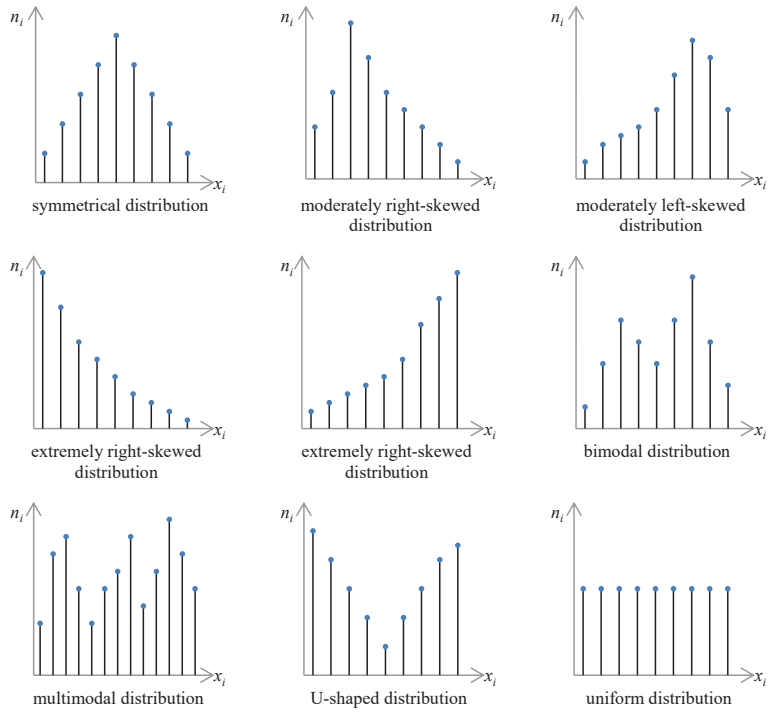
The cumulative histogram can be used to determine **the cumulative frequency polygon**. This is the line connecting the points with coordinates:

- $(x_{01}, 0)$ – lower endpoint of the first interval and 0,
- $(x_{1i}, n_{sk,i})$ – upper endpoints of subsequent intervals and the corresponding cumulative class counts (or cumulative frequencies, cumulative percentages, or cumulative density ratios).

The analysis of a variable's structure should be preceded by an analysis of its empirical distribution, as the type of distribution affects the choice of variables relevant to describe given population. The **empirical distribution** refers to the way in which counts are distributed across intervals. A distinction is made between empirical distributions for discrete and continuous variables, and these are divided into (Figures 2.3 and 2.4):

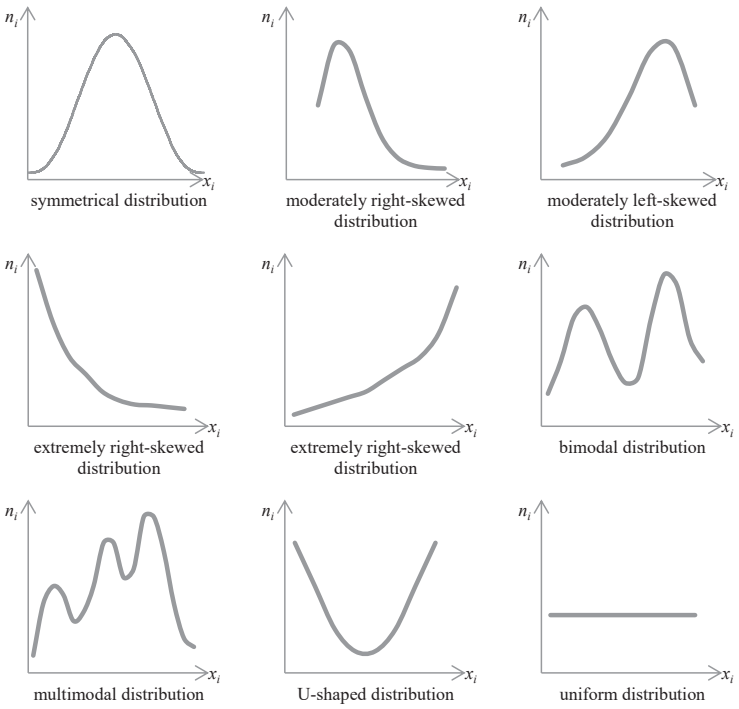
- symmetrical and skewed,
- u-shaped,
- unimodal and multimodal,
- uniform.

Figure 2.3. Types of empirical distributions for a discrete variable



Source: own elaboration based on Hozer 1994.

Figure 2.4. Types of empirical distributions for a continuous variable



Source: own elaboration based on Hozer 1994.

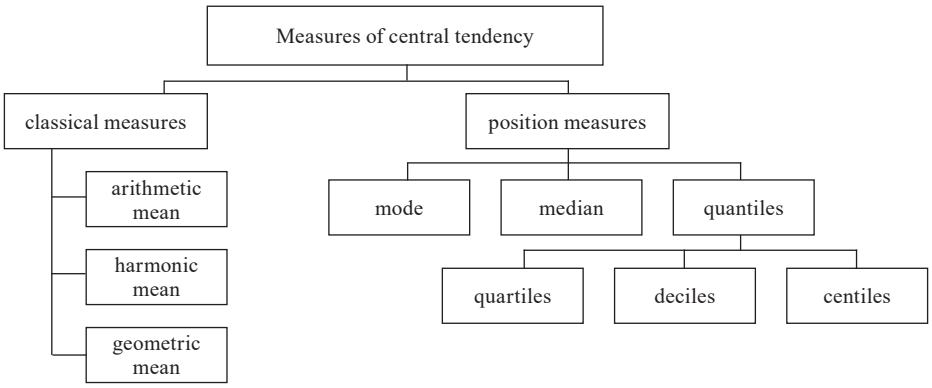
2.2. Measures of central tendency

The descriptive analysis of the structure regularity begins with determining **measures of central tendency,** which are divided into classical and position measures (Figure 2.5). Measures of central tendency are numerical characteristics assigned to appropriate units of the population and represent the resultant values of the examined characteristic of all units in a given population.

Classical measures are calculated from all values of the population, thus they are mutually exclusive. In a statistical study, only one of them is cognitively meaningful.

Positional measures, on the other hand, refer to different points in the distribution of a variable. They are complementary and not mutually exclusive, offering a description of the population from various perspectives.

Figure 2.5. Summary of measures of central tendency



Source: own study.

2.2.1. Arithmetic mean

The arithmetic mean is a measure of the average level of a variable in a population with a symmetric or near-symmetric distribution and with the following properties:

- It is calculated based on all values of the variable, so all values of the variable must be known.
- The variable must be additive.
- The statistical population should be relatively homogeneous.
- The mean is a denominated measure, i.e., it is expressed in units of the variable (kg, m, EUR).
- It satisfies the relation $x_{\min} < \overline{x} < x_{\max}$, where $x_{\min}$ and $x_{\max}$ denote the smallest and largest value of the variable in the data set respectively.
- The result of the arithmetic mean may correspond to one of the values of the variable, but more often it is an abstract value, not equal to any value of the variable.
- It is sensitive to extreme values of the variable's individual units.
- If class intervals (first or last) are open-ended, the arithmetic mean can be calculated:
  - when there are grounds for closing the class interval,
  - if the number of units contained in an open-ended interval is small (less than 1% of the total population), the interval may be disregarded, and the arithmetic mean calculated based on the remaining intervals only.

A distinction is made between the unweighted (simple) arithmetic mean and the weighted arithmetic mean. The **simple arithmetic mean** is calculated for data presented in a detailed series:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.20}$$

where:

x_i – values of the variable, for $i = 1, 2, ..., n$,

n – number of units.

Example 2.6

The following percentage of the population knowing one foreign language (Number of foreign languages spoken) were found in the countries of the Baltic Region[1] in 2022:

29.2; 41.4; 19.3; 26.3; 38.3; 47.3; 20.4; 50.8

Source: Eurostat (edat_aes_l26), accessed 12/02/2024.

Task: Calculate the average proportion of the population speaking one foreign language in the studied countries.

The data on the share of the population speaking one foreign language is presented in a detailed series, so Formula (2.20) must be applied to determine the average share:

$$\overline{x} = \frac{29.2 + 41.4 + 19.3 + 26.3 + 38.3 + 47.3 + 20.4 + 50.8}{8} = 34.13$$

On average, in 2022, 34.13% of the population in the Baltic States spoke one foreign language.

The **weighted arithmetic mean** is calculated for data presented in:

– An ungrouped frequency distribution – then the weights of the variable X are the values n_i, which represents the counts of individual values of the characteristic x_i:

[1] Denmark, Germany, Estonia, Latvia, Lithuania, Poland, Finland, Sweden.

$$\overline{x} = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_k n_k}{n} = \frac{\sum_{i=1}^{k} x_i n_i}{n} \tag{2.21}$$

where k – the number of variants of the variable.

– A grouped frequency distribution – in such case the weights of the variable X are the values of n_i – the counts of individual classes which are expressed by the midpoints of the intervals $\dot{x}_i$:

$$\overline{x} = \frac{\dot{x}_1 n_1 + \dot{x}_2 n_2 + \ldots + \dot{x}_k n_k}{n} = \frac{\sum_{i=1}^{k} \dot{x}_i n_i}{n} \tag{2.22}$$

where k – the number of bins of the variable.

Example 2.7

Based on 2011 Population and Housing Censuses, in 27 European countries families were counted by the number of members and then a summary was prepared (Table 2.8).[2]

Table 2.8. Distribution of families by size in selected European countries in 2011 (in millions) with supporting calculations

Number of persons in family	Number of families (in millions)	$x_i\, n_i$
x_i	n_i	
2	69.91	139.82
3	35.91	107.73
4	27.77	111.08
5	7.29	36.45
6	1.62	9.72
7	0.38	2.66
8	0.12	0.96
9	0.04	0.36
10	0.02	0.20
11	0.01	0.11
∑	143.07	409.09

Source: own compilation based on Eurostat data (cens_11fts_r3), accessed 12/02/2024.

Task: What was the average number of persons in a European family in 2011?

[2] In the book, when source data and supporting calculations are tabulated, the former are highlighted with an additional frame.

The data in Table 2.8 is presented as a point series, so Formula (2.21) is used to calculate the arithmetic mean:

$$\bar{x} = \frac{409.09}{143.07} = 2.86$$

In selected European countries in 2011, the average number of family members was 2.86.

Example 2.8

The distribution of live births (in thousands) by mother's age in the EU-27 in 2021 is shown in Table 2.9. Calculate the average mother's age.

Table 2.9. Age structure of women giving birth in 2021 in the EU27 with supporting calculations

Mother's age (years) x_{0i}–x_{1i}	Live births (thousands) n_i	$\dot{x}_i$	$\dot{x}_i n_i$	$\frac{n_i}{n}$	$\dot{x}_i \frac{n_i}{n}$
10–15	1.66	12.5	20.75	0.0004	0.0050
15–20	82.69	17.5	1,447.08	0.0202	0.3535
20–25	396.96	22.5	8,931.60	0.0971	2.1848
25–30	1,050.19	27.5	28,880.23	0.2569	7.0648
30–35	1,458.71	32.5	47,408.08	0.3568	11.5960
35–40	865.96	37.5	32,473.50	0.2118	7.9425
40–45	214.91	42.5	9,133.68	0.0526	2.2355
45–50	15.83	47.5	751.93	0.0039	0.1853
50–55	1.36	52.5	71.40	0.0003	0.0158
∑	4,088.27	×	129,118.23	1.0000	31.5832

Source: own compilation based on Eurostat data (demo_fordagec), accessed 12/02/2024.

Age is a continuous variable, and the structure of births is shown as a grouped frequency distribution. To determine the average age of a woman giving birth, all intervals must be closed-ended. This condition is met.

The midpoints of the intervals are estimated as in Example 2.3 using Formula (2.22):

$$\bar{x} = \frac{129,118.23}{4,088.27} = 31.58$$

The average age of a woman giving birth in 2021 in the EU-27 was 31.58.

The arithmetic mean can also be determined based on relative frequencies $\left(\frac{n_i}{n}\right)$, which are considered weights instead of the counts (n_i). The formula for the arithmetic mean takes the form (for a point series, instead of the interval midpoints $\dot{x}_i$ the variable's variances x_i should be used):

$$\overline{x} = \sum_{i=1}^{k} \dot{x}_i \frac{n_i}{n} \tag{2.23}$$

Hence (last column in Table 2.9):

$$\overline{x} = 31.5832 \approx 31.58$$

Regardless of the method of calculation, the value of the arithmetic mean remains the same, but in the latter case, the figures had to be more precise; therefore, the calculations were made to the nearest ten-thousandths.

The arithmetic mean is characterised by the following properties:

– When the sum of the deviations of a variable's values from the arithmetic mean is 0, a relationship occurs:

$$\sum_{i=1}^{n} \left(x_i - \overline{x}\right) = 0 \qquad \text{for detailed series} \tag{2.24}$$

$$\sum_{i=1}^{k} \left(x_i - \overline{x}\right) n_i = 0 \qquad \text{for point series} \tag{2.25}$$

$$\sum_{i=1}^{k} \left(\dot{x}_i - \overline{x}\right) n_i = 0 \qquad \text{for binned series} \tag{2.26}$$

– The sum of the squares of the deviations of the variable's values from the mean value is minimal:

$$\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2 = \min \qquad \text{for detailed series} \tag{2.27}$$

$$\sum_{i=1}^{k} \left(x_i - \overline{x}\right)^2 n_i = \min \qquad \text{for point series} \tag{2.28}$$

$$\sum_{i=1}^{k} \left(\dot{x}_i - \overline{x}\right)^2 n_i = \min \qquad \text{for binned series} \tag{2.29}$$

– The value of the arithmetic mean does not depend on the absolute size of the weights but on their proportion:

x_{1i}	n_{1i}
1	6
5	10
10	9
∑	25

x_{2i}	n_{2i}
1	60
5	100
10	90
∑	250

$$\overline{x}_1 = \frac{1 \cdot 6 + 5 \cdot 10 + 10 \cdot 9}{25} = 5.84 \qquad \overline{x}_2 = \frac{1 \cdot 60 + 5 \cdot 100 + 10 \cdot 90}{250} = 5.84$$

$$\overline{x}_1 = \overline{x}_2$$

– The arithmetic mean of a population divided into subgroups is equal to the weighted average of the arithmetic means in the subgroups, where the weights are the sizes of the subgroups:

$$\overline{x} = \frac{\sum_{i=1}^{r} \overline{x}_i n_i}{n} \tag{2.30}$$

where:

$\overline{x}_i$ – arithmetic mean of the i-th subgroup (i = 1, 2, ..., r),

n_i – numerosity of the i-th subgroup.

Example 2.9

The average number of employees and number of companies in Denmark, Sweden, and Norway in 2021 are shown in Table 2.10. Calculate the average number of employees in companies in the Nordic countries in 2021.

Table 2.10. Average number of persons employed and number of companies in Denmark, Sweden and Norway in 2021

Specification	Average number of employees $\overline{x}_i$	Number of companies n_i
Denmark	9.94	291,195
Sweden	7.58	873,865
Norway	7.73	433,358

Source: own calculations based on Eurostat data (sbs_sc_ovw), accessed 12/02/2024.

Table 2.10 contains the arithmetic means of the number of people employed in Nordic companies and the size of each group (i.e. the number of companies). On average, the largest average number of employed persons was in Danish companies (9.94), while Swedish and Norwegian companies had a similar average number of employees (7.58 and 7.73 respectively). To calculate the average number of persons employed in companies in the Nordic countries combined, Formula (2.30) should be used:

$$\overline{x} = \frac{9.94 \cdot 291{,}195 + 7.58 \cdot 873{,}865 + 7.73 \cdot 433{,}358}{291{,}195 + 873{,}865 + 433{,}358} = 8.05$$

The average number of employees in the Nordic companies in 2021 was 8.05.

2.2.2. Harmonic mean

The harmonic mean is a measure that is rarely used in the analysis of economic phenomena and as such, has become obsolete and is sometimes erroneously replaced by the arithmetic mean. **The harmonic mean** is indispensable for calculating the variable's average value that is not an observed physical quantity, but it is the ratio of two different quantities correlated in a logical manner $\left(x = \frac{y}{z}\right)$ – i.e., it is an indicator. The harmonic mean is calculated to determine, for example:

- average fuel consumption (l/km),
- average vehicle speed (km/h),
- average output (items/hour),
- average population density (persons/km^2).

The harmonic mean is the reciprocal of the arithmetic mean of the reciprocals of the individual values of the units in a statistical population (Zając 1994).

The **harmonic mean** is estimated according to the formulae:

1. For data presented as a detailed series – **simple harmonic mean**:

$$\overline{x}_H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_i}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \tag{2.31}$$

2. For data presented as a point series – **weighted harmonic mean**:

$$\bar{x}_H = \frac{n_1 + n_2 + \ldots + n_k}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + \ldots + \frac{n_k}{x_k}} = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}} \tag{2.32}$$

$$\bar{x}_H = \frac{y_1 + y_2 + \ldots + y_k}{\frac{y_1}{x_1} + \frac{y_2}{x_2} + \ldots + \frac{y_k}{x_k}} = \frac{\sum_{i=1}^{k} y_i}{\sum_{i=1}^{k} \frac{y_i}{x_i}} \tag{2.33}$$

If the characteristic x is expressed by the ratio y/z, then to determine the average value of x, we use:

- a harmonic mean when y is constant or known,
- an arithmetic mean when z is constant or known.

The harmonic mean is only calculated for positive quantities greater than 0.

Example 2.10

a) The driver was carrying a load from Świecko to Berlin. On the way there, there was little traffic on the road, and the average speed of the lorry was 100 km/h. On the return journey, however, the traffic was considerably heavier, and the average speed of the lorry was 80 km/h.

Task: At what average speed did the driver cover the route from Świecko to Berlin and back?

The average speed of the lorry is determined by the formula:

$$\text{vehicle speed}(x) = \frac{\text{distance}(y)}{\text{time}(z)}$$

In the case, the distance (y) in both directions was the same, i.e., y was constant. To determine the average speed of the car, we use Formula (2.31):

$$\bar{x}_H = \frac{2}{\frac{1}{100} + \frac{1}{80}} = \frac{2}{\frac{4}{400} + \frac{5}{400}} = \frac{2}{\frac{9}{400}} = 2 \cdot \frac{400}{9} = \frac{800}{9} = 88\frac{8}{9} \approx 88.89$$

The average speed at which the driver drove to Berlin and returned to Świecko was 88.89 km/h.

The correctness of the calculations was checked as follows:

The route from Świecko to Berlin is 100 km. On the way to Berlin, the driver spent 1 hour in transit, while on his way back: 1 hour and 15 minutes (1.25 h). On the way to Berlin the driver drove at a speed of:

$$x_p = \frac{100\text{ km}}{1\text{ h}} = 100\text{ km/h}$$

while on the way back:

$$x_n = \frac{100\text{ km}}{1.25\text{ h}} = 80\,\text{km/h}$$

That is, he covered the 200 km route in 2 hours and 15 minutes (2.25 h), giving an average speed of:

$$x = \frac{200\text{ km}}{2.25\text{ h}} = 88.89\text{ km/h}$$

If the arithmetic mean was used in this case, the average travel speed from Świecko to Berlin and back would be:

$$\overline{x} = \frac{x_p + x_n}{2} = \frac{100 + 80}{2} = 90\,\text{km/h}$$

which is not the correct value.

b) The driver covered two routes by car, each in 2 hours. The first route was covered at a speed of 100 km/h, the second route with a speed of 80 km/h. At what speed did the driver cover both routes?

Here, the constant is the time (z). In this case, the arithmetic mean must be used to calculate the average speed of the car on the two routes:

$$\overline{x} = \frac{100 + 80}{2} = 90\text{ km/h}$$

The average speed of the car on both routes was 90 km/h.

Checking the calculations:

Route I: Speed $x_{\text{I}} = 100$ km/h, in 2 hours the car travelled 200 km.

Route II: Speed $x_{\text{II}} = 80$ km/h, in 2 hours the car travelled 160 km.

In total, the driver travelled 360 km in 4 hours, i.e., the driver travelled at a speed of:

$$x = \frac{360}{4} = 90\text{ km/h}$$

Example 2.11

At Amazon's warehouse near Dobroviz in the Czech Republic, three workers were packing one box each. Lucie did it in 0.4 minutes, Honza in 0.125 minutes, and Adam in 0.2 minutes.

Task: How long did it take on average to pack one box? How many boxes in total did the workers pack in one shift?

The time to pack one box (x) is calculated as the working time (y) divided by the number of completed boxes (z):

$$x = \frac{y}{z}$$

One shift lasts 8 hours (480 min) and 3 workers are under observation (y = const.), hence:

$$x = \frac{3 \cdot 480}{z}$$

To calculate the average time spent by three workers to pack a box, we use Formula (2.31):

$$\overline{x}_H = \frac{3}{\frac{1}{0.4} + \frac{1}{0.125} + \frac{1}{0.2}} = 0.193548 \approx 0.19$$

The average time spent by three workers packing one box was about 0.19 minutes.

The total number of boxes packed by three workers during one shift will be:

$$z = \frac{3 \cdot 480}{0.193548} = 7440 \text{ boxes}$$

To test the correctness of the calculations, it was checked how many boxes each of the three workers packed in one hour and during one shift (Table 2.11).

Table 2.11. Supporting calculations for Example 2.11

Employee	Average time to pack 1 box (min)	Number of boxes packed in 1 hour (pcs)	Number of boxes packed during 1 shift (pcs)
Lucie	0.400	150	1,200
Honza	0.125	480	3,840
Adam	0.200	300	2,400
		∑	7,440

Source: own calculations.

Example 2.12

Task: Calculate the average population density in Hungary based on the data in Table 2.12.

Table 2.12. Population size and density in the NUTS3 regions of Hungary in 2022

Region	Population size (thousands) y_i	Population density (persons/km^2) x_i
Közép-Magyarország	3.032	446.6
Dunántúl	2.918	82.0
Alföld és Észak	3.739	76.9

Source: own calculations based on Eurostat data (demo_r_d3dens), accessed 12/02/2024.

Population density (x) is an indicator calculated as the ratio of population (y) to area (z). Since y is known and z is different and unknown for each region, the harmonic mean must be used – Formula (2.33):

$$\overline{x}_H = \frac{3032 + 2918 + 3739}{\frac{3032}{446.6} + \frac{2918}{82} + \frac{3739}{76.9}} = \frac{9689}{91} = 106.47$$

The average population density in Hungary in 2022 was 106.47 persons/km^2.

2.2.3. Geometric mean

The geometric mean is used to measure the average level of a variable when the variable is multiplicative. It is particularly applicable to variables expressed as indicators. These can be observed over time, i.e., they are presented in a time series. A more extensive explanation of the use of this mean is provided in Chapter 4.

One of the conditions to be satisfied by an average measure for a multiplicative variable is that the product of the values of the studied variable remains unchanged. That means that replacing one of the variable values with the average value does not change the product of those variable values. Other properties of the geometric mean are:

- It is only applicable for values greater than zero.
- If one of the observed variables is equal to 0, then the geometric mean will also be equal to 0.
- It is less sensitive to extreme values than the arithmetic mean.

The geometric mean is expressed by the formula:

$$\overline{x}_G = \sqrt[k]{x_1 \cdot x_2 \cdot \ldots \cdot x_k} = \sqrt[k]{\prod_{i=1}^{k} x_i} \tag{2.34}$$

When successive variants of a variable occur at different frequencies (n_i), these frequencies become the weights, and the geometric mean is determined by the formula:

$$\overline{x}_G = \sqrt[\sum_{i=1}^{k} n_i]{x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_k^{n_k}} = \sqrt[n]{\prod_{i=1}^{k} x_i^{n_i}} \tag{2.35}$$

It should be emphasised that each of the means outlined above indicates an average magnitude of the phenomenon. The choice of the appropriate measure depends on the type of variable, as in statistical surveys only one of the classical measures can be used.

Example 2.13

Table 2.13 shows the change in prices of selected food groups in the second half of 2023 (monthly data) in the EU-27.

Task: Calculate what was the average change in prices of selected products over the studied period.

Table 2.13. Monthly price dynamics of selected food products in the second half of 2023 (previous month = 100)

Selected groups	Months					
	07	08	09	10	11	12
Bread and cereals	100.21	99.65	99.86	100.11	100.07	100.02
Meat	100.38	100.23	99.85	100.06	100.03	99.94
Coffee, tea, and cocoa	100.22	100.05	100.07	99.8	100.02	99.38
Beer	100.04	100.30	100.24	99.92	100.47	99.41

Source: own calculations based on Eurostat data (prc_fsc_idx), accessed 12/02/2024.

Price dynamics is a multiplicative variable, expressed as the ratio of the price over the studied period to the previous year (or base year – more in Chapter 4.1). The change in the price of meat in July 2023 was calculated by the formula:

$$_M x_{07-2023} = \frac{_M p_{07-2023}}{_M p_{06-2023}} \cdot 100 = \frac{199.76}{199.00} \cdot 100 \approx 100.38$$

where p is the price of the product.

This means that meat was 0.38% more expensive in July 2023 than June 2023 (100.38 – 100 = 0.38).

To determine the average rate of change in the prices of selected food products, we use the geometric mean Formula (2.34):

$$_B \overline{x}_G = \sqrt[6]{100.21 \cdot 99.65 \cdot 99.86 \cdot 100.11 \cdot 100.07 \cdot 100.02} = 99.99$$

$$_M \overline{x}_G = \sqrt[6]{100.38 \cdot 100.23 \cdot 99.85 \cdot 100.06 \cdot 100.03 \cdot 99.94} = 100.10$$

$$_C \overline{x}_G = \sqrt[6]{100.22 \cdot 100.05 \cdot 100.07 \cdot 99.8 \cdot 100.02 \cdot 99.38} = 99.92$$

$$_B \overline{x}_G = \sqrt[6]{100.04 \cdot 100.3 \cdot 100.24 \cdot 99.92 \cdot 100.47 \cdot 99.41} = 100.06$$

Interpretation:

For meat and beer, during the studied period (July–December 2023):

- meat was on average 0.1% more expensive month-on-month (100.10 – 100 = 0.10),
- beer was 0.06% higher month-on-month (100.06 – 100 = 0.06).

For bread and cereals and coffee, tea, and cocoa:

- bread and cereals recorded a slight average price decrease of 0.01% (99.99 – 100 = −0.01),
- coffee, tea, and cocoa recorded an average price decrease of 0.08% (99.92 − 100 = −0.08).

2.2.4. Mode

Mode – or most frequently occurring value or modal value – is the value of a variable that occurs most frequently in the given population. The mode is the value of the variable of the largest count. It is a positional measure, which means that it is not the resultant of all values of the variable, but its value depends on the counts of the

adjacent intervals, their span, and the distribution of the variable. The mode can be determined:

- directly from the detailed or point statistical series,
- graphically based on a histogram drawn up for a continuous or discrete variable presented in an interval series,
- analytically when the variable is represented as a binned series.

Depending on the type of series in which the observed variable is presented, the mode is determined by the formula:

1. For a detailed series:

$$D = x_i \tag{2.36}$$

where x_i is the most frequent variant of the variable.

2. For a point series:

$$D = x_i \tag{2.37}$$

where x_i is the characteristic variant of the largest count.

3. For a binned series:

$$D = {}_0x_D + \frac{n_D - n_{D-1}}{\left(n_D - n_{D-1}\right) + \left(n_D - n_{D+1}\right)} \cdot h_D \tag{2.38}$$

where:

${}_0x_D$ – the lower endpoint of the interval with a mode,
n_D – the count of the interval with a mode,
$n_{D\text{-}1}$ – the count of the interval preceding the interval with a mode,
n_{D+1} – the count of the interval following the interval with a mode,
h_D – the span of the interval with a mode.

To determine the mode, the following conditions must be met:

- The distribution of the variable must be unimodal and moderately skewed.
- The interval with the largest count must not be the first or the last, as that would indicate a U-shaped or extremely skewed distribution of the variable.
- The span of the interval with the mode and the spans of the adjacent intervals should be the same.

If this is not the case, Formula (2.38) is applied, but instead of the interval counts, the density of counts is used, i.e., the ratio of the interval count to its span:

$$f_i = \frac{n_i}{h_i}$$

The formula for determining the mode then becomes:

$$D = {}_0x_D + \frac{f_D - f_{D-1}}{(f_D - f_{D-1}) + (f_D - f_{D+1})} \cdot h_D \tag{2.39}$$

where:

${}_0x_D$ – the lower endpoint of the interval with a mode,
f_D – the density of the interval with a mode,
$f_{D\text{-}1}$ – the density of the interval preceding the interval with a mode,
f_{D+1} – the density of the interval following the interval with a mode,
h_D – the span of the interval with a mode.

Example 2.14

The number of theatres operating in selected Belgian cities in 2021:

97; 42; 31; 27; 10; 2; 4; 5; 16; 4; 1; 9; 4

Source: Eurostat (urb_ctour), accessed 12/02/2024.

Task: Calculate what was the most common number of theatres in a single city in Belgium.

The data is presented in a detailed series. Therefore, Equation (2.36) applies: $D = 4$. In 2021, the most common number of theatres in a single Belgian city was 4.

Example 2.15

Table 2.14 shows the structure of the number of foreign languages known (self-reported) by the Danish population in 2022.

Table 2.14. Distribution of the number of foreign languages known by the Danish population in 2022

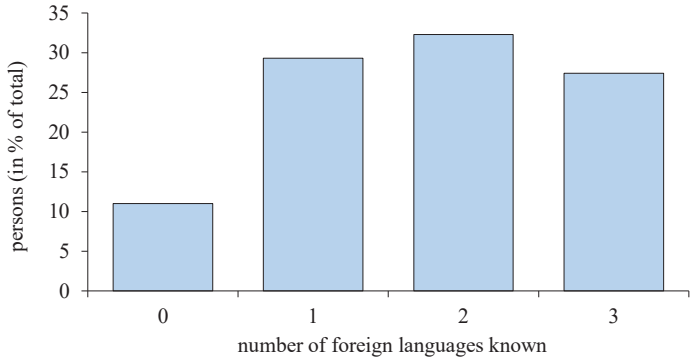
	Number of foreign languages x_i	Population share (%) n_i
	0	11.0
	1	29.3
D	2	32.3
	3	27.4

Source: Eurostat (edat_aes_l21), accessed 12/02/2024.

Task: Determine the most common number of foreign languages spoken by Danish residents in 2022.

The determination of the mode begins with assessing the distribution of the variable presented in a point series. For this purpose, Figure 2.6 was created. The variable is characterised by a unimodal distribution with moderate skewness, so it is possible to determine the mode.

Figure 2.6. Number of foreign languages known by the Danish population in 2022



Source: own elaboration based on Table 2.14.

Having used Formula (2.37), we found out that the mode is equal to the variant of the characteristic to which the highest count corresponds:

$$D = 2$$

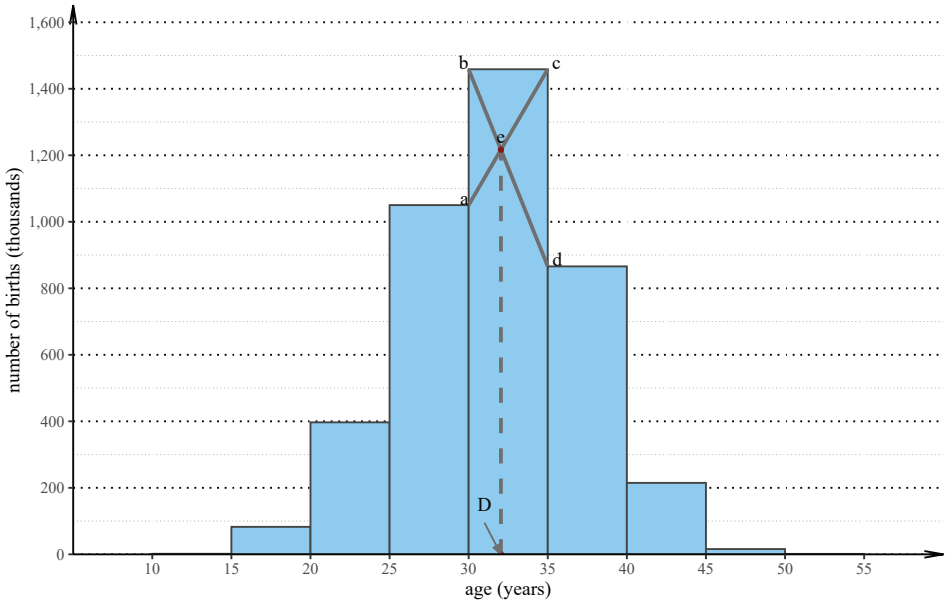
In 2022, the knowledge of two foreign languages was dominant in the Danish population.

Example 2.16

Using the data in Table 2.9 (Example 2.8) on the number of children born in the EU-27 in 2021 by mother's age, determine the modal value of mothers' ages. Determine the mode graphically and analytically.

To determine the mode graphically, a histogram of the variable should be created – Figure 2.7. The distribution is unimodal and moderately skewed. Therefore, the most frequent value can be established.

Figure 2.7. Graphical delineation of the mode – age distribution of EU-27 women who gave birth in 2021



Source: own elaboration based on Table 2.9.

The largest number of women who gave birth in 2021 in the EU27 were aged between 30 and 35, i.e., the dominant age meets inequality: $30 \leq D < 35$.

The graphical determination of the mode consists of finding the abscissa of the point (point e) of intersection of the lines connecting the points with the coordinates:

- The upper endpoint of the modal interval (c) and the upper endpoint of the preceding interval (a) with corresponding abundances.
- The upper endpoint of the modal interval (c) and the upper endpoint of the preceding interval (a) with the corresponding counts.

Thus, the coordinates of the points are:

- Point a (30; 1,050.19),
- Point b (30; 1,458.71)
- Point c (35; 1,458.71),
- Point d (35; 865.96).

Figure 2.7 shows a graphical determination of the modal age of women giving birth in 2021.

In an analytical manner, the dominant age of women who give birth in 2021 is determined by formula (2.38):

$$D = 30 + \frac{1458.71 - 1050.19}{(1458.71 - 1050.19) + (1458.71 - 865.96)} \cdot 5 = 32.04$$

The dominant age of women giving birth in 2021 in the EU-27 was 32.04 years.

Example 2.17

Task: What was the most common job-seeking time for women and for men in Poland during the 4th quarter of 2020?

Table 2.15. Distribution of time spent job-seeking by unemployed men and women in 4th quarter of 2020 in Poland

Job-seeking time (months) x_{0i}–x_{1i}	Men (thousands) n_{Mi}	Women (thousands) n_{Ki}
0–2	50	55
2–4	72	53
4–7	68	49
7–13	60	43
13–25	25	23
25–34	16	18

Source: own elaboration based on TABL. 21 (180) Bezrobotni według miejsca zamieszkania i czasu poszukiwania pracy w IV kwartale 2020r – by BAEL. *Rocznik Statystyczny Rzeczypospolitej Polskiej* 2021, 259.

The analysis of the counts of male job search intervals indicates the presence of a modal value in the second interval: from 2 months to 4 months, $n_{Mi} = 72$. However, the interval with the modal value, and the adjacent ones have different spans:

- $h_1 = 2$,
- $h_2 = 2$,
- $h_3 = 3$.

In this case, the count density of the variable (f_i) should be determined. The supporting calculations are summarised in Table 2.16.

Table 2.16. Supporting calculations for Example 2.17

	x_{0i}–x_{1i}	n_{Mi}	n_{Ki}	h_i	f_{Mi}	f_{Ki}
	0–2	50	55	2	25	27
D_M	2–4	72	53	2	**36**	26
	4–7	68	49	3	22	16
	7–13	60	43	6	10	7
	13–25	25	23	12	2	1
	25–34	16	18	9	1	2

Source: own calculations based on Table 2.15.

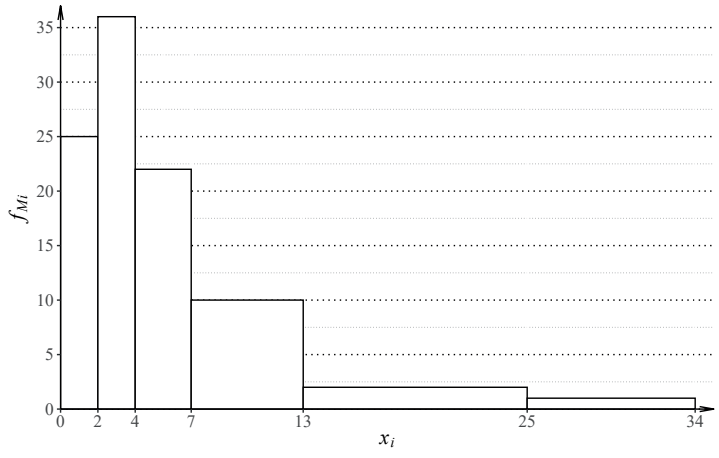
In the next step, job-seeking time density histograms were produced for men (Figure 2.8) and women (Figure 2.9).

The analysis of the density distribution of men's job-seeking time counts indicates a unimodal distribution, and the mode occurs in the second interval, i.e., the mode is $2 \leq D < 4$. The mode of men's job-seeking time is determined based on Equation (2.39) and Table 2.16:

$$D_M = 2 + \frac{36 - 25}{(36 - 25) + (36 - 22)} \cdot 2 = 2.8$$

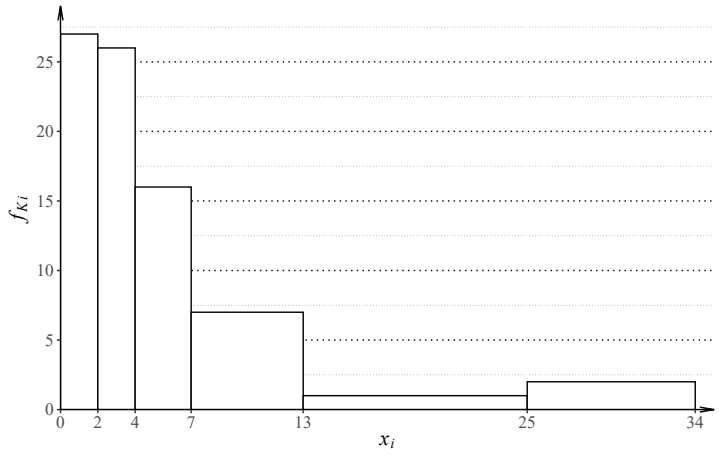
The most frequent time the unemployed Polish men spent on seeking jobs in Q4 2020 was 2.8 months.

Figure 2.8. Distribution of men's job-seeking time count density in Q4 2020 in Poland



Source: Table 2.16.

Figure 2.9. Distribution of the women's job-seeking time count density in Q4 2020 in Poland



Source: Table 2.16.

The distribution of the density of counts of job-seeking time by unemployed women (Figure 2.9) is a U-shaped/uniform distribution (the count of the first interval is greater than the count of the next interval, and the count of the last interval is greater than the count of the preceding interval). It is not possible to determine the most frequent value.

2.2.5. Median

Similarly to the mode, the **median** is an average positional measure, and is determined based only on certain units of the study population that are selected according to their position in this population.

The median is a particular kind of quantile; it is the middle value of an ordered population (from the smallest values to the largest or from the largest to the smallest). Half of the observations in a given population take a value less than or equal to the median, and half take a value greater than or equal to the median. The median acts as the equivalent of the arithmetic mean among positional measures.

The median can be determined directly from the series (detailed or point series), graphically, or analytically (in the case of an interval series). For a variable presented in a detailed series, the position of the median depends on the population size:

– when n is odd, the median is equal to the median value of the observation:

$$M = x_{\frac{n+1}{2}} \tag{2.40}$$

– when n is even, the median is the arithmetic mean of the two middle values:

$$M = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} \tag{2.41}$$

For data ordered in a point series, the median is calculated using the following formulas:

– when n is odd:

$$M = x_{\frac{n+1}{2}} \tag{2.42}$$

– when n is even:

$$M \approx x_{\frac{n}{2}} \tag{2.43}$$

For a continuous or discrete variable ordered in a binned series, the median should be determined by the formula:

$$M = {}_0x_M + \frac{\frac{n}{2} - n_{sk,i-1}}{n_M} \cdot h_M \tag{2.44}$$

where:

${}_0x_M$ – the lower endpoint of the interval containing the median,

$n_{sk,i-1}$ – the cumulative count of the interval preceding the interval containing the median,

n_M – the count of the interval containing the median,

h_M – the span of the interval containing the median,

n – size of the population.

Example 2.18

Task: Based on the data in Example 2.14 on the number of theatres in selected Belgian cities in 2021, determine the median value of the number of theatres.

The data are presented in a detailed series. The series must first be ordered (usually in ascending order):

1; 2; 4; 4; 4; 5; **9**; 10; 16; 27; 31; 42; 97

The number of observations is odd ($n = 13$), hence, following Formula (2.40):

$$M = x_{\frac{13+1}{2}} = x_7 = 9 \tag{2.40}$$

The median is equal to the value of the seventh observation in the ordered series, i.e., it equals 9.

Interpretation: In 50% of the surveyed Belgian cities, the number of theatres did not exceed 9 (less than or equal to), and in 50% of the cities the number of theatres was no less than 9 (greater than or equal to).

Example 2.19

Task: Using the data in Table 2.14 (Example 2.15), determine the median number of foreign languages declared to be spoken by Danish residents in 2022.

The determination of the median should start with ordering the statistical series. The data are presented in a point series, so in this case, the ordering is the equivalent of determining the count of cumulated observations (see Table 2.17).

Table 2.17. Supporting calculations for Example 2.19

	x_i	n_i	$n_{sk,i}$
	0	11.0	11.0
	1	29.3	40.3
M	2	32.3	72.6
	3	27.4	100.0
	$\sum$	100.0	×

Source: Table 2.14.

Since the size of the population is an even number, the median is determined by Formula (2.43):

$$M \approx x_{\frac{100}{2}} = x_{50} = 2$$

The median value is equal to the value of the observation numbered 50. This observation should be found in the column with the cumulative counts. Observation nº 50 falls in the third class (all observations with consecutive numbers from 41 to 73 belong to this class), so $M = 2$, which is read from the column where the variants of the variable (x_i) are recorded.

Interpretation: Half of the Danes surveyed declared knowledge of at most two foreign languages (two or fewer) in 2022, while the other half declared speaking at least two foreign languages (two or more).

The median value, unlike the arithmetic mean, does not depend on the marginal values of the variable, meaning that the median is not sensitive to outliers (atypical values). When the variable is characterised by high variation and significant skewness, it is recommended to apply the median instead of the arithmetic mean. On the other hand, when the distribution of a variable is symmetric or only slightly skewed, the arithmetic mean is preferred, as all observations are included in its calculation.

The median can be calculated even when not all observations are known; for instance, it can be estimated from open-ended interval series. The intervals do not necessarily have to be of the same span, but the smaller the class intervals, the greater the accuracy of the median.

If the middle observation ($n/2$) is in the first interval of an interval series, then, to determine the median, the interval must be close-ended. In thiscase, the cumulative count of the interval preceding the interval containing the median is 0 ($n_{sk,i-1} = 0$).

2.2.6. Quantiles

Quantiles are a group of measures that report the values of a variable at specific points within a population. They are positional measures, applied in ordered statistical series, and can be determined:

- directly from the statistical series – detailed or point ones,
- graphically on the basis of a cumulative frequency polygon drawn up for a continuous or discrete variable presented in a resolution series,
- analytically.

The most commonly used quantiles are: quartiles, deciles, and centiles (percentiles). The general formulae for the analytical determination of quantiles are as follows:

1. For detailed and point series:

$$Q_{k.m} \approx x_{\frac{k \cdot n}{m}} \tag{2.45}$$

where:

k – quantile number,

m – the number of parts into which the total population is divided,

n – the size of the population.

2. For binned series:

$$Q_{k.m} = {}_0x_{Q_{k.m}} + \frac{\frac{k \cdot n}{m} - n_{sk,i-1}}{n_{Q_{k.m}}} \cdot h_{Q_{k.m}} \tag{2.46}$$

where:

${}_0x_{Q_{k.m}}$ – lower endpoint of the quantile interval,

$n_{sk,i-1}$ – the cumulative count of the interval preceding the interval with the quantile,

$n_{Q_{k.m}}$ – the count of the interval with the quantile,

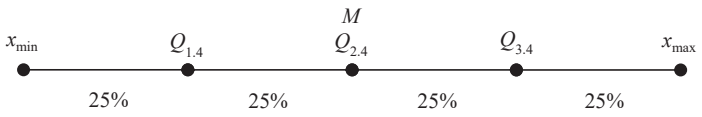
$h_{Q_{k.m}}$ – the span of the interval with the quantile,

Other designations are as in Formula (2.42).

In economics, the most common measure are the **quartiles**. They divide the population into four parts, with 25 per cent of the population between each quartile (Figure 2.10):

- Below the value of the first quartile (sometimes called the lower quartile), $Q_{1.4}$ accounts for 25% of the population, with 75% in the quartiles above it.
- Quartile 2 is equal to the median $Q_{2.4} = M$, dividing the population into two equal parts.
- Below the value of the third (upper) quartile, $Q_{3.4}$, there is 75% of the population, with 25% remaining above it.

Figure 2.10. Graphical representation of the division of the population into quartiles



Source: own study.

Like the median, the value of quartiles is obtained by applying an interpolation formula that depends on the type of series in which the variable is presented and on the number of observations. Quartiles, like medians, are always determined from series of ordered values (ascending or descending). The first quartile is estimated from the following formulae (Bąk et al. 2019):

1. For a detailed series:
   – when n is divisible by 4:

$$Q_{1.4} = \frac{x_{\frac{n}{4}} + x_{\frac{n}{4}+1}}{2} \tag{2.47}$$

   – when $(n + 1)$ is divisible by 4:

$$Q_{1.4} = x_{\frac{n+1}{4}} \tag{2.48}$$

   – when $(n + 2)$ is divisible by 4:

$$Q_{1.4} = x_{\frac{n}{4}+0.5} \tag{2.49}$$

   – when $(n + 3)$ is divisible by 4:

$$Q_{1.4} = \frac{x_{\frac{n+1}{4}-0.5} + x_{\frac{n+1}{4}+0.5}}{2} \tag{2.50}$$

2. For a point series:

$$Q_{1.4} \approx x_{\frac{n}{4}} \tag{2.51}$$

3. For a binned series:

$$Q_{1.4} = {}_0x_{Q_{1.4}} + \frac{\frac{n}{4} - n_{sk,i-1}}{n_{Q_{1.4}}} \cdot h_{Q_{1.4}} \tag{2.52}$$

where the symbols are the same as in Formula (2.46).

We estimate the second quartile using the following formulae:

1. For a detailed series:
   – when n is divisible by 4:

$$Q_{3.4} = \frac{x_{\frac{3n}{4}} + x_{\frac{3n}{4}+1}}{2} \tag{2.53}$$

   – when $(n + 1)$ is divisible by 4:

$$Q_{3.4} = x_{\frac{3(n+1)}{4}} \tag{2.54}$$

   – when $(n + 2)$ is divisible by 4:

$$Q_{3.4} = x_{\frac{3n}{4}+0.5} \tag{2.55}$$

– when $(n + 3)$ is divisible by 4:

$$Q_{3.4} = \frac{x_{\frac{3(n+1)}{4}-0.5} + x_{\frac{3(n+1)}{4}+0.5}}{2} \tag{2.56}$$

2. For a point series:

$$Q_{3.4} \approx x_{\frac{3n}{4}} \tag{2.57}$$

3. For an interval series:

$$Q_{3.4} = {}_0x_{Q_{3.4}} + \frac{\frac{3n}{4} - n_{sk.i-1}}{n_{Q_{3.4}}} \cdot h_{Q_{3.4}} \tag{2.58}$$

Example 2.20

Based on the data in Example 2.14, determine the value of the first and second quartile of the number of theatres in selected Belgian cities in 2021.

The number of observations is $n = 13$. The condition of $(n + 3)$ divisible by 4 is met:

$$\frac{(13+3)}{4} = 4$$

so the lower quartile is determined from Equation (2.50):

$$Q_{1.4} = \frac{x_{\frac{13+1}{4}-0.5} + x_{\frac{13+1}{4}+0.5}}{2} = \frac{x_3 + x_4}{2}$$

The values of the third and fourth observations in the ordered series are respectively, 4 and 4, hence:

$$Q_{1.4} = \frac{4+4}{2} = 4$$

We estimate the third quartile with Formula (2.56):

$$Q_{3.4} = \frac{x_{\frac{3(13+1)}{4}-0.5} + x_{\frac{3(13+1)}{4}+0.5}}{2} = \frac{x_9 + x_{10}}{2}$$

Since $x_9 = 16$ and $x_{10} = 27$ (in an ordered series), we have:

$$Q_{3.4} = \frac{16+27}{2} = 22$$

Interpretation:

- In 25% of the selected Belgian cities in 2021, the number of theatres was no more than 4, and in 75% it was no less than 4 theatres.
- In 75% of the surveyed Belgian cities in 2021, the number of theatres was no more than 22, and in 25% it was no less than 22 theatres.

Example 2.21

Task: Based on the data in Example 2.15, calculate the first and the third quartile of the number of foreign languages declared to be spoken by Danish citizens in 2022.

The number of foreign languages spoken by the Danes is presented in a point series. Using Formulae (2.51) and (2.57), we determine the number of observations for which the values of the variable should be read:

$$Q_{1.4} \approx x_{\frac{100}{4}} = x_{25} \text{ i } Q_{3.4} \approx x_{\frac{3\cdot 100}{4}} = x_{75}$$

Observations number 25 and 75 are to be found in the column of cumulative counts (which is equivalent to the ordering of the values in the detailed series). The supporting calculations are summarised in Table 2.18.

Table 2.18. Supporting calculations for example 2.21

	x_i	n_i	$n_{sk,i}$
	0	11.0	11.0
$Q_{1.4}$	1	29.3	40.3
	2	32.3	72.6
$Q_{3.4}$	3	27.4	100.0
	$\sum$	100.0	×

Source: Table 2.14.

The first sought number relates to the second value of the variable, hence: $Q_{1.4} = 1$, while the second number relates to the fourth value, hence: $Q_{3.4} = 3$.

Interpretation:

- 25% of the Danes surveyed in 2022 declared that they spoke no more than 1 foreign language, while 75% declared that they knew at least one foreign language.
- In 2022, 75% of the surveyed Danes declared speaking no more than 3 foreign languages, and 25% declared knowledge of at least 3 foreign languages.

Example 2.22

Task: Using the data in Table 2.9 (Example 2.8) on the age of mothers giving birth in 2020 in the EU-27, determine:

(a) Graphically

(b) Analytically

the values of the first, second, and third quartiles.

(a) Graphical Determination of Quartiles

The graphical determination of quartiles should begin by plotting the cumulative frequency polygon. The coordinate points forming this line are the upper bounds of the intervals and the corresponding cumulative counts of the intervals (x_{1i}, $n_{sk,i}$), as shown in Table 2.19.

For example:

- (15; 1.66), (20; 84.35), (25; 481.31) etc. (the cumulative counts are shown in the right column of Table 2.19).

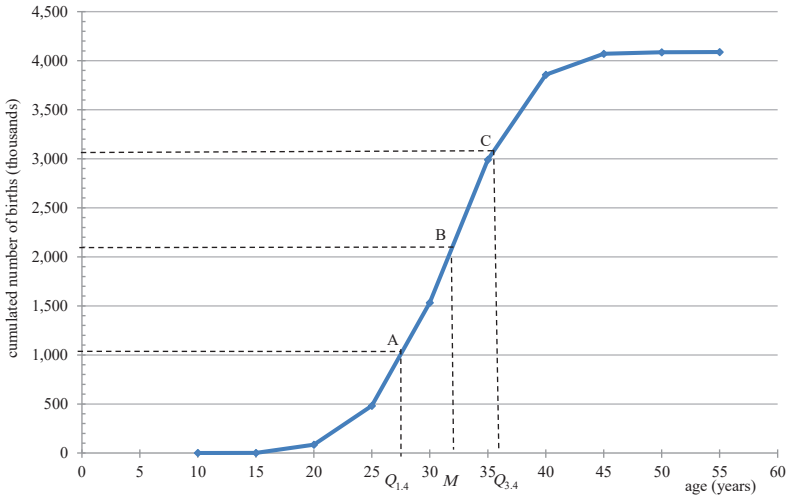
The exception is the starting point, whose coordinates are the lower bound of the first interval (x_{01}) and 0, i.e., (10; 0). The cumulative frequency polygon is shown in Figure 2.11.

Table 2.19. Supporting calculations for example 2.22

	x_{0i}–x_{1i}	n_i	$n_{sk,i}$
	10–15	1.66	1.66
	15–20	82.69	84.35
	20–25	396.96	481.31
$Q_{1.4}$	25–30	1,050.19	**1,531.50**
M	30–35	1,458.71	**2,990.21**
$Q_{3.4}$	35–40	865.96	**3,856.17**
	40–45	214.91	4,071.08
	45–50	15.83	4,086.91
	50–55	1.36	4,088.27
	∑	4,088.27	×

Source: Table 2.9.

Figure 2.11. Graphical presentation of the first, second and third quartiles



Source: own elaboration based on Table 2.19.

1. On the *y*-axis, mark the points for which the quartiles will be determined:

$$\frac{n}{4} = \frac{4{,}088.27}{4} = 1{,}022.07 \text{ for } Q_{1.4}$$

$$\frac{4{,}088.27}{2} = 2{,}044.14 \text{ for } M\,(Q_{2.4})$$

$$\frac{3 \cdot 4{,}088.27}{4} = 3{,}066.20 \text{ for } Q_{3.4}$$

Then, for these points, mark the segments parallel to the *x*-axis until they intersect with the cumulative frequency polygon – points A, B, and C. Points A, B, and C should be projected perpendicularly onto the *x*-axis, thus determining the values of the quartiles sought.

(b) Analytical Determination of Quartiles

The analytical determination of quartiles should begin with calculating the values of the cumulative frequencies – ($n_{sk,i}$) of the variable (see Table 2.19), followed by determining the bins in which the measures sought will be found.

1. For the first quartile, find the observation number:

$$\frac{4{,}088.27}{4} = 1{,}022.07$$

2. For the median:

$$\frac{4{,}088.27}{2} = 2{,}044.14$$

3. For the third quartile:

$$\frac{3 \cdot 4{,}088.27}{4} = 3{,}066.20$$

The observation 1,022.07 falls into the bin between 25–30, i.e., $Q_{1.4} \in \langle 25; 30)$, the observation 2,044.14 and M is in the bin $\langle 30; 35)$, and the observation 3,066.20 and $Q_{3.4}$ are in the bin $\langle 35; 40)$.

Calculations:

- The lower quartile is calculated by Formula (2.52):

$$Q_{1.4} = 25 + \frac{\frac{4{,}088.27}{4} - 481.31}{1{,}050.19} \cdot 5 = 27.57$$

- The median is calculated by Formula (2.44):

$$M = 30 + \frac{\frac{4{,}088.27}{2} - 1{,}531.50}{1{,}458.71} \cdot 5 = 31.76$$

- The upper quartile is calculated using Formula (2.58):

$$Q_{3.4} = 35 + \frac{\frac{3 \cdot 4{,}088.27}{4} - 2{,}990.21}{865.96} \cdot 5 = 35.44$$

Interpretation of the Measures:

- $Q_{1.4}$: 25% of EU-27 women giving birth in 2020 were no older than 27.57 (i.e. younger or of that age), and 75% of women were no younger than 27.57 (i.e. older or of that age).
- M: 50% of EU-27 women giving birth in 2020 were no older than 31.76 (i.e. younger or that age), and 50% of women were no younger than 31.76 (i.e. older or that age).

- $Q_{3.4}$: 75% of EU-27 women giving birth in 2020 were no older than 35.44 (i.e. younger or of that age), and 25% of women were no younger than 35.44 (i.e. older or of that age).

2.3. Measures of variability

Measures of central tendency, as discussed in Section 2.2, are not sufficient to adequately describe the observed variable. What is also important to know is the arrangement of individual units around these measures. The units may be clustered around the mean or be distant from it in varying degrees (Examples 2.23 and 2.24). Measures that are useful at this stage of statistical analysis are called **measures of variability** or **dispersion**.

Example 2.23

A developer built two blocks of flats, BI and BII. The flats had a different number of rooms, ranging from 1 to 7. The average number of rooms in both buildings was the same – 4 rooms ($\bar{x}_{BI} = 4$ i $\bar{x}_{BII} = 4$). Does this mean that the room structure in both buildings was the same?

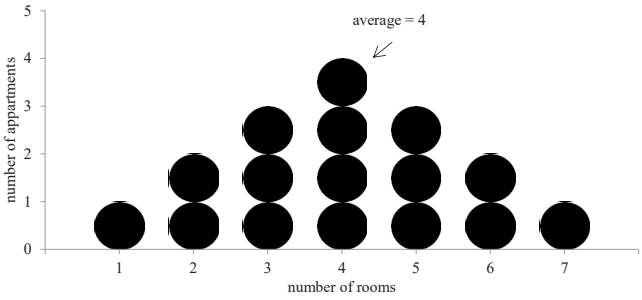
The structure of flats in the two buildings in terms of the number of rooms they contained is shown in Table 2.20, and a graphical representation of the distribution is shown in Figures 2.12 (BI) and 2.13 (BII).

Table 2.20. Structure of flats in Building I and II by number of rooms

Number of rooms x_i	Number of flats in BI n_{1i}	Number of flats in BII n_{2i}
1	1	4
2	2	3
3	3	
4	4	2
5	3	
6	2	3
7	1	4
∑	16	16

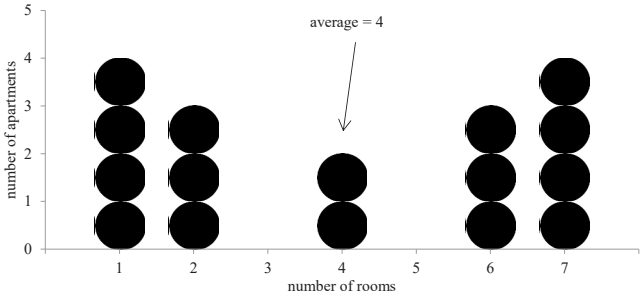
Source: exemplary data.

Figure 2.12. Structure of the number of rooms in BI



Source: Table 2.20.

Figure 2.13. Structure of the number of rooms in BII



Source: Table 2.20.

It becomes apparent that information about the arithmetic mean value for the studied variable is not sufficient. The same number of flats – 16 – were built in both buildings, and the median was the same, i.e., 4 rooms. However, the distribution of units around the mean in the two blocks of flats was different, and so was the structure of the two variables.

Example 2.24

The productivity of two shifts of seasonal workers picking strawberries was compared. Both shifts had the same number of workers – 12. The number of strawberry baskets picked by the workers on both shifts was:

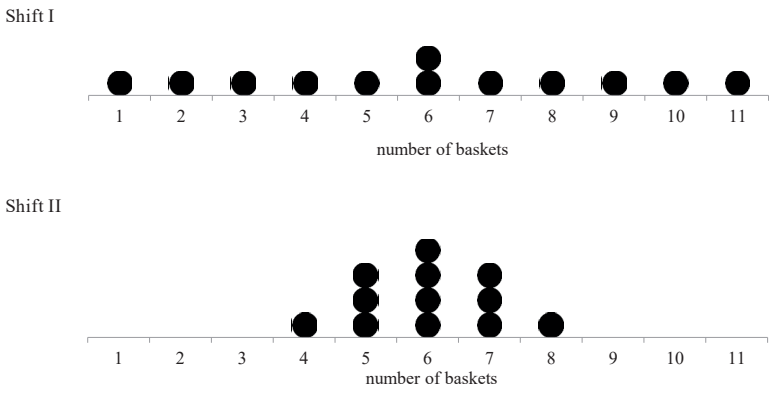
Shift I:	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11
Shift II:	4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8

Source: exemplary data.

The average measures of central tendency (mean, median, mode) of the number of baskets collected on both shifts were the same at 6. Did the workers work equally efficiently?

The distribution of the harvested strawberry baskets is shown in Figure 2.14.

Figure 2.14. Distribution of baskets harvested by the workers on two shifts



Source: exemplary data.

The arrangement of the variables around the mean reveals that on two shifts, the workers worked with varying productivity. On Shift I, the slowest worker filled only one basket of strawberries, while the fastest worker picked 11 baskets. On Shift II, the slowest worker picked 4 baskets, while the fastest worker picked 8 baskets.

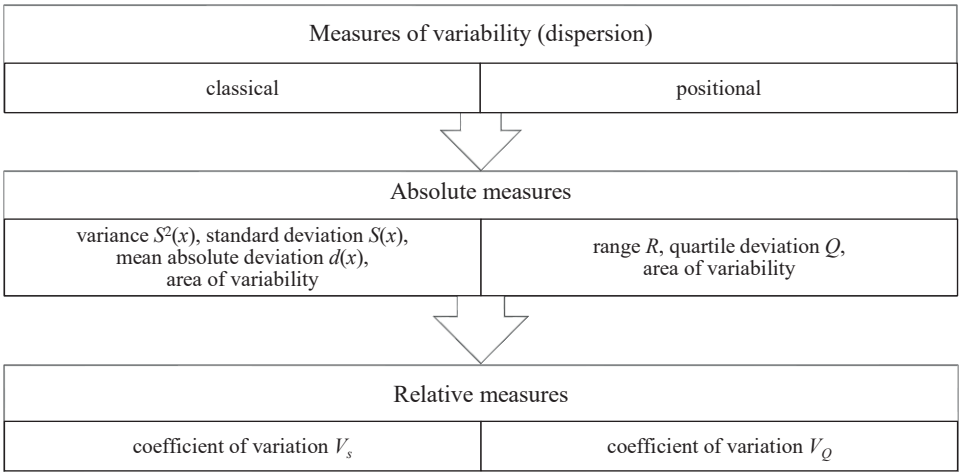
There are many measures of dispersion (Figure 2.15). They can be classified according to different criteria. The first is to examine the dispersion of units around the arithmetic mean – **classical measures** – or around the median – **positional measures**. Many classical measures have their counterparts among positional measures, e.g., standard deviation versus quartiles, areas of variation, or coefficients of variation.

Measures of variability can also be divided into **absolute** and **relative** measures:

- Absolute measures examine by how much units vary. They are expressed in units of measurement (i.e., in units of the observed variable). The applicability of these measures to assess the degree of variability of a given population is constrained to the same variable.
- Relative measures are expressed in dimensionless units (in this case, in percent), and they answer the question of how big the differences between specific units are.

Relative measures are widely used to compare the variability of not only different populations due to one selected variable, but also of a single population due to several variables.

Figure 2.15. Summary of variability measures



Source: own study.

2.3.1. Variance and standard deviation

Variance is a classical measure, which means that it examines the dispersion of units around the arithmetic mean. The basis for its calculation is the difference between the values of the variable of individual units and the arithmetic mean $\left(x_i - \overline{x}\right)$. This difference has a number of properties (see Formulae 2.24–2.29).

The variance is defined as the arithmetic mean of the square deviations of the individual variable values from the arithmetic mean of that variable. Depending on the statistical series in which the variable is ordered, the variance is determined by the formula:

– Detailed series:

$$S^2(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \tag{2.59}$$

– Point series:

$$S^2(x) = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i}{n} \tag{2.60}$$

– Binned series:

$$S^2(x) = \frac{\sum_{i=1}^{k}(\dot{x}_i - \bar{x})^2 n_i}{n} \tag{2.61}$$

The variance can also be determined by the formula:

$$S^2(x) = \overline{x^2} - \bar{x}^2 \tag{2.62}$$

where:

$\bar{x}$ – the arithmetic mean determined, depending on the type of series, using Formulae (2.20)–(2.22),

$\overline{x^2}$ – the average of the square values of the variable, determined by the formulas:

– for a detailed series:

$$\overline{x^2} = \frac{\sum_{i=1}^{n} x_i^2}{n} \tag{2.63}$$

– for a point series:

$$\overline{x^2} = \frac{\sum_{i=1}^{k} x_i^2 n_i}{n} \tag{2.64}$$

– for a binned series:

$$\overline{x^2} = \frac{\sum_{i=1}^{k} \dot{x}_i^2 n_i}{n} \tag{2.65}$$

The computation of variance for data presented as an interval series with equal interval spans may be subject to clustering error, resulting in overestimation. It is the midpoints of the intervals rather than the values of the arithmetic means of individual classes that are used to calculate the variance, because the arithmetic means of classes cannot be determined since the distributions within the individual classes are not known. Too small a number of intervals ($k < 12$) is often accompanied by too large interval spans, which results in greater overestimation of the variance. To reduce the effect of clustering error on the variance value, the **Sheppard correction** is applied (Yule, Kendall 1966). For interval series with 12 or more classes, the Sheppard correction is not applied. The formula for the corrected variance is as follows:

$$S_*^2(x) = S^2(x) - \frac{h^2}{12} \tag{2.66}$$

where:

$S_*^2(x)$ – corrected variance value,
$S^2(x)$ – the variance determined from Equation (2.61),
h – the span of the class interval.

If the population is divided into sub-groups (sub-populations), the variance of the total population can be determined from the variance of the sub-groups and their size. The variance of a population is expressed as the sum of the arithmetic mean of the intra-group variances of the variable value (the intra-group variance) and the variances of group means of this variable (the inter-group variance). This relationship is called **the variational equality**:

$$S^2(x) = \overline{S_j^2(x)} + S^2(\overline{x}_j) = \frac{\sum_{j=1}^{m} S_j^2(x) n_{.j}}{\sum_{j=1}^{m} n_{.j}} + \frac{\sum_{j=1}^{m} (\overline{x}_j - \overline{x})^2 n_{.j}}{\sum_{j=1}^{m} n_{.j}} \tag{2.67}$$

where:

$\overline{S_j^2(x)}$ – mean variance from intra-group variances (intra-group),
$S^2(\overline{x}_j)$ – variance between group mean values of a variable (inter-group),
$S_j^2(x)$ – the variance of the j-th group,
$\overline{x}_j$ – the arithmetic mean of the variable values in the j-th group,
$\overline{x}$ – the arithmetic mean of the total population,
$n_{.j}$ – the size of the j-th group.

Variational equality is the basis for variance analysis, which is justified when all groups are drawn from the same population and their statistical measures are not significantly different from each other.

The variance is a non-negative number ($S_2(x) \geq 0$). The value of the variance indicates the diversity of the population – the more dispersed the population, the higher the value of the variance. The variance is an absolute measure expressed in the squared variable unit (the square of the physical unit in which the variable is measured) and is therefore difficult to interpret. The root of the variance gives the **standard deviation**, which is expressed in the variable's original unit of measurement and has a clear statistical interpretation:

$$S(x) = \sqrt{S^2(x)} \tag{2.68}$$

Properties of the standard deviation:

- The standard deviation is calculated on the basis of all the units of the population and can be subjected to algebraic transformations.
- The value of the standard deviation will not change if the numbers in the series are expressed in relative terms (percentages).
- The value of the standard deviation will not change when the same number is added to or subtracted from all the observations in a series.
- If each observation in the series is multiplied (or divided) by the same positive number, the standard deviation will also increase (or decrease) by the same factor.

The standard deviation is a very common measure of dispersion. It indicates the dispersion of observations is in relation to the arithmetic mean, so it has statistical significance when the arithmetic mean is known. The standard deviation is also referred to as the mean error or the mean squared error.

Associated with standard deviation is the so-called three-sigma rule, which is based on Chebyshev's inequality:

$$\overline{x} - 3S(x) < x_i < \overline{x} + 3S(x) \tag{2.69}$$

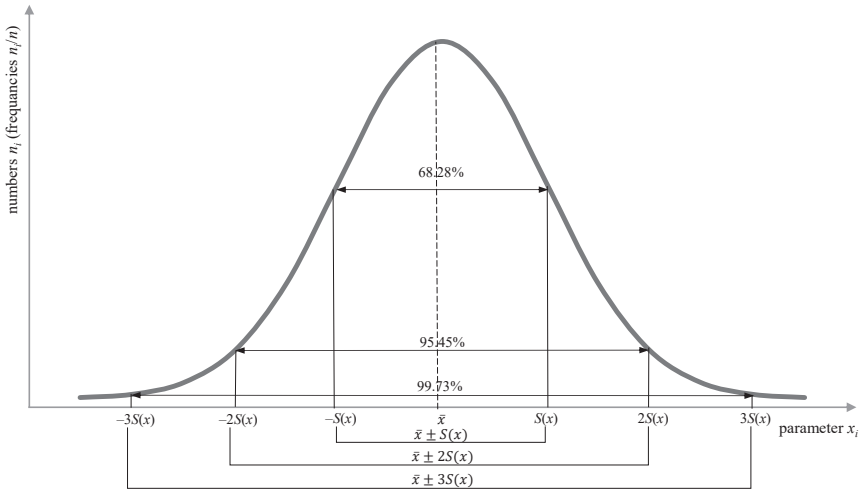
In the case of a normal or near-normal distribution (slightly skewed), only about 0.27% of the observations fall outside the interval determined by Formula (2.69):

- approximately 68.28% of observations do not differ from the arithmetic mean by more than ± one standard deviation,

– approximately 95.45% of observations do not differ from the arithmetic mean by more than ± two standard deviations,
– approximately 99.73% of observations do not differ from the arithmetic mean by more than ± three standard deviations.

A graphical interpretation of the three-sigma rule is shown in Figure 2.16. The three-sigma rule has practical applications, allowing for the identification of abnormal observations or anomalies for the observed variable.

Figure 2.16. Graphical interpretation of three-sigma rule



Source: own study.

Example 2.25

Using the data from Example 2.14, calculate the standard deviation of the number of theatres in selected Belgian cities in 2021.

The data about the number of theatres in Belgian cities are summarised in a detailed series. To calculate the variance and then the standard deviation, it is necessary to know the arithmetic mean, which is calculated by using Formula (2.20):

$$\bar{x} = \frac{97+42+31+27+10+2+4+5+16+4+1+9+4}{13} = \frac{252}{13} = 19.38$$

The variance of the number of theatres is calculated by Formula (2.59):

$$S^2(x) = \frac{(97-19.38)^2 + (42-19.38)^2 + \ldots + (9-19.38)^2 + (4-19.38)^2}{13} = 653.31$$

The variance is a measure expressed in the squared unit of the studied variable. In this case, it is (theatre^2) and this measure has no direct interpretation. What does have an interpretation is the standard deviation, calculated using Formula (2.68):

$$S(x) = \sqrt{653.31} = 25.56$$

The actual values of the number of theatres in selected Belgian cities differ from the mean number of theatres by ±25.56 facilities.

Example 2.26

Using the data in Example 2.7 (Table 2.8), calculate, basing on the standard deviation, the variation in the number of persons in families in selected European countries in 2011.

The standard deviation is calculated using Formula (2.68), while the variance for data presented in a point series is calculated using Formula (2.60). The supporting calculations are summarised in Table 2.21. For the variance calculation, the arithmetic mean is assumed to be 2.86.

Table 2.21. Supporting calculations for Example 2.26

x_i	n_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2 n_i$	x_i^2	$x_i^2 n_i$
2	69.91	–0.86	51.71	4	279.64
3	35.91	0.14	0.70	9	323.19
4	27.77	1.14	36.09	16	444.32
5	7.29	2.14	33.39	25	182.25
6	1.62	3.14	15.97	36	58.32
7	0.38	4.14	6.51	49	18.62
8	0.12	5.14	3.17	64	7.68
9	0.04	6.14	1.51	81	3.24
10	0.02	7.14	1.02	100	2.00
11	0.01	8.14	0.66	121	1.21
∑	143.07	×	150.73	×	1,320.47

Source: own calculations based on Table 2.8.

The variance is:

$$S^2(x) = \frac{150.73}{143.07} = 1.05$$

and the standard deviation is:

$$S(x) = \sqrt{1.05} = 1.02$$

The actual values of the number of persons in families in selected European countries in 2011 differed from the average value by ±1.02 persons.

The value of the variance was also calculated using Formula (2.62). To this end, the mean of the squared number of persons in families in selected European countries was calculated by Formula (2.64):

$$\overline{x^2} = \frac{1320.47}{143.07} = 9.23$$

The variance was:

$$S^2(x) = 9.23 - 2.86^2 = 1.05$$

and the standard deviation was:

$$S(x) = \sqrt{1.05} = 1.02$$

In several cases, there may be a difference between the variance values calculated with the two methods at the level of decimals. This is due to the approximations of numbers at each stage of the calculation. In this particular case, the standard deviation values are the same – 1.02 persons.

Example 2.27

The survey shows how the number of operating companies in selected European countries evolved in 2020. Information was collected for regions (NUT-1) in the former Eastern Bloc countries (B regions) and the so-called Western part of Europe (A regions). The distribution of the regions A and B by number of companies in operation in 2020 is shown in Table 2.22.

Calculate the variance and the standard deviation of the number of operating enterprises in regions A, B, and in the total of the regions of the surveyed countries.

Table 2.22. Structure of the number of businesses in operation (in thousands) in European countries in regions A and B in 2020

Number of companies (in thousands)	Number of regions	
	A	B
x_{0i}–x_{1i}	${}_An_i$	${}_Bn_i$
0–250	16	8
250–500	15	10
500–750	5	2
750–1,000	3	0
1,000–1,250	3	1
1,250–1,500	2	0
∑	44	21

Source: own compilation based on Eurostat (bd_hgnace2_r3), accessed 14/02/2024.

The number of companies is a discrete characteristic. However, due to a large number of variants (reaching more than 1.2 million), it is presented in an interval series. The calculation of the variance and the standard deviation should start with determining the average value. Formula (2.22) was used for this purpose and the supporting calculations are summarised in Table 2.23.

Table 2.23. Supporting calculations for Example 2.27

x_{0i}–x_{1i}	${}_An_i$	${}_Bn_i$	$\dot{x}_i$	$\dot{x}_i\,{}_An_i$	$\dot{x}_i\,{}_Bn_i$	$\left(\dot{x}_i - {}_A\overline{x}\right)^2 \cdot {}_An_i$
0–250	16	8	125	2,000	1,000	1,619,816.20
250–500	15	10	375	5,625	3,750	69,727.69
500–750	5	2	625	3,125	1,250	165,292.56
750–1,000	3	0	875	2,625	0	559,405.54
1,000–1,250	3	1	1,125	3,375	1,125	1,394,635.54
1,250–1,500	2	0	1,375	2,750	0	1,736,577.02
∑	44	21	×	19,500	7,125	5,545,454.55

Source: own calculations based on Table 2.22.

The number of operating companies in the regions of the West European countries (Group A) in 2020 was:

$${}_A\overline{x} = \frac{19{,}500}{44} = 443.18 \text{ thousand companies}$$

while in the regions of the Eastern Bloc B (Group B) the number was:

$${}_B\overline{x} = \frac{7{,}125}{21} = 339.29 \text{ thousand companies}$$

The variance of the number of firms operating in 2020 in the regions of the Western European countries (A) was calculated using Formula (2.61), and the standard deviation using Formula (2.68):

$$_{A}S^2(x) = \frac{5{,}545{,}454.55}{44} = 126{,}033.06\,,\quad _{A}S(x) = \sqrt{126{,}033.06} = 355.01$$

The actual number of operating companies in each region of the Western European countries in 2020 differed from the average number of companies by ±355.01 thousand companies.

The variance and standard deviation for Group B countries were determined by Formulae (2.62) and (2.68). The supporting calculations are included in Table 2.24. The mean of the squares of the variable was calculated by Formula (2.65):

$$_{B}\overline{x^2} = \frac{3{,}578{,}125}{21} = 170{,}386.90$$

The variance of the number of companies in Region B is:

$$_{B}S^2(x) = 170{,}386.90 - 339.29^2 = 55{,}269.2$$

and the standard deviation is

$$_{B}S(x) = \sqrt{55{,}269.2} = 235.09$$

The actual number of companies operating in each region of the Eastern European Bloc countries in 2020 differed from the average number of companies by ±235.09 thousand companies.

Table 2.24. Supporting calculations for Example 2.27 (Continued)

x_{0i}–x_{1i}	$\dot{x}_i$	$\dot{x}_i^2$	$\dot{x}_i^2 \cdot {}_{B}n_i$	n_i	$\dot{x}_i^2 \cdot n_i$
0–250	125	15,625	125,000	24	375,000
250–500	375	140,625	1,406,250	25	3,515,625
500–750	625	390,625	781,250	7	2,734,375
750–1,000	875	765,625	0	3	2,296,875
1,000–1,250	1,125	1,265,625	1,265,625	4	5,062,500
1,250–1,500	1,375	1,890,625	0	2	3,781,250
∑	×	×	3,578,125	65	17,765,625

Source: own calculations based on Table 2.22.

The variational equation (Formula 2.67) was used to determine the variance for the two groups of countries combined. To apply the variational equation, it is first necessary to determine the average number of companies operating in 2020 in the surveyed European regions combined. This is determined from the arithmetic means of the subgroups (Regions A and B) using Formula (2.30):

$$\overline{x} = \frac{443.18 \cdot 44 + 339.29 \cdot 21}{44 + 21} = 409.62$$

The intra-group variance is:

$$\overline{S_j^2(x)} = \frac{126{,}033.06 \cdot 44 + 55{,}272.11 \cdot 21}{44 + 21} = 103{,}171.83$$

and the intergroup variance:

$$S^2(\overline{x}_j) = \frac{(443.18 - 409.62)^2 \cdot 44 + (339.29 - 409.62)^2 \cdot 21}{44 + 21} = 2{,}360.44$$

The variance for the European regions of Group A and B countries combined is:

$$S^2(x) = 103{,}171.83 + 2{,}360.44 = 105{,}532.27$$

while the standard deviation is:

$$S(x) = \sqrt{105{,}532.27} = 324.86$$

To check the validity of the calculations, the variance for both groups combined was also determined by Formula (2.62). The mean of the squared numbers of companies in both groups combined was calculated by Formula (2.65). The supporting calculations are summarised in Table 2.24:

$$\overline{x^2} = \frac{17{,}765{,}625}{65} = 273{,}317.31$$

$$S^2(x) = 273{,}317.31 - 409.62^2 = 105{,}528.77$$

$$S(x) = \sqrt{105{,}528.77} = 324.85$$

The actual number of companies operating in 2020 in the regions of the European countries combined differed from the average number of companies by 324.86 (324.85) thousand companies.

2.3.2. Mean absolute deviation

The mean absolute deviation (MAD) is an absolute and classic measure. It indicates by how much the individual units of the surveyed population differ in terms of a given variable from its arithmetic mean. It is the arithmetic mean of the absolute values of the deviations of a variable from the arithmetic mean. Depending on the type of series in which the variable is ordered, the mean absolute deviation is calculated using the following formulae:

– for a detailed series:

$$d(x) = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n} \tag{2.70}$$

– for a point series:

$$d(x) = \frac{\sum_{i=1}^{k} |x_i - \overline{x}| n_i}{n} \tag{2.71}$$

– for a binned series:

$$d(x) = \frac{\sum_{i=1}^{k} |\dot{x}_i - \overline{x}| n_i}{n} \tag{2.72}$$

The MAD has no equivalent among positional measures, as one of the properties of this measure is that it must be calculated in reference to the arithmetic mean. It is rather uncommon measure, sometimes applied as an alternative to the standard deviation $S(x)$.

Example 2.28

Given the data in Example 2.27, determine the mean absolute deviation for the number of companies operating in 2020 in the regions of the Western European countries.

To calculate the average deviation in Group A, the arithmetic mean of the number of companies must be determined. The detailed calculation of the arithmetic mean is shown in Example 2.27. The value of the arithmetic mean of the number of companies in Group A is: ${}_A\overline{x} = 443.18$.

The variable is ordered in an interval series, so Formula (2.72) must be used to determine the mean absolute deviation. The supporting calculations are summarised in Table 2.25.

Table 2.25. Supporting calculations for Example 2.28

$\dot{x}_i$	${}_An_i$	$\left\|\dot{x}_i - {}_A\bar{x}\right\|$	$\left\|\dot{x}_i - {}_A\bar{x}\right\| \cdot {}_An_i$
125	16	318.18	5,090.88
375	15	68.18	1,022.70
625	5	181.82	909.10
875	3	431.82	1,295.46
1,125	3	681.82	2,045.46
1,375	2	931.82	1,863.64
∑	44	×	12,227.24

Source: own calculations based on Table 2.22.

The deviation of the average number of companies in Western European regions in 2020 is:

$$ {}_Ad(x) = \frac{12{,}227.24}{44} = 277.89 $$

The actual number of companies operating in the Western European regions in 2020 differed on average from the average number of companies by ±277.89 companies.

2.3.3. Quarterly variation

The quarterly deviation is a positional absolute measure. It is defined as half of the difference between the third and the first quartiles:

$$ Q = \frac{Q_{3.4} - Q_{1.4}}{2} \tag{2.73} $$

where: $Q_{3.4} - Q_{1.4}$ is an interquartile range.

This measure is based on the lower ($Q_{1.4}$) and upper ($Q_{3.4}$) quartiles. It measures the dispersion of only 50% of the observations contained between these quartiles (hence the wording in the interpretation: "in the narrowed area of variation"). The value of the quarterly deviation is not affected by observations with values smaller than the lower quartile ($x_i < Q_{1.4}$) and observations with values larger than the upper quartile ($x_i > Q_{3.4}$). Additionally, the value of the quartile deviation is not affected by atypical (outlier) observations, making this measure particularly recommended for analysing variables where such units are observed.

Example 2.29

Let us continue with Example 2.27 and determine the quarterly deviation of the number of companies operating in 2020 in the regions of Western European countries (Group A).

The calculation of the quartile deviation should start with determining the lower and upper quartiles based on the formulae: in the case of the lower quartile – (2.52) and in the case of the upper quartile – (2.58). The supporting calculations are summarised in Table 2.26.

Table 2.26. Supporting calculations for Example 2.29

	x_{0i}–x_{1i}	${}_An_i$	${}_An_{sk,i}$
${}_AQ_{1.4}$	0–250	16	**16**
	250–500	15	31
${}_AQ_{3.4}$	500–750	5	**36**
	750–1,000	3	39
	1,000–1,250	3	42
	1,250–1,500	2	44
	∑	44	×

Source: own calculations based on Table 2.22.

The quartile values for Group A are:

$$ {}_AQ_{1.4} = 0 + \frac{\frac{44}{4} - 0}{16} \cdot 250 = 171.88 $$

$$ {}_AQ_{3.4} = 500 + \frac{\frac{3 \cdot 44}{4} - 31}{5} \cdot 250 = 600.00 $$

Next, the quartile deviation is determined based on Equation (2.73):

$$ {}_AQ = \frac{600.00 - 171.88}{2} = 214.06 $$

In the narrowed area of variation, the actual number of operating companies in the regions of the Western European countries differed from the middle value (median) by an average of ±214.06 companies.

2.3.4. Areas of variation

A special case of the area of variation is the **range**, also called the **empirical area of variation**. It is an absolute and positional measure calculated as the difference between the maximum and minimum values of a variable:

$$R = x_{\max} - x_{\min} \tag{2.74}$$

It is a measure with a very limited cognitive scope, as its value depends on the values of observations that are often random and atypical for the variable. Accurate determination of this measure is possible for detailed and point series. For an interval series, however, the range is an approximation, as it is calculated as the difference between the upper endpoint of the last interval and the lower endpoint of the first interval:

$$R = x_{1k} - x_{01} \tag{2.75}$$

When the first or last interval in an interval series is not close-ended, the range cannot be determined.

On the basis of the known measures of dispersion, it is also possible to designate two **areas of variation**: classical and positional, depending on whether we examine dispersion around the arithmetic mean or the median. Based on the determined areas of variability, the typical value of the observed variable can be established. Areas of variation are calculated according to the formulae:

– for classical measures:

$$\overline{x} - S(x) < x_{typ} < \overline{x} + S(x) \tag{2.76}$$

– for positional measures:

$$M - Q < x_{typ} < M + Q \tag{2.77}$$

Example 2.30

Given the data from Example 2.6, determine the range and the classic area of variation.

To determine the range according to Formula (2.74), find the country where, in 2022, one foreign language was spoken by the smallest and the largest percentage of the population: $x_{\min} = 19.3$ and $x_{\max} = 50.8$, hence: $R = 50.8 - 19.3 = 31.5$.

The difference between the countries with the highest and the lowest population percentage speaking one foreign language was 31.5 per cent.

The establishment of the classical area of variation of speaking one foreign language should start with calculating the arithmetic mean and the standard deviation. The former measure has already been calculated in Example 2.6: $\bar{x} = 34.13$. The variance is determined by Formula (2.59):

$$S^2(x) = \frac{(29.2-34.13)^2 + (41.4-34.13)^2 + \ldots + (20.4-34.13)^2 + (50.8-34.13)^2}{8} = \\ = 126.95$$

and the standard deviation – by Formula (2.68):

$$S(x) = \sqrt{126.95} = 11.27$$

The classical area of variation – Formula (2.76) – takes the form:

$$34.13 - 11.27 < x_{typ} < 34.13 + 11.27$$

hence:

$$22.86 < x_{typ} < 45.4$$

In 2022, the percentage of people speaking a single foreign language in the Baltic Region countries that were typical in terms of the observed variable ranged from 22.86% to 45.4%.

Example 2.31

Estimate, within a narrowed area of variation, the typical age of a woman giving birth in the EU-27 in 2020, using the information summarised in Table 2.19.

In order to determine a typical level of a variable, we need to know the median and the quartile deviation. In Example 2.22, positional measures of central tendency have been determined (and discussed in detail): $Q_{1.4} = 27.57$, $M = 31.76$, $Q_{3.4} = 35.44$. Therefore, according to Formula (2.73) the quarterly deviation is:

$$Q = \frac{35.44 - 27.57}{2} = 3.94$$

The positional area of variation is established by Formula (2.77):

$$31.76 - 3.94 < x_{typ} < 31.76 + 3.94$$

having previously calculated:

$$27.82 < x_{typ} < 35.70$$

In the narrowed area of variation, the typical age of a woman giving birth in the EU-27 in 2020 ranged from 27.82 to 35.70.

2.3.5. Coefficients of variation

Coefficients of variation are relative measures of dispersion. They are the only measures of variation that are expressed in dimensionless units and, when multiplied by 100, are expressed as percentages. Because they are expressed in relative units, they can be used when comparing different populations regarding one (or many) variables and one population in terms of several variables. Coefficients of variation are the ratio of a variability measure to the average measure; depending on the measures applied (classical or positional) and the type of variation, these measures can be calculated by the formulae:

– for classical measures:

$$V_S = \frac{S(x)}{\bar{x}} \cdot 100\% \tag{2.78}$$

$$V_d = \frac{d(x)}{\bar{x}} \cdot 100\% \tag{2.79}$$

– for positional measures:

$$V_Q = \frac{Q}{M} \cdot 100\% \tag{2.80}$$

$$V_D = \frac{Q_D}{M} \cdot 100\% \tag{2.81}$$

where:

V_D – coefficient of decile variation,

Q_D – decile deviation: $Q_D = \frac{D_{9.10} - D_{1.10}}{2}$,

$D_{9.10}$ – decile 9, quantile 9/10, $D_{9.10} = C_{90.100}$,

$D_{1.10}$ – decile 1, quantile 1/10, $D_{1.10} = C_{10.100}$.

When analysing the dispersion of a variable based on coefficients of variation, the following thresholds for the measure of coefficients are adopted:

<20% – population of low diversity,
20–40% – population of moderate diversity,
40–80% – population of high diversity,
>80% – population of very high diversity.

The higher the values of the coefficient of variation, the greater the dispersion of the variable. If the coefficient of variation takes on very high values, this indicates heterogeneity in the population.

The value of the coefficient of variation can be a rationale for basing the statistical analysis on classical or positional measures. Low values of the coefficient indicate a homogeneous population with no outliers, so the use of classical measures is justified. In contrast, high values of the coefficient of variation indicate a heterogeneous population with atypical observations. With these types of observations, classical measures are not reliable; therefore, statistical analysis should rather be based on positional measures.

Example 2.32

Using the data given in Example 2.27, determine the coefficients of variation: the classical and positional coefficients of variation of the number of companies operating in the regions of Western European countries in 2020.

To determine the classical coefficient of variation (Formula (2.78)), information on the arithmetic mean and standard deviation is needed. Example 2.27 provides the detailed description of the calculation of these measures: $_A\overline{x} = 443.18$ and $_AS(x) = 355.01$. Thus, the classical coefficient of variation is:

$$_AV_S = \frac{355.01}{443.18} \cdot 100\% = 80.11\%$$

The standard deviation of companies operating in the regions of Western European countries in 2020 accounted for 80.11% of the average number of companies in Group A. In terms of the studied variable, the level of variability in Group A was very high.

To determine the positional coefficient of variation, we need to know the median and the quarterly deviation. The median is calculated based on Table 2.26 and Equation (2.44):

$$_{A}M = 250 + \frac{\frac{44}{2} - 16}{15} \cdot 250 = 350$$

The procedure for calculating the quarter deviation is shown in Example 2.29:

$$_{A}Q = 214.06$$

The positional coefficient of variation determined from Equation (2.80) is:

$$_{A}V_{Q} = \frac{214.06}{350.00} \cdot 100\% = 61.16\%$$

In the narrowed area of variation, the quarterly deviation of the number of companies operating in the regions of the Western European countries accounted for 61.16% of the median value. Also within the narrowed area of variation, the variation was high.

Positional coefficients of variation take on smaller values than the classical ones because they examine dispersion in a smaller (more narrowed) area of variation.

2.4. Measures of asymmetry

A complementary, yet equally significant, step in the statistical analysis of variables is the analysis of asymmetry. The knowledge of mean and dispersion measures is not sufficient to fully assess the structure of statistical variables – see: Example 2.33.

Example 2.33

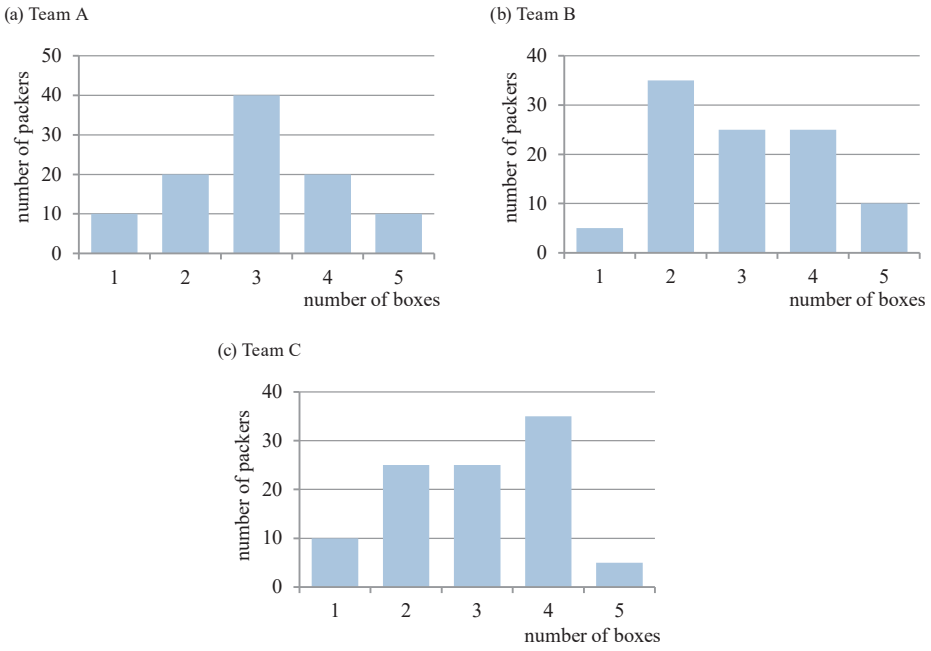
The productivity of box packers in three teams in one of Amazon's warehouses was compared. The distribution of the number of boxes packed by the three teams in 10 minutes is shown in Figure 2.17 and Table 2.27.

Table 2.27. Labour productivity distribution of the three teams of packers in the Amazon warehouse

Number of boxes	Number of packers		
x_i	${}_An_i$	${}_Bn_i$	${}_Cn_i$
1	10	5	10
2	20	35	25
3	40	25	25
4	20	25	35
5	10	10	5
∑	100	100	100

Source: exemplary data.

Figure 2.17. Distribution of labour productivity of teams A, B and C in the Amazon warehouse



Source: Table 2.27.

Each team had 100 packers. Workers packed between 1 to 5 boxes in 10 minutes. The basic measures of structure were determined for each team: arithmetic mean, median, mode, and variance. The results are summarised in Table 2.28.

Table 2.28. Supporting calculations for Example 2.33

Measures	Team		
	A	B	C
$\overline{x}$	3.0	3.0	3.0
M	3.0	3.0	3.0
D	3.0	2.0	4.0
$S^2(x)$	1.2	1.2	1.2

Source: own calculations based on Table 2.27.

The average number of boxes packed by each team was the same at 3 boxes. The median was also at the same level ($M = 3$), as well as the measure of variation – the variance ($S^2(x)$). Seemingly, it would appear that the workers in the three teams worked with the same productivity. However, an analysis of the empirical distributions (Figure 2.17) indicates that this was not the case.

In Team A, all measures of central tendency were equal to each other:

$$_A\overline{x} = {}_AM = {}_AD$$

In Team B, the productivity of most workers was below the arithmetic mean:

$$_BD < {}_B\overline{x}$$

In Team C, the productivity of most workers was above the arithmetic mean:

$$_C\overline{x} < {}_CD$$

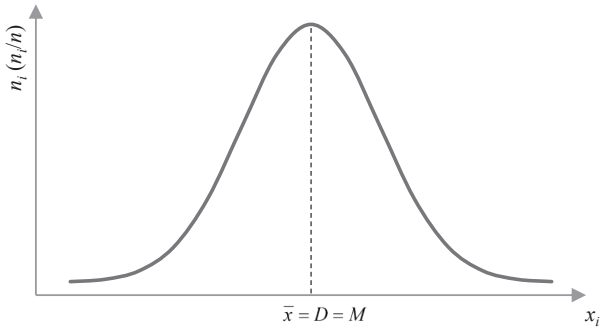
Issues related to the relative position of measures of central tendency – arithmetic mean, median, and dominant – fall within the domain of the **asymmetry (skewness)** of the **distribution**.

- The distribution of the observed variable is symmetrical when values of the variable that are equally distant from the median are accompanied by equal counts.
- The area under the count curve to the left of the median is equal to the area under the count curve to the right of the median (Figure 2.18).
- The value of the variable with the highest count falls at the midpoint of the variable.

For a symmetric distribution, the relationship holds:

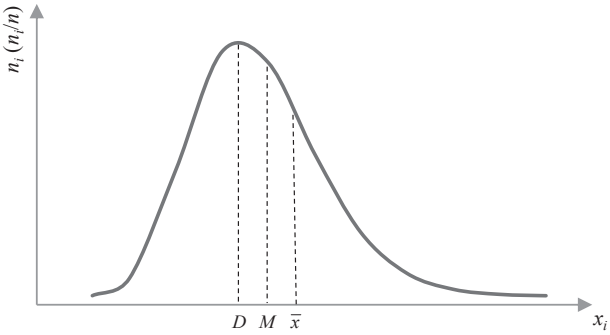
$$\overline{x} = D = M \tag{2.82}$$

Figure 2.18. Symmetrical distribution



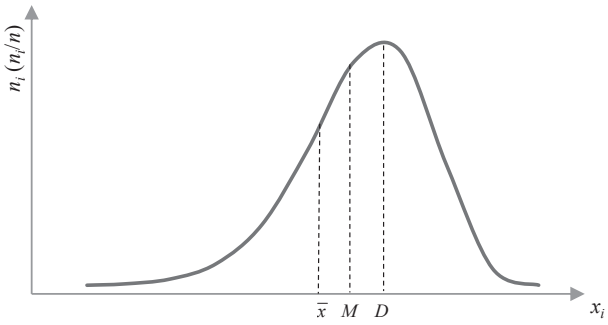
Source: own study.

Figure 2.19. Right-asymmetric distribution



Source: own study.

Figure 2.20. Left-asymmetric distribution



Source: own study.

A distribution where the average measures are not equal to each other is said to be an asymmetric distribution. There are specific relationships between the average measures when the asymmetry is:

– Right (positive):

$$D < M < \overline{x} \tag{2.83}$$

– Left (negative):

$$D > M > \overline{x} \tag{2.84}$$

For a variable with a right-asymmetric distribution, most units take values below the arithmetic mean (Figure 2.19). For a left-asymmetric distribution, most observations are above the arithmetic mean value (Figure 2.20).

The stronger the asymmetry, the greater the difference between the value of the mode and the arithmetic mean.

When examining asymmetry, it is important to determine its strength and direction. The constructed coefficients make this possible. The coefficients of skewness are dimensionless numbers. As with the previous groups of measures, a distinction is made between classical and positional coefficients – depending on the measures of mean and variation employed to calculate them.

The classical asymmetry coefficient A_1 is expressed by the formula:

$$A_1 = \frac{\mu_3}{S^3(x)} \tag{2.85}$$

where:

$S^3(x)$ – the cube of the standard deviation,

μ_3 – third-order central moment, determined according to the type of series in which the variable is ordered:

– for a detailed series:

$$\mu_3 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{n} \tag{2.86}$$

– for a point series:

$$\mu_3 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^3 n_i}{n} \tag{2.87}$$

– for a binned series:

$$\mu_3 = \frac{\sum_{i=1}^{k}(\dot{x}_i - \overline{x})^3 n_i}{n} \tag{2.88}$$

The classical coefficient of asymmetry is a fully accurate measure, but it can only be used in series where all values of the variable are known (the intervals are closed). It takes values:

- In most cases $-2 \leq A_1 \leq 2$,
- The absolute value of the coefficient $|A_1|$ indicates the strength of asymmetry:
  - $0 \leq |A_1| < 0.7$ – weak asymmetry,
  - $0.7 \leq |A_1| < 1.4$ – moderate asymmetry,
  - $1.4 \leq |A_1|$ – strong asymmetry.
- The sign of the coefficient marks the direction of the asymmetry:
  - $A_1 = 0$ indicates a symmetrical distribution,
  - $A_1 < 0$ indicates left (negative) asymmetry,
  - $A_1 > 0$ indicates right (positive) asymmetry.

The positional coefficient of asymmetry A_2 is calculated by the formula:

$$A_2 = \frac{(Q_{3.4} - M) - (M - Q_{1.4})}{(Q_{3.4} - M) + (M - Q_{1.4})} = \frac{Q_{3.4} + Q_{1.4} - 2 \cdot M}{2 \cdot Q} \tag{2.89}$$

Or, when deciles are used in the calculation:

$$A_2' = \frac{D_1 + D_9 - 2 \cdot M}{D_9 - D_1} \tag{2.90}$$

Positional measures are used to calculate A_2. Therefore, similarly to positional measures of variation, this measure also relates to a narrowed area of variation and is recommended to be used for distributions where with open-ended intervals and/or outliers (extreme values). The positional coefficient of asymmetry takes values (applies to both positional coefficients):

- In most cases $-1 \leq A_2 \leq 1$,
- The absolute value of the coefficient $|A_2|$ indicates the strength of asymmetry:
  - $0 \leq |A_2| < 0.3$ – weak asymmetry,
  - $0.3 \leq |A_2| < 0.6$ – moderate asymmetry,
  - $0.6 \leq |A_2|$ – strong asymmetry.
- The sign of the coefficient value marks the direction of asymmetry in the narrowed area of variation:
  - $A_2 = 0$ indicates a symmetrical distribution,
  - $A_2 < 0$ indicates left asymmetry,
  - $A_2 > 0$ indicates right asymmetry.

A third measure, **the coefficient of asymmetry A_3**, is applied to analytically determine the direction and strength of asymmetry. This is a classic-positional measure, called the Pearson measure. It is determined by the formula:

$$A_3 = Ws = \frac{\overline{x} - D}{S(x)} \tag{2.91}$$

or

$$A_3 = Ws = \frac{\overline{x} - D}{d(x)} \tag{2.92}$$

The coefficient A_3 does not have clearly defined bounds, although it usually takes values between – 1 and 1:

– The absolute value of the coefficient $|A_3|$ indicates the strength of asymmetry:
  - $0 \leq |A_3| < 0.3$ – weak asymmetry,
  - $0.3 \leq |A_3| < 0.6$ – moderate asymmetry,
  - $0.6 \leq |A_3|$ – strong asymmetry.
– The sign of the coefficient value marks the direction of asymmetry:
  - $A_3 = 0$ indicates a symmetrical distribution,
  - $A_3 < 0$ indicates left (negative) asymmetry,
  - $A_3 > 0$ indicates right (positive) asymmetry.

The use of the coefficient of asymmetry is limited to unimodal distributions, as only for such distributions is it possible to determine the mode. In series with open-ended intervals, this measure cannot be determined.

The coefficients A_1 and A_3 are equivalent as they test for asymmetry across the entire area of variation – they have the same logical foundation. In a statistical analysis, it is recommended to use only one of them.

For distributions with moderate asymmetry, the following relationship holds:

$$\overline{x} - D = 3(\overline{x} - M) \tag{2.93}$$

From this, the approximate value of the mode can be determined as:

$$D = \overline{x} - 3(\overline{x} - M) \tag{2.94}$$

Example 2.34

Based on classical and positional measures, determine the asymmetry of the distribution of the number of companies operating in the regions of Western European countries in 2020 (group A) using the data in Table 2.22 (Example 2.27).

First, the classical coefficient of asymmetry A_1 is determined by Formula (2.85). It will be necessary to determine the third central moment – Formula (2.88). The value of the arithmetic mean and standard deviation have been calculated in Example 2.27: ${}_A\overline{x} = 433.18$, ${}_AS(x) = 355.01$. The supporting calculations are summarised in Table 2.29.

Table 2.29. Supporting calculations for Example 2.34

$\dot{x}_i$	${}_A n_i$	$\dot{x}_i - {}_A\overline{x}$	$\left(\dot{x}_i - {}_A\overline{x}\right)^3 \cdot {}_A n_i$
125	16	–318.18	–515,393,118.01
375	15	–68.18	–4,754,033.63
625	5	181.82	30,053,493.62
875	3	431.82	241,562,499.07
1,125	3	681.82	950,890,401.97
1,375	2	931.82	1,618,177,203.25
∑	44	×	2,320,536,446.27

Source: own calculations based on Table 2.22.

The third central moment for Group A is:

$${}_A\mu_3 = \frac{2,320,536,446.27}{44} = 52,739,464.69$$

The classical coefficient of asymmetry:

$${}_AA_1 = \frac{52,739,464.69}{355.01^3} = 1.18$$

The distribution of the number of companies operating in the observed regions in countries from Group A was characterised by moderate right asymmetry.

The positional coefficient of asymmetry was determined by Formula (2.89). The determination of the necessary measures of mean and variation are discussed in Examples 2.29 and 2.32:

$${}_AQ_{1.4} = 171.88, \quad {}_AM = 350, \quad {}_AQ_{3.4} = 600, \quad {}_AQ = 214.06$$

The positional coefficient of asymmetry is:

$${}_AA_2 = \frac{600.00 + 171.88 - 2 \cdot 350.00}{2 \cdot 214.06} = 0.17$$

The distribution of the number of operating firms in the surveyed regions of the Western European countries was characterised by a varying asymmetry in the entire area of variation and in the central quadrants. In the narrowed area of variation, the number of operating firms in Group A was characterised by a weak right asymmetry.

Example 2.35

A visual assessment of the distribution of the number of foreign languages spoken in Denmark in 2022 indicated a weak left skewness (Figure 2.6 in Example 2.15). Calculate the strength of asymmetry in an analytical manner. On the basis of the coefficient of asymmetry, determine whether in 2022 most Danes spoke more or fewer foreign languages than the average.

The coefficient of asymmetry is determined by Formula (2.91). The value of the mode has already been calculated in Example 2.15: $D = 2$. The average number of foreign languages known to Danes was determined by Formula (2.21):

$$\overline{x} = \frac{176.10}{100} = 1.76$$

The variance of the number of foreign languages spoken by Danes was found by Formula (2.60):

$$S^2(x) = \frac{94.98}{100} = 0.95$$

while the standard deviation – by Formula 2.68):

$$S(x) = \sqrt{0.95} = 0.97$$

The detailed calculations necessary to determine the above measures are summarised in Table 2.30.

Table 2.30. Supporting calculations for Example 2.35

x_i	n_i	$x_i n_i$	$(x_i - \overline{x})^2 n_i$
0	11.0	0.0	34.07
1	29.3	29.3	16.92
2	32.3	64.6	1.86
3	27.4	82.2	42.13
∑	100.0	176.1	94.98

Source: own calculations based on Table 2.14.

By inserting the individual measures into the coefficient of asymmetry formula, the following is obtained:

$$A_3 = Ws = \frac{1.76 - 2}{0.97} = -0.25$$

The distribution of the number of foreign languages spoken by Danes in 2022 was characterised by weak left asymmetry. Therefore, the conclusions drawn from the observations of the distribution are correct.

The left-skewed distribution indicates that in 2022 the declared number of foreign languages spoken by the Danes was higher than the average.

2.5. Measures of kurtosis and concentration

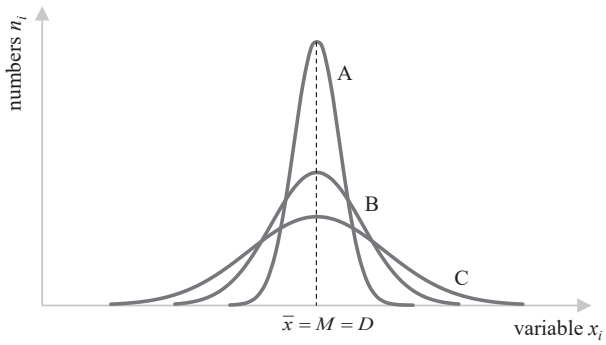
The last group of measures in a comprehensive structure analysis are measures of kurtosis and concentration. With these measures, it is possible to check whether the distribution of the variable's values among the individual units of the population is normal. Depending on the distribution of the given variable, different measures of kurtosis are used:

- when the distribution is symmetrical or weakly asymmetric, we examine the distribution of individual units around the arithmetic mean using the **measures of kurtosis**,
- when the distribution is strongly asymmetric, **measures of concentration** are applied.

2.5.1. Measures of kurtosis

Distributions for which the arithmetic mean, mode, and median are equal (such a relationship is observed for symmetric distributions) may differ in the distribution of individual units around the mean value (Figure 2.21). The measures that examine the variation of units around the mean value are called **kurtosis measures**. Measures of kurtosis are sensitive to outliers and atypical observations – the more outliers there are, the larger the measures will be (Westfall 2014).

Figure 2.21. Symmetric distributions with varying kurtosis around the arithmetic mean



A – Peaked distribution, B – Normal distribution, C – Flattened distribution.

Source: own study.

The degree of variation of a variable directly affects its kurtosis – the greater the variation, the lower the kurtosis – and this, in turn, determines the shape of the distribution. The following distinctions are made:

- Peaked (leptokurtic) distribution – shows the least variation around the average value and exhibits higher kurtosis.
- Normal (mesokurtic) distribution – serves as a model distribution for which the relationship holds:

$$\frac{\mu_4}{S^4(x)} = 3 \tag{2.95}$$

where:

μ_4 – fourth central moment,

$S^4(x)$ – standard deviation raised to the fourth power.

- Flattened (platykurtic) distribution – demonstrates the greatest variation around the average value and has the lowest kurtosis.

Kurtosis can be examined across the whole area of variation using the classical measure, or within a narrowed area of variation using the positional measure. **The classical measure of kurtosis** is expressed by the formula:

$$\alpha = \frac{\mu_4}{S^4(x)} \tag{2.96}$$

with designations as in Formula 2.95.

Depending on the series in which the variable is ordered, the fourth central moment is calculated by the formulae:

– for a detailed series:

$$\mu_4 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^4}{n} \tag{2.97}$$

– for a point series:

$$\mu_4 = \frac{\sum_{i=1}^{k}\left(x_i - \bar{x}\right)^4 n_i}{n} \tag{2.98}$$

– for a binned series:

$$\mu_4 = \frac{\sum_{i=1}^{k}\left(\dot{x}_i - \bar{x}\right)^4 n_i}{n} \tag{2.99}$$

The α coefficient is a dimensionless measure (it is not expressed in physical units of the variable), so it can be used to compare distributions of selected variables. If the value of the *α coefficient* is:

– if $\alpha < 3$, then the distribution is flattened (platykurtic),
– if $\alpha = 3$, then the distribution is normal (mesokurtic),
– if $\alpha > 3$, then the distribution is peaked (leptokurtic).

The positional measure of kurtosis is expressed using the formula:

$$K_P = \frac{Q_{3.4} - Q_{1.4}}{2\left(C_{90.100} - C_{10.100}\right)} \tag{2.100}$$

where:

$Q_{3.4}$ – third (top) quartile,
$Q_{1.4}$ – first (bottom) quartile,
$C_{90.100}$ – centile 90, quantile 90/100, $C_{90.100} = D_{9.10}$,
$C_{10.100}$ – centile 10, quantile 10/100, $C_{10.100} = D_{1.10}$.

This measure is expressed in dimensionless units. Its interpretation is as follows:

– if $K_P > 0.263$ the distribution is flattened (platykurtic),
– if $K_P = 0.263$ distribution is normal (mesokurtic),
– if $K_P < 0.263$ then the distribution is peaked (leptokurtic).

Example 2.36

An analysis of the asymmetry of the distribution of the number of foreign languages spoken by the Danish population in 2022 showed that the distribution is close to symmetrical (refer to Example 2.35). On the basis of the data summarised in Table 2.14 (see Example 2.15), determine the degree of kurtosis of the number of foreign languages spoken by the Danes.

The determination of the classical measure of kurtosis expressed using Formula (2.96), requires the knowledge of the fourth central moment. This is calculated using Formula (2.98) (as the data are presented in a point series), and the standard deviation, which requires knowing the average value of the variable. The method for determining the arithmetic mean and standard deviation has been discussed in detail in Example 2.35: $\bar{x} = 1.76$, $S(x) = 0.97$. Table 2.31 summarises the supporting calculations.

Table 2.31. Supporting calculations for Example 2.36

x_i	n_i	$x_i - \bar{x}$	$(x_i - \bar{x})^4 n_i$
0	11.0	–1.76	105.55
1	29.3	–0.76	9.78
2	32.3	0.24	0.11
3	27.4	1.24	64.78
∑	100.0	×	180.00

Source: own calculations based on Table 2.14.

The fourth central moment for the population of the surveyed Danes is:

$$\mu_4 = \frac{180.00}{100} = 1.80$$

The classical coefficient of kurtosis is equal to:

$$\alpha = \frac{1.80}{0.97^4} = 2.03$$

The distribution of the number of foreign languages that the Danes spoke in 2022 is characterised as a flattened (platykurtic) distribution.

The positional flattening coefficient for the population of the surveyed Danes was determined using Formula (2.100). The method of determining the quartiles (both lower and upper) was outlined in Example 2.21:

$$Q_{1.4} = 1,\ Q_{3.4} = 3.$$

The values of the variable in the 10th and 90th centiles were calculated based on Table 2.18 and Formula (2.45):

$$C_{10.100} \approx x_{\frac{10 \cdot 100}{100}} = x_{10} = 0\,,\ C_{90.100} \approx x_{\frac{90 \cdot 100}{100}} = x_{90} = 3$$

The positional coefficient of kurtosis is:

$${}_{A}\mathrm{K}_{P} = \frac{3-1}{2(3-0)} = 0.33$$

The analysed distribution, even within the narrowed area of variation, is also characterised as a platykurtic (flattened) distribution.

2.5.2. Measures of concentration

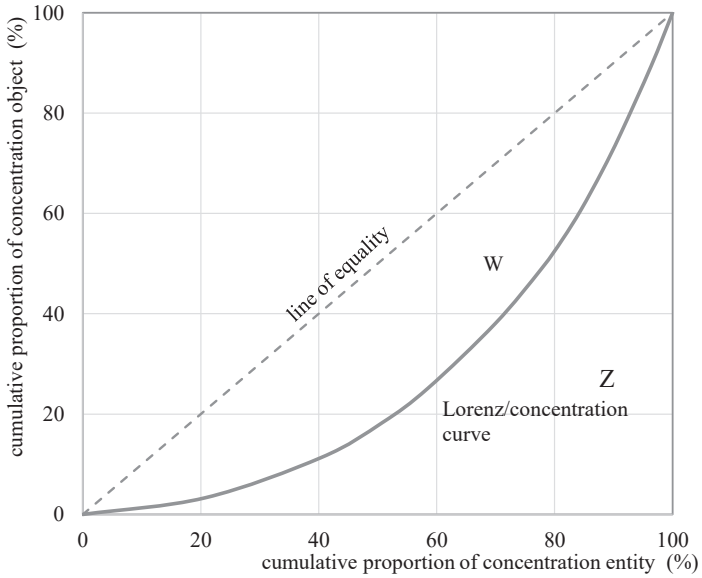
In the case of extremely asymmetric distributions, an interesting point of analysisis how unevenly the total sum of values is distributed between the individual units, or what the concentration of the variable is. Concentration (or uneven distribution) is evident in many socio-economic phenomena: the distribution of population across localities, market shares for the sale of product X, the distribution of wealth, capital, income, land, production, and many others. According to the authors of the *World Inequality Report* (2022), in 2021, the wealthiest 0.1 per cent of individuals held 3 per cent of the world's wealth, while 50 per cent of the population owned just 2 per cent. This demonstrates that the world's wealth was not evenly distributed among its population. On the contrary, a large proportion of wealth was concentrated within a small percentage of people. Such a phenomenon is said to be characterised by high concentration.

Concentration can be examined graphically and analytically. The **graphical method** involves plotting and estimating the Lorenz concentration curve (Figure 2.22).

In the graph, the cumulative share of the population (cumulative relative frequencies in percent) of the variable is plotted on the *x*-axis, and the cumulative share

of the variable total value is plotted on the *y*-axis. Connecting the points forms a polygon (if the points are connected by segments) or the Lorentz curve. The coordinate system representing the cumulative percentages on both axes, is a quadratic system with a side equal to 100%. The main diagonal of this square is the **line of equality**. The area between the line of equality and the concentration curve is called the concentration field (*W-field*). The larger the *W-field*, the higher the concentration.

Figure 2.22. Graphical representation of the Lorentz curve



Source: own study.

In extreme cases, no concentration is observed (the concentration curve coincides with the line of uniform distribution – Figure 2.23a) or complete (total) concentration is observed, when the Lorentz curve is maximally distant from the line of equality (Figure 2.23d). However, these cases are very rare, with weak or moderate concentration being observed for most variables (Figure 2.23b–c).

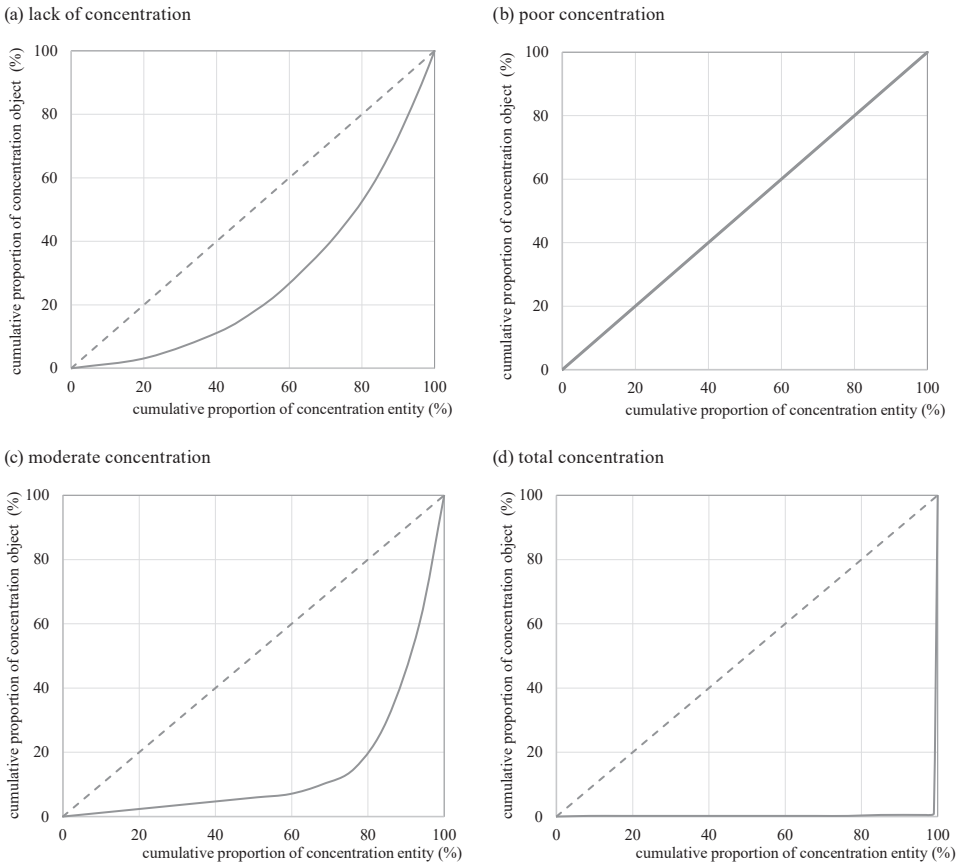
The **analytical method** of determining the level of concentration of a variable involves calculating the ratio of the *W-field* to the area of the triangle delimited by the line of equality, i.e., the *W* + *Z field*. The formula was proposed by Gini (1912) hence it is called the **Gini coefficient** (although it is sometimes referred to as the **Lorentz coefficient)**. The formula is written as:

$$K_L = \frac{W}{W+Z} = \frac{5{,}000 - Z}{5{,}000} \tag{2.101}$$

The figure $W + Z$ is a right-angled triangle defined by the line of equality and two perpendiculars sides, both equal to 100. Based on the formula for the area of a triangle, the area of the figure $W + Z$ is calculated as:

$$W + Z = \frac{1}{2} \cdot a \cdot h = \frac{1}{2} \cdot 100 \cdot 100 = 5{,}000$$

Figure 2.23. Exemplary positions of Lorentz concentration curve



Source: own study.

To calculate the *Z-field*, it is necessary to determine the analytical form of the concentration curve function and then calculate the area under the curve. In practice, the *Z-field* is determined approximately (Figure 2.24 provides a graphical representation). Each point forming the Lorentz curve is projected onto the x-axis. This projection divides the *Z-field* into a series of fields P_i:

$$Z = \sum_{i=1}^{k} P_i \tag{2.102}$$

here, P_1 is a triangle, while the remaining *Ps* are trapezoids. For example, the trapezoid P_2, has two bases: the upper one is U_{sk1} and the lower one is U_{sk2}, with the height of this trapezoid given as Y_2. Using the formula for the area of a trapezoid, the area of the figure P_i is calculated as:

$$P_i = \frac{U_{sk,i} + U_{sk,i-1}}{2} \cdot Y_i \tag{2.103}$$

where:

Y_i – percentage of count for the *i*-th value of the variable:

$$Y_i = \frac{n_i}{\sum_{i=1}^{k} n_i} \cdot 100\% \tag{2.104}$$

$U_{sk,i}$ – cumulative percentage of the sum of values for the *i*-th value of the variable,

$U_{sk,i-1}$ – cumulative percentage of the sum of values for the previous variable value,

U_i – percentage of the sum of values for the *i*-th value of the variable. If this value cannot be established from the individual observations, it can be calculated depending on the series in which the data are presented, using the following formulas:

– for a point series:

$$U_i = \frac{x_i n_i}{\sum_{i=1}^{k} x_i n_i} \cdot 100\% \tag{2.105}$$

– for a binned series:

$$U_i = \frac{\dot{x}_i n_i}{\sum_{i=1}^{k} \dot{x}_i n_i} \cdot 100\% \tag{2.106}$$

If there are open-ended intervals in the interval series, the use of Formula (2.104) may lead to errors. In such a case, the total value of the variable must be determined from the individual observations.

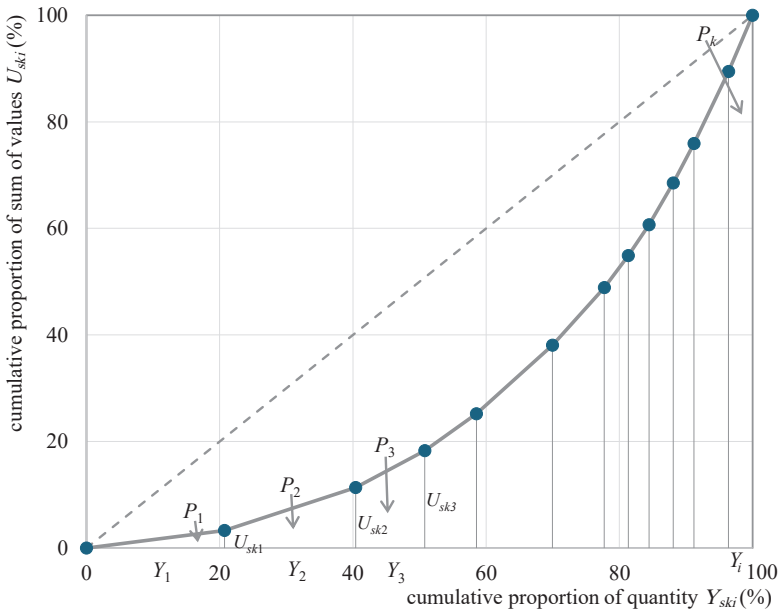
The Gini concentration coefficient is a dimensionless measure and takes values:

$$0 \leq K_L \leq 1$$

The level of measure K_L indicates:

$K_L = 0$ – no concentration
$0 < K_L < 0.3$ – poor concentration,
$0.3 < K_L < 0.6$ – moderate concentration,
$0.6 < K_L < 1$ – strong concentration,
$K_L = 1$ – total concentration.

Figure 2.24. Graphical explanation of the determination of the *Z-field*



Source: own study.

Example 2.37

Information on the structure in the regions (NUT2) in selected European countries, based on the area of land allocated to vine-growing holdings in 2020, is presented in Table 2.32. Assess graphically and analytically whether there was a concentration of land allocated for this purpose in these regions.

Table 2.32. Structure of vineyard areas in regions of selected European countries in 2020

Area under vines (thousand ha)	Number of regions	Total area under vines (thousand ha)
0–2	51	17.7
2–5	18	58.8
5–10	24	166.1
10–20	13	186.6
20–50	28	850.6
50–100	14	1,005.9
100–200	2	257.4
Over 200	2	641.8
∑	152	3,184.9

Source: Eurostat (vit_t1), accessed 16/02/2024.

The total area under vines (third column in Table 2.32) was calculated based on individual data obtained from the regions by summing the land area in each class.

To determine the intensity of concentration, it is necessary to calculate the percentage $\frac{n_i}{n} \cdot 100\%$ (or relative frequencies $\frac{n_i}{n}$) of the number of regions Y_i (formula 2.104) and the total land area U_i. Subsequently, the cumulative values of both measures are calculated. The supporting calculations are summarised in Table 2.33.

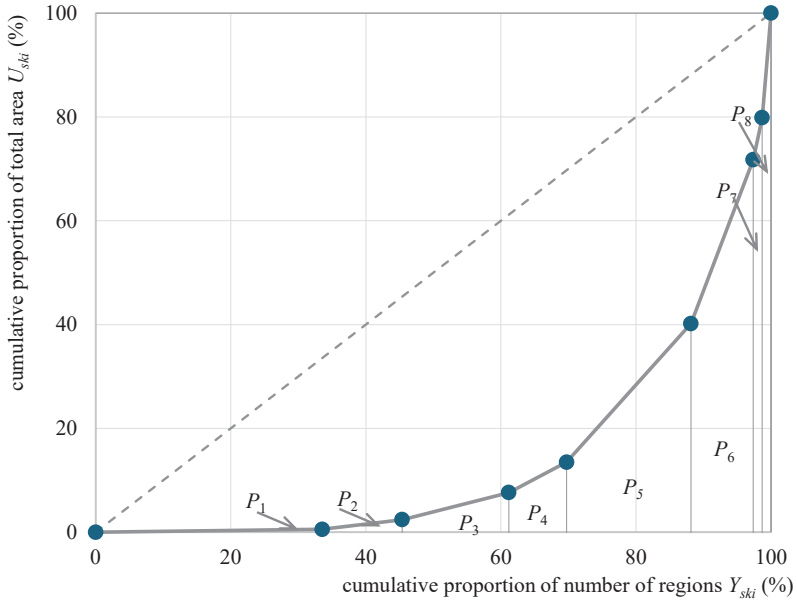
Table 2.33. Supporting calculations for Example 2.37 – calculation of Lorentz curve

Land area (thousand ha)	Number of regions	Total land area (thousand ha)	Percentage		Cumulative percentage	
			of number of regions	of total land area	of number of regions	of total land area
			Y_i	U_i	$Y_{sk,i}$	$U_{sk,i}$
0–2	51	17.7	33.55	0.56	33.55	0.56
2–5	18	58.8	11.84	1.85	45.39	2.41
5–10	24	166.1	15.79	5.21	61.18	7.62
10–20	13	186.6	8.55	5.86	69.73	13.48
20–50	28	850.6	18.42	26.71	88.15	40.19
50–100	14	1,005.9	9.21	31.58	97.36	71.77
100–200	2	257.4	1.32	8.08	98.68	79.85
Over 200	2	641.8	1.32	20.15	100.00	100.00
∑	152	3,184.9	100.00	100.00	×	×

Source: own calculations based on Table 2.32.

The last two columns in Table 2.33 represent the coordinates of the points which, when plotted on the coordinate system and connected, form the Lorenz concentration polygon (curve). These data should be supplemented with a starting point (0, 0) – Figure 2.25.

Figure 2.25. Lorentz curve of land under vine in regions of selected European countries in 2020



Source: Table 2.33.

The position of the Lorentz curve with respect to the line of uniform distribution indicates a strong concentration of land under vine in European regions in 2020.

The graphical method of determining the concentration will be supplemented by an analytical method. To determine the Ginni coefficient (Formula 2.101), the points from which the Lorentz polygon has been determined should be projected onto the *x*-axis. The resulting geometrical figures include triangles (the first figure) and trapezoids (the others). The number of geometrical figures determined corresponds to the number of classes in the interval series (Figure 2.25).

The area of the figures P_i is calculated by Formula (2.103). For example:

$$P_1 = \frac{0.56}{2} \cdot 33.55 = 9.39\,, \quad P_2 = \frac{0.56 + 2.41}{2} \cdot 11.84 = 17.58\,, \quad \text{etc.}$$

Detailed calculations of the areas of the remaining figures are shown in Table 2.34.

Table 2.34. Supporting calculations for Example 2.37 – Calculation of areas of figures

Land area (thousand ha)	Y_i	$U_{sk,i}$	$\frac{U_{sk,i} + U_{sk,i-1}}{2}$	$\frac{U_{sk,i} + U_{sk,i-1}}{2} \cdot Y_i$
0–2	33.55	0.56	0.28	9.39
2–5	11.84	2.41	1.49	17.58
5–10	15.79	7.62	5.02	79.19
10–20	8.55	13.48	10.55	90.2
20–50	18.42	40.19	26.84	494.30
50–100	9.21	71.77	55.98	515.58
100–200	1.32	79.85	75.81	100.07
Over 200	1.32	100.00	89.93	118.70
∑	100.00	×	×	1,425.01

Source: own calculations based on Table 2.32.

The sum of the areas of the designated figures is 1,425.01, i.e.

$$Z = 1{,}425.01$$

The concentration factor is:

$$K_L = \frac{5{,}000 - 1{,}425.01}{5{,}000} = 0.72$$

The analytical method of determining concentration confirmed the conclusions drawn from the graphical method – in 2020, there was a strong concentration of land under vine across regions of selected European countries. In many regions (more than 33%), there were vineyards, but their share of their total area was small (0.56%). However, there were also regions, constituting a marginal percentage of 1.32%, where more than 20% of all land was under vine.

2.6. Comparison and comprehensive analysis of structures

When comparing two or more structures of a variable in different populations, it can be interesting to see how similar (or different) the structures under consideration are. There is a number of measures designed to test the similarity of structures (Ostasiewicz 2011), one of which is the structure similarity coefficient written:

$$w_p = \sum_{i=1}^{k} \min\left(w_{1i}, w_{2i}\right) \tag{2.107}$$

where:

w_p – the structure similarity coefficient,

w_i – the population structure index (relative frequency): $w_i = \frac{n_i}{n}$, $i = 1, 2, ..., k$.

A prerequisite for applying this particular measure is that the structures under study are grouped in the same way. The structure similarity coefficient takes values from 0 to 1 ($0 \leq w_p \leq 1$); the closer the value is to zero, the less similar the structures are. In contrast, the higher the value of the coefficient and the closer it is to unity, the more similar the structures are to each other.

Example 2.38

Table 2.35 shows the floor area structure of flats listed and sold in 2018 in Szczecin, Poland.

Table 2.35. Structure of the floor area of flats listed and sold in Szczecin in 2018 with supporting calculations

Floor area (m^2) x_{0i}–x_{1i}	Flats listed n_{1i}	Flats sold n_{2i}	w_{1i}	w_{2i}	$\min(w_{1i}, w_{2i})$
1	2	3	4	5	6
Less than 30	21	213	0.032	0.083	0.032
30–40	50	465	0.077	0.181	0.077
40–50	81	691	0.125	0.269	0.125
50–60	101	448	0.156	0.174	0.156
60–70	100	350	0.154	0.136	0.136
70–80	97	180	0.150	0.070	0.070
80–90	66	96	0.102	0.037	0.037

1	2	3	4	5	6
90–100	39	42	0.060	0.016	0.016
100–110	35	42	0.054	0.016	0.016
110–120	20	13	0.031	0.005	0.005
120–130	11	7	0.017	0.003	0.003
130–140	7	12	0.011	0.005	0.005
140–150	9	5	0.014	0.002	0.002
150–160	2	1	0.003	0.000	0.000
160–170	3	2	0.005	0.001	0.001
170–180	2	2	0.003	0.001	0.001
Above 180	4	1	0.006	0.000	0.000
∑	648	2,570	1.000	1.000	0.682

Source: Gdakowicz, Putek-Szeląg 2020.

The structure similarity coefficient (Formula (2.107)) will be used to compare the similarity of the structures. For this purpose, the relative frequencies of flats listed and sold must be determined, and then the smaller value of the relative frequency from each class must be determined.

The structure similarity coefficient, $w_p = 0.682$, indicates a significant similarity in the area of flats listed and sold in Szczecin in 2018.

Comprehensive structure analysis not only aims, as the name implies, to analyse, but should also allow comparison – whether of a single population across different variables or of different populations for a selected variable. In both cases, comparability of results can be challenging. Therefore, it is important that the measures chosen for comparison meet the following criteria:

– when analysing a single population for different variables, often expressed in different units, **relative measures** should be used,
– when comparing the value of a selected variable across different populations, and the variables are presented in the same units, **all measures** of structure analysis can be used, as will be interpretable.

Table 2.36 summarises the most commonly used measures in structure analysis. The choice of measures depends on the scope of the study and the variables under analysis.

Table 2.36. Summary of structure analysis measures

Measures	Classic	Positional
I. Measures of central tendency (average)		
absolute	arithmetic mean $\overline{x}$	Mode D quantiles: median M lower quartile $Q_{1.4}$ upper quartile $Q_{3.4}$
II. Measures of diversity (variability)		
absolute	(variance $S^2(x)$) standard deviation $S(x)$ typical area of variation x_{TYP}	*R-distribution* quarterly deviation Q typical area of variation x_{TYP}
relative	coefficient of variation V_S	coefficient of variation V_Q
III. Measures of asymmetry		
relative	asymmetry factor A_1	asymmetry factor A_2
	asymmetry factor A_3	
IV. Measures of kurtosis*		
relative	classical measure α	positional measure K_P
V. Measures of concentration*		
relative	the Lorentz concentration curve and the Gini coefficient K_L	

* The use of Group IV and V measures depends on the strength of asymmetry.

Source: own study.

Example 2.39

The monthly wages of workers in two Italian factories, A and B, were analysed. What measures can be used to compare wages in the two factories?

Both A and B are Italian factories, so workers' wages are paid in euros (the same unit). For a comprehensive comparison of the workers' monthly wages in the two factories, all the measures of structure analysis presented in Table 2.35 can be applied.

Example 2.40

In 2022, more than 30,000 residential properties were sold in Estonia. Real estate agents are interested in information on the distribution of sold properties by floor area (m^2) and price per m^2 (EUR). What measures should they use for comparative analysis?

This analysis involves comparing one set (residential properties sold in Estonia in 2020) based on two variables: floor area and price of m^2. The variables are expressed in different units: m^2 and EUR. The measures that will be applicable (and interpretable) in this case are all dimensionless (ratio) measures. Out of the group of variation measures, V_S and V_Q (Table 2.36), as well as measures of asymmetry, kurtosis, and concentration can be used.

Chapter 3

Interdependence analysis

This chapter deals with another type of statistical regularity – regularities in the interdependence of phenomena (the previous one dealt with regularities in the structure of phenomena). **The interdependence analysis** is the study of relationships between statistical variables. It consists of two parts – correlation analysis and regression analysis (Weinbach, Grinnell 2007; Hozer 1998). Correlation analysis involves the determination of different types of correlation coefficients, while in regression analysis a mathematical function is constructed to describe the interdependence between the studied variables. Any relationship between statistical variables can be assessed in terms of its strength and direction. In terms of strength, relationships can be weak, moderate or strong. When assessing the direction, it is indicated whether the relationship is positive, i.e. whether the values of the variables change in the same direction, or negative – changes occur in opposite directions. An example of a positive interdependence is the relationship between the number of paid working hours and the monthly salary of an employee. The more hours worked, the higher the salary. An example of a negative relationship is the one between paid working hours and hours spent on housework. The longer the paid working hours, the shorter the housework time.

3.1. Correlation series and table

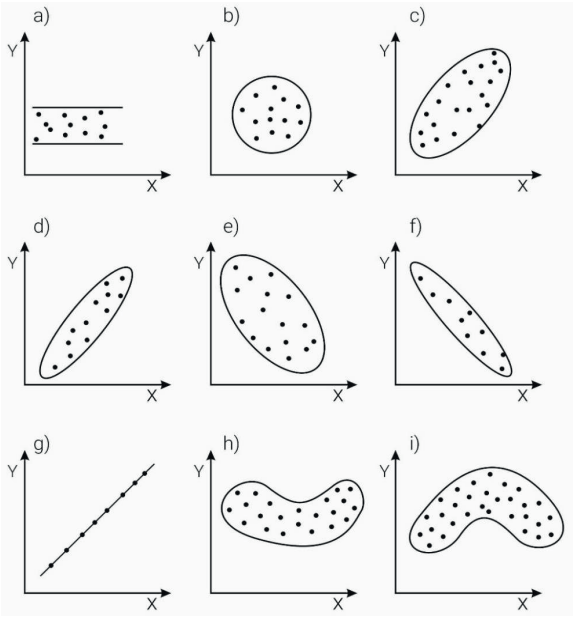
3.1.1. Correlation series

Data in correlation analysis can be presented as a correlation series or a correlation table. The **correlation series** consists of two columns of values of the studied characteristics for each observed unit. One value relates to the variable Y (dependent), the

other to the variable X (independent). When we are dealing with regularity regarding the interdependence in space, we denote the empirical values by: (x_i, y_i). When we are investigating regularity in interdependence over time the values are denoted: (x_t, y_t). Examples of the correlation series can be found in Tables 1.13–1.14.

The graphical representation of the correlation series is called **a correlation diagram** or empirical scatterplot. Pairs of points are marked in the coordinate system (x_i, y_i), which represent each unit of the population. The correlation diagram is the first step in analysing the interdependence of statistical characteristics. Figure 3.1 shows the basic research situations that are observed when the points are plotted on the diagram.

Figure 3.1. Types of dependencies between the two variables *Y* and *X*



Source: own study.

The diagrams a and b indicate the lack of correlation. This happens when the points (x_i, y_i) can be delimited by parallel lines or a circle. The dependency is linear when the points can be bounded by an ellipse (c, d, e, f, g), while a non-linear dependency occurs when the points are not aligned in a straight line. The dependency is positive when the characteristic values evolve in the same direction (c, d, g), and it is

negative when the values change in opposite directions (e, f). An example of a functional dependency is Figure g where the points are plotted exactly on a straight line.

Example 3.1

The statistics presented in Table 3.1 refer to working time in hours per week (Y) and GDP per capita in PPS (X) in a selection of 16 European countries in 2022. A correlation diagram describing the dependency between the studied variables is presented below.

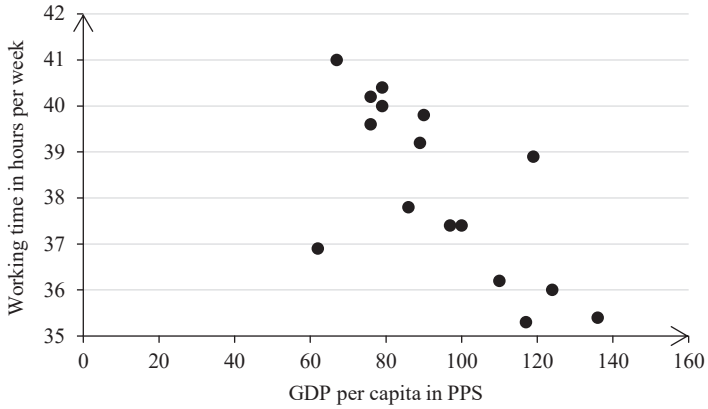
Table 3.1. Weekly working time (in hours) and GDP per capita (Purchasing Power Standards, PPS) in selected European countries in 2022

No.	Country	Working time (hours per week) y_i	GDP per capita (PPS) x_i
1	Belgium	36.9	62
2	Czechia	39.8	90
3	Denmark	35.4	136
4	Germany	35.3	117
5	Greece	41.0	67
6	Spain	37.8	86
7	France	37.4	100
8	Italy	37.4	97
9	Lithuania	39.2	89
10	Hungary	39.6	76
11	Austria	36.0	124
12	Poland	40.4	79
13	Portugal	40.0	79
14	Romania	40.2	76
15	Finland	36.2	110
16	Sweden	38.9	119

Source: own compilation based on Eurostat data (lfsa_ewhun2 and tec00114), accessed 20/03/2024.

A correlation diagram describing the dependency between the studied variables is presented in Figure 3.2. The dependent variable Y shows paid working hours, while the independent variable X is GDP per capita. The dependency between the variables is negative, which means that the higher the GDP per capita, the shorter the weekly working time. In other words, people work shorter hours in the wealthier European countries.

Figure 3.2. Correlation diagram of weekly working time in hours and GDP per capita in PPS in Europe in 2022



Source: own elaboration based on Table 3.1.

3.1.2. Correlation table

The second form of presenting statistical data is a **correlation table**. It consists of rows showing the values or variants of one characteristic, columns showing the values or variants of the other characteristic, and the counts of the units in the centre of the table. The counts n_{ij} represent the number of units having the i-th variant/value of characteristic X and the j-th variant/value of characteristic Y. A diagram of a correlation table is shown in Table 3.2.

Table 3.2. Diagram of correlation table

Characteristic X ($X = x_i$)	Characteristic Y ($Y = y_i$)					
	y_1	y_2	y_3	...	y_l	$n_{i.}$
x_1	n_{11}	n_{12}	n_{13}	...	n_{1l}	$n_{1.}$
x_2	n_{21}	n_{22}	n_{23}	...	n_{2l}	$n_{2.}$
x_3	n_{31}	n_{32}	n_{33}	...	n_{3l}	$n_{3.}$
...	...	...	...	...	...	...
x_k	n_{k1}	n_{k2}	n_{k3}	...	n_{kl}	$n_{k.}$
$n_{.j}$	$n_{.1}$	$n_{.2}$	$n_{.3}$	...	$n_{.l}$	n

Source: own study.

The following designations have been adopted:

x_i – values/variants of characteristic X, i – row number ($i = 1, 2, ..., k$),
y_j – values/variants of characteristic Y, j – column number ($j = 1, 2, ..., l$),
k – number of rows,
l – number of columns,
$n_{i.}$ – count of characteristic values $X = x_i$,
$n_{.j}$ – count of characteristic values $Y = y_j$,
n_{ij} – number of units with characteristic value $X = x_i$ and characteristic value $Y = y_j$,
n – size of the total population.

If the counts n_{ij} line up along the main diagonal – the interdependency of the variables is positive. If the counts line up along the other diagonal – the dependency is negative.

When the data are presented in the form of a correlation table, the study of interdependence is also conducted by analysing **marginal distributions** and **conditional distributions.** The marginal distributions represent the structure of the population by the characteristic X independently of the value of characteristic Y and by the characteristic Y independently of the value of characteristic X. Each correlation table therefore contains two marginal distributions. A graphical presentation of these distributions is shown in Table 3.3. They contain individual values of the characteristics X and Y and their counts. The marginal distributions are used to analyse the structure of the statistical variables.

Table 3.3. Diagram of marginal distributions of characteristics X and Y

Empirical values of characteristic $X = x_i$	Row count $n_{i.}$
x_1	$n_{1.}$
x_2	$n_{2.}$
x_3	$n_{3.}$
...	...
x_k	$n_{k.}$
$\sum$	n

Empirical values of characteristic $Y = y_j$	Column count $n_{.j}$
y_1	$n_{.1}$
y_2	$n_{.2}$
y_3	$n_{.3}$
...	...
y_l	$n_{.l}$
$\sum$	n

Source: own study.

Conditional distributions characterise the interdependency of the two examined characteristics. The number of conditional distributions is ($k + l$), which is the

sum of the rows and columns of the correlation table. Table 3.4 shows examples of the distribution of characteristic Y conditional on characteristic X taking one of its k values and the distribution of characteristic X conditional on characteristic Y taking one of its l values.

Table 3.4. The distribution of characteristic Y when characteristic X takes value x_2 ($X = x_2$) and the distribution of characteristic X when characteristic Y takes value y_3 ($Y = y_3$)

Characteristic Y	Number of units when $X = x_2$
y_1	n_{21}
y_2	n_{22}
y_3	n_{23}
...	...
y_l	n_{2l}
$\sum$	$n_{2.}$

Characteristic X	Number of units when $Y = y_3$
x_1	n_{13}
x_2	n_{23}
x_3	n_{33}
...	...
x_k	n_{k3}
$\sum$	$n_{.3}$

Source: own study.

The analysis of the interdependency of two characteristics without determining specific correlation measures can be conducted based on empirical regression lines (the diagram for a correlation table). They are constructed by assigning the conditional means of one characteristic to the values of the other characteristic. This implies that the conditional mean of characteristic Y is a function of the successive values of characteristic X:

$$\bar{y}_i = f(x_i) \tag{3.1}$$

Similarly, the conditional mean of characteristic X is a function of characteristic Y:

$$\bar{x}_j = f(y_j) \tag{3.2}$$

The graphical presentation of empirical regression functions includes plotting two lines on the coordinate system. The first line is an empirical regression of characteristic Y against characteristic X, showing the course of the conditional means of characteristic Y when characteristic X changes. This means that a curve with coordinates is drawn on the graph: $(x_1, \bar{y}_1)$, $(x_2, \bar{y}_2)$, …, $(x_k, \bar{y}_k)$. The second line is the empirical regression of characteristic X against characteristic Y, describing the course of average changes of characteristic X when the values of characteristic Y change. Points are therefore plotted on the graph: $(y_1, \bar{x}_1)$, $(y_2, \bar{x}_2)$, …, $(y_l, \bar{x}_l)$.

By analysing the position of the empirical regression lines, it is possible to determine whether the correlation exists, how strong it is, and what direction of the dependency it takes. The smaller the angle of the intersection, the stronger the relationship. If the lines intersect at a single point and the angle of intersection approximates 90 degrees, there is no correlation between the characteristics.

Example 3.2

A correlation table was constructed to assess the interdependency of the population of cities (Y) and their area (X). The study covered 58 cities in Poland in 2022. Based on the marginal distributions and empirical regression lines, the structure of the statistical characteristics and their relationship were examined. A graph of the empirical regression lines was also drawn. The data are summarised in Table 3.5.

Table 3.5. Urban population in thousands relative to urban area in km^2 in Poland in 2022

Urban area in km^2 x_i	Urban population in thousands y_j				
	50–100	100–200	200–500	500–1000	$n_{i.}$
25.0–49.9	10	1			11
50.0–99.9	10	13			23
100.0–299.9	2	10	7	3	22
300.0–499.9			1	1	2
$n_{.j}$	22	24	8	4	58

Source: own research based on data from Rocznik Statystyczny Rzeczypospolitej Polskiej 2023 (Statistical Yearbook of the Republic of Poland 2023).

The first characteristic X (urban area) is a continuous quantitative characteristic. The Y (urban population) is a quantitative discrete characteristic. The values of both characteristics are expressed in intervals. The marginal distributions can be read from the main correlation table, but are recorded separately in Table 3.6 for ease of reference.

Table 3.6. Marginal distributions of population and area of Polish cities in 2022

Urban area $X = x_i$	Row count $n_{i.}$
25.0–49.9	11
50.0–99.9	23
100.0–299.9	22
300.0–499.9	2
∑	58

Urban population $Y = y_j$	Column count $n._j$
50–100	22
100–200	24
200–500	8
500–1000	4
∑	58

Source: own elaboration based on Table 3.5.

The sum of the counts of the two marginal distributions is the same at 58.

Empirical regression lines are constructed from the formulas for conditional means for the characteristics expressed as intervals (Table 3.7):

$$\overline{y}_i = \frac{\sum_{j=1}^{r} \dot{y}_j n_{ij}}{n_i.} \tag{3.3}$$

$$\overline{x}_j = \frac{\sum_{i=1}^{k} \dot{x}_i n_{ij}}{n._j} \tag{3.4}$$

The interval midpoints are used in the calculations. The mean value of the area X in the first interval (37.5 km^2) is calculated by determining the midpoint of the interval from the formula: $\frac{25.0 + 50.0}{2} = 37.5$.

The variable X (urban area) is a continuous variable and the values of the interval endpoints do not coincide with the initial values of the subsequent intervals. Taking this fact into account, the midpoint of the first interval is determined using the arithmetic mean of the initial values of the first and second intervals.

The midpoints of the intervals for the second variable Y (urban population) are also calculated. As the endpoint of an interval coincides with the beginning of the next interval, the interval means are determined as the arithmetic mean of the beginning and the end of the respective interval. The first interval midpoint for urban population (in thousands) is calculated as follows: $\frac{50 + 100}{2} = 75$.

All conditional means are then calculated. In order to calculate the average population in cities of area between 25.0 and 49.9 km^2, the conditional distribution of the characteristic Y is taken when the characteristic X takes values between 25.0 and 49.9.The mean $\overline{y}_1$ sought amounts to: $\frac{75 \cdot 10 + 150 \cdot 1 + 350 \cdot 0 + 750 \cdot 0}{11} = 81.8$.

This means that cities with a relatively small area between 25.0 and 49.9 km^2 have a mean population of 81,800.

To calculate the first conditional mean $\overline{x}_1$, i.e. the mean urban area (X) with the population between 50,000 and 100,000 (the interval midpoint at 75), the conditional distribution of the characteristic X is taken when Y takes values in this interval. The first conditional mean $\overline{x}_1$ is 69.3 km^2. It is calculated from the formula: $\frac{37.5 \cdot 10 + 75 \cdot 10 + 200 \cdot 2 + 400 \cdot 0}{22} = 69.3$. This result informs us that the cities with the smallest population, i.e. between 50 and 100 thousand, have a mean area of 69.3 km^2.

Table 3.7. Empirical regression lines of urban population against urban area in km^2 (I) and urban area against population (in thousands), (II)

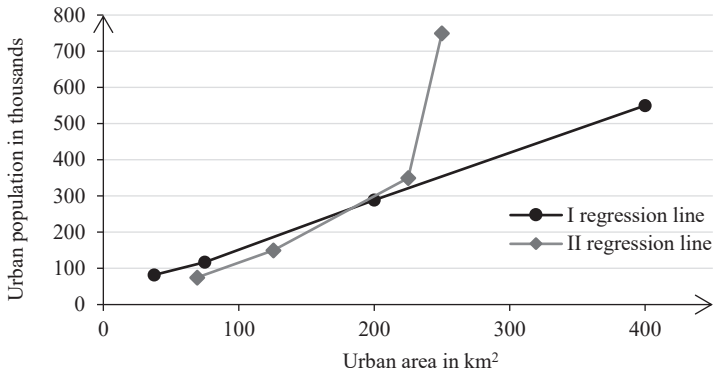
$\dot{x}_i$	$\overline{y}_i$
37.5	81.8
75.0	117.4
200.0	288.6
400.0	550.0

$\dot{y}_j$	$\overline{x}_j$
75.0	69.3
150.0	125.5
350.0	225.0
750.0	250.0

Source: own study.

A graphical presentation of the empirical regression lines of urban population against urban area and urban area against population is shown in Figure 3.3.

Figure. 3.3. Empirical regression lines of urban population against urban area (I), and urban area against population, (II)



Source: own elaboration based on Table 3.7.

The interdependency of the studied variables is positive because the regression curves are rising. This means that as the area of cities X increases, their population Y increases. This is a dependency of medium strength because the angle of intersection of the lines is relatively narrow. When the empirical regression lines intersect at an angle close to a right angle (90 degrees), then the interdependency is weak.

3.2. Measures of the strength of statistical interdependence

The relationship between two characteristics can be measured using a variety of coefficients. The choice of the appropriate one depends on the scale on which the observed characteristics are measured and whether the data are presented as a series or a correlation table (Weinbach, Grinnell 2007; Hozer 1998; Gdakowicz, Putek-Szeląg 2020). Most measures in interdependence analysis measure the strength and direction of a correlation. However, there are some measures that assess the strength only, e.g., the Tschuprow coefficient. The most used measures of statistical interdependence are:

- Tschuprow's coefficient (T_{yx}),
- Spearman's rank correlation coefficient (R_{yx}),
- Pearson's correlation coefficient (r_{yx}),
- correlation ratios (e_{yx} and e_{xy}).

Tschuprow's coefficient is used to produce a correlation table when one or both characteristics are qualitative, or presented on a nominal or ordinal scale. Spearman's rank correlation coefficient is used for quantitative, or qualitative but orderable characteristics (the scale is at least ordinal). Pearson's correlation coefficient and correlation ratios are applied for interval and ratio scales. In such a case, the characteristics are quantitative (as regards correlation ratios, at least the dependent characteristic is quantitative). At the end of Section 3.2, a tabular summary of all described measures of statistical interdependence is provided.

3.2.1. Tschuprow's coefficient

Tschuprow's coefficient T_{xy} is based on χ^2 (chi-square test statistic). It is designed to quantify the strength of the relationship of qualitative characteristics. However, the coefficient does not measure its direction as it cannot take negative values. It is defined by the formula:

$$T_{xy} = T_{yx} = \sqrt{\frac{\chi^2}{n\sqrt{(k-1)\cdot(r-1)}}} \tag{3.5}$$

where:

χ^2 – chi-square test statistic,
n – number of observations,
k – number of the row in the contingency table,
r – number of the column in the contingency table.

The chi-square test statistic is defined by the formula:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{r}\frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} = \sum_{i=1}^{k}\sum_{j=1}^{r}\frac{n_{ij}^2}{\hat{n}_{ij}} - n \tag{3.6}$$

$$\hat{n}_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n} \tag{3.7}$$

where:

n_{ij} – partial counts in the contingency table,
$\hat{n}_{ij}$ – theoretical counts calculated under the assumption that both characteristics are independent,
$n_{i\cdot}$ – marginal counts due to the first characteristic,
$n_{\cdot j}$ – marginal counts of the second characteristic,
n – number of units,
$i = 1, 2, ..., k$ – row number in the contingency table,
$j = 1, 2, ..., r$ – column number in the contingency table.

Each of the coefficients has specific properties. Tschuprow's T takes values in the interval <0, 1>. It is symmetric, i.e., $T_{xy} = T_{yx}$. It does not indicate the direction of the dependency, as it cannot take negative values. It is primarily applied for qualitative characteristics (nominal and ordinal scale), but can be used in the analysis of the interdependence of quantitative characteristics (interval and ratio scales).

Example 3.3

Examining whether there is a relationship between gender and place of residence in the population aged 15+ in Poland in 2021. In other words, checking whether the population structure by gender is related to the place of residence. The information collected was used to build a correlation table (Table 3.8).

Table 3.8. Gender and place of residence of the Polish population (in thousands), in 2021

Gender	Place of residence		
	urban	rural	$n_{i\cdot}$
Female	10,387.5	6,370.1	16,757.6
Male	9,079.9	6,260.8	15,340.7
$n_{\cdot j}$	19,467.4	12,630.9	32,098.3

Source: own research based on data from Rocznik Statystyczny Rzeczypospolitej Polskiej 2023 (Statistical Yearbook of the Republic of Poland 2023).

The relationship between the two qualitative variables can be assessed with Tschuprow's T. The calculation is shown in Table 3.9, where the first column contains counts n_{ij} taken from Table 3.8, while the data in subsequent columns are obtained with Equation (3.6) and (3.7). Note that the sum of empirical n_{ij} and theoretical $\hat{n}_{ij}$ is the same and amounts to 32,098.3 thousand individuals.

Table 3.9. Supporting calculations for Example 3.3

	n_{ij}	$\hat{n}_{ij}$	$\frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}}$
n_{11}	10,387.5	10,163.4	4.9
n_{21}	9,079.9	9,304.0	5.4
n_{12}	6,370.1	6,594.2	7.6
n_{22}	6,260.8	6,036.7	8.3
$\sum$	32,098.3	32,098.3	26.3

Source: own study.

The middle column was constructed based on the following calculations:

$$\hat{n}_{11} = \frac{n_{1\cdot} \cdot n_{\cdot 1}}{n} = \frac{19,467.4 \cdot 16,757.6}{32,098.3} = 10,163.4$$

$$\hat{n}_{12} = \frac{n_{1\cdot} \cdot n_{\cdot 2}}{n} = \frac{12,630.9 \cdot 16,757.6}{32,098.3} = 6,594.2$$

$$\hat{n}_{21} = \frac{n_{2\cdot} \cdot n_{\cdot 1}}{n} = \frac{19,467.4 \cdot 15,340.7}{32,098.3} = 9,304.0$$

$$\hat{n}_{22} = \frac{n_{2\cdot} \cdot n_{\cdot 2}}{n} = \frac{12{,}630.9 \cdot 15{,}340.7}{32{,}098.3} = 6{,}036.7$$

By substituting appropriate values into Formula (3.5), Tschuprow's T was determined:

$$T_{xy} = \sqrt{\frac{26.3}{32{,}098.3\sqrt{(2-1)\cdot(2-1)}}} = 0.03$$

Tschuprow's T at 0.03 shows that there is no relationship between gender and the place of residence of the population in Poland. This means that the gender structure is very similar in both rural and urban areas.

3.2.2. Spearman's rank correlation coefficient

Spearman's rank correlation coefficient R_{yx} is applied for assessing the strength and direction of a relationship when characteristics are quantitative, or qualitative, but they can be represented on an ordinal scale. The individual variants/values of the observed characteristic are given ranks, i.e., ordinal numbers. For example, the variants of the characteristic named 'attractiveness of investment location' are low, average, and high. Thus, the ranks are 1, 2, 3, respectively. Spearman's correlation coefficient is defined as:

$$R_{xy} = R_{yx} = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n^3 - n} \tag{3.8}$$

$$d_i = \text{rank}(x_i) - \text{rank}(y_i) \tag{3.9}$$

where:

d_i – difference in ranks of the two characteristics,

n – number of observations.

As noted before, the coefficient measures the strength and direction of a correlation ratio. It takes values in the range <–1, 1>. Spearman's rank correlation coefficient is symmetrical, and therefore $R_{xy} = R_{yx}$. This means that when we measure the dependence of characteristic Y on characteristic X, we follow the same procedure as for the dependence of characteristic X on characteristic Y. However, the researcher needs to know what the correct direction of dependence is.

Example 3.4

In order to answer the question of whether a country's wealth affects female fertility, the relationship between Total Fertility Rate (TFR) and the country's wealth is examined. The variable TFR represents the number of children per an average woman. It is a quantitative continuous variable. Wealth is expressed using three variants: low, medium, and high. It is a quasi-quantitative variable, i.e., its variants are expressed in words but can be ordered. The survey was conducted in 2022 for 15 European countries. Spearman's rank correlation coefficient was employed. Wealth statistics (X), TFR (Y) and calculations are summarised in Table 3.10.

Table 3.10. Wealth and Total Fertility Rate in selected European countries in 2022, and supporting calculations

No.	Country	TFR y_i	Wealth x_i	Rank y_i	Rank x_i	d_i	d_i^2
1	Czechia	1.64	low	13	1) 4	–9.0	81.0
2	Denmark	1.55	high	11	11) 13	2.0	4.0
3	Germany	1.46	high	9	12) 13	4.0	16.0
4	Greece	1.32	low	5) 5.5	2) 4	–1.5	2.3
5	Spain	1.16	average	1	8) 9	8.0	64.0
6	France	1.79	average	15	9) 9	–6.0	36.0
7	Italy	1.24	average	2	10) 9	7.0	49.0
8	Lithuania	1.27	low	3	3) 4	1.0	1.0
9	Hungary	1.56	low	12	4) 4	–8.0	64.0
10	Austria	1.41	high	7	13) 13	6.0	36.0
11	Poland	1.29	low	4	5) 4	0.0	0.0
12	Portugal	1.43	low	8	6) 4	–4.0	16.0
13	Romania	1.71	low	14	7) 4	–10.0	100.0
14	Finland	1.32	high	6) 5.5	14) 13	7.5	56.3
15	Sweden	1.53	high	10	15) 13	3.0	9.0
						$\sum$	534.5

Source: own compilation based on Eurostat data (variables demo_frate and tec00114), accessed 20/04/2024.

The ranks of characteristic x_i and y_i are the ordered from the lowest intensity of the characteristic to the highest. In the example, there are the so-called tied ranks, which occur when two or more values (in the case of a quantitative characteristic) or variants (in the case of a qualitative or quasi-quantitative characteristic) have the

same rank. Two countries (no. 4 and no. 14) had the same TFR values of 1.32 children per woman. They were listed as the 5th and 6th, and they are given a rank equal to the arithmetic mean of the two, i.e. $\frac{5+6}{2}=5.5$.

The repeated variants of the second variable, i.e. wealth, were ranked in a similar way. The countries no. 1, 4, 8, 9, 11, 12, and 13 had the same variant of wealth (low) and were ranked 4th because $\frac{1+2+3+4+5+6+7}{7}=4$. Another repetition occurred for countries listed as the 5th, 6th, and 7th (medium wealth). They were assigned a rank of 9 because $\frac{8+9+10}{3}=9$. The last repetition occurred regarding the third variant (high wealth) for countries no. 2, 3, 10, 14, and 15. They were ranked 13th because $\frac{11+12+13+14+15}{5}=13$.

In the column labelled d_i, the difference between the ranks of the two characteristics was calculated. In the last column d_i^2, the difference between the ranks was squared. By substituting the corresponding values into Formula (3.8), Spearman's rank correlation coefficient was calculated: $R_{xy}=1-\frac{6\cdot 534.5}{15^3-15}=0.05$. The coefficient R_{xy} = 0.05 indicates that there was no relationship between fertility and country wealth – high female fertility rates were reported in both poor (Romania) and rich (Denmark) European countries.

3.2.3. Pearson's linear correlation coefficient

Pearson's correlation coefficient r_{xy}, is the most used measure of the strength and direction of the linear relationship between two quantitative characteristics. This means that the shape of the relationship is best described by a straight line. Pearson's coefficient is calculated both when the data are in the form of a correlation series, and a correlation table. It is calculated from the formula:

$$r_{xy}=r_{yx}=\frac{\operatorname{cov}(x,y)}{S(x)\cdot S(y)} \tag{3.10}$$

where:

$\operatorname{cov}(x,y)$ – covariance between observed characteristics X and Y,
$S(x)$ – standard deviation of characteristic X,
$S(y)$ – standard deviation of characteristic Y.

Covariance means the difference between the mean of the product of the variables, and the product of the arithmetic means. It has no independent interpretation. It is calculated from the formula:

$$\operatorname{cov}(x,y) = \overline{x \cdot y} - \overline{x} \cdot \overline{y} \tag{3.11}$$

where the mean of the product is defined as:

– for a correlation series:

$$\overline{x \cdot y} = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{n} \tag{3.12}$$

– for a correlation table:

$$\overline{x \cdot y} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{r} x_i y_j n_{ij}}{n} \tag{3.13}$$

If the values of any of the characteristics in the correlation table are expressed by an interval, we use interval midpoints ($\dot{x}$, $\dot{y}$) in Formula 3.13. The values of Pearson's correlation coefficient always fall in the range <–1, 1>. The coefficient measures the strength and direction of the relationship. Its negative value indicates a negative relationship, while a positive value indicates a positive relationship. When Pearson's coefficient is equal to 0, it means that there is no correlation. The coefficient is symmetrical, i.e. $r_{xy} = r_{yx}$. In other words, it is a measure of the relationship when X influences Y and when Y influences X. The squared Pearson's correlation coefficient (r^2_{xy}) is called the coefficient of linear determination. It measures how good the fit of the theoretical regression function is to the empirical data (see Section 3.3). Expressed as a percentage, it tells you what percentage of changes in one characteristic can be attributed to changes in the other characteristic.

Example 3.5

Determine Pearson's correlation coefficient for GDP per capita Y (in PPS), and weekly working time in hours X, in 16 selected European countries in 2022. Statistics for the studied variables are summarised in Table 3.11.

Table 3.11. GDP per capita (PPS) and weekly working time (hours) in Europe in 2022, and supporting calculations

No.	Working time y_i	GDP per capita x_i	$x_i y_i$
1	36.9	62	2,287.80
2	39.8	90	3,582.00
3	35.4	136	4,814.40
4	35.3	117	4,130.10
5	41.0	67	2,747.00
6	37.8	86	3,250.80
7	37.4	100	3,740.00
8	37.4	97	3,627.80
9	39.2	89	3,488.80
10	39.6	76	3,009.60
11	36.0	124	4,464.00
12	40.4	79	3,191.60
13	40.0	79	3,160.00
14	40.2	76	3,055.20
15	36.2	110	3,982.00
16	38.9	119	4,629.10
Σ	611.5	1,507	57,160.20

Source: own compilation based on Eurostat data (lfsa_ewhun2 and tec00114), accessed 20/03/2024.

Based on Formula (2.20) and (2.59), the structure analysis parameters needed to calculate Pearson' correlation coefficient were determined. The mean level of GDP per capita was:

$$\overline{x} = \frac{1{,}507}{16} = 94.19 \text{ PPS}$$

The mean working time for the 16 countries was:

$$\overline{y} = \frac{611.50}{16} = 38.22 \text{ hours per week}$$

The standard deviation of GDP per capita was $S(x) = 21.03$ PPS, while the deviation of working time amounted to $S(y) = 1.84$ hours per week.

The covariance between the studied variables was determined with Formula (3.11):

$$\operatorname{cov}(x,y)=\frac{57{,}160.20}{16}-94.19\cdot 38.22=-27.43$$

The calculations needed for determining the mean of the product of the characteristics are shown in Table 3.11. Pearson's correlation coefficient was calculated using Formula (3.10):

$$r_{xy}=\frac{\operatorname{cov}(x,y)}{S(x)\cdot S(y)}=\frac{-27.43}{21.03\cdot 1.84}=-0.71$$

There was thus a moderately strong negative correlation between GDP per capita and working time in selected European countries in 2022, i.e., the higher the GDP per capita, the shorter the working time. Over half, $r_{xy}^2\cdot 100\%=50.41\%,$ of changes in working time can be explained by changes in wealth, as measured by GDP per capita.

Example 3.6

Calculating Pearson's correlation coefficient between the city area in km^2 (X) and the city population in thousands (Y), in 58 Polish cities in 2022. The statistics are presented as a correlation table in Table 3.12.

Table 3.12. City area in km^2 and city population (in thousands), and Poland in 2022

Area in km^2 x_i	$\dot{x}_i$	Population in thousands y_j				$n_{i.}$		
		50–100	100–200	200–500	500–1000			
		$\dot{y}_j$						
		75.0	150.0	350.0	750.0		$\dot{x}_i n_{i.}$	$(\dot{x}_i-\bar{x})^2 n_{i.}$
25.0–49.9	37.5	10	1	–	–	11	412.50	87,147.88
50.0–99.9	75.0	10	13	–	–	23	1,725.00	61,022.17
100.0–299.9	200.0	2	10	7	3	22	4,400.00	118,821.62
300.0–499.9	400.0	–	–	1	1	2	800.00	149,595.07
$n_{.j}$		22	24	8	4	58	7,337.50	416,586.75
$\dot{y}_j n_{.j}$		1,650	3,600	2,800	3,000	11,050.00		
$(\dot{y}_j-\bar{y})^2 n_{.j}$		293,573.13	39,399.52	203,478.00	1,252,083.83	1,788,534.48		

Source: own research based on data from Rocznik Statystyczny Rzeczypospolitej Polskiej 2023 (Statistical Yearbook of the Republic of Poland 2023).

First, the arithmetic means of the observed variables and their standard deviations were determined. Those measures were needed to calculate the covariance, which is the numerator in Formula (3.10) for Pearson's correlation coefficient.

When they have an interval series, as in the case of both variables in the example, the means of the intervals should be used in the calculations. The intermediate steps to determine the structure parameters are shown in Table 3.11. The arithmetic mean of the population was:

$$\overline{y} = \frac{\sum_{j=1}^{r} \dot{y}_j n_{\cdot j}}{n} = \frac{11{,}050}{58} = 190.52$$

The arithmetic mean area of the cities was calculated from the formula:

$$\overline{x} = \frac{\sum_{i=1}^{k} \dot{x}_i n_{i\cdot}}{n} = \frac{7{,}337.50}{58} = 126.51$$

The standard deviations of city area $S(x)$ and city population $S(y)$ were defined using the following formulas:

$$S(x) = \sqrt{\frac{\sum_{i=1}^{k} \left(\dot{x}_i - \overline{x}\right)^2 n_{i\cdot}}{n}} = \sqrt{\frac{416{,}586.75}{58}} = 84.75$$

and

$$S(y) = \sqrt{\frac{\sum_{j=1}^{r} \left(\dot{y}_j - \overline{y}\right)^2 n_{\cdot j}}{n}} = \sqrt{\frac{1{,}788{,}534.48}{58}} = 175.60$$

The mean of the product of the observed characteristics was:

$$\overline{x \cdot y} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{r} \dot{x}_i \dot{y}_j n_{ij}}{n} = \frac{1{,}946{,}250.00}{58} = 33{,}556.03$$

The supporting calculations to determine this value are summarised in Table 3.13.

Table 3.13. Supporting calculations to determine the mean of the product of city population (in thousands), and city area in km^2 in Poland in 2022

$\dot{x}_i$	$\dot{y}_j$	n_{ij}	$\dot{x}_i \cdot \dot{y}_j \cdot n_{ij}$
1	2	3	4
37.5	75.0	10	28,125.0
75.0	75.0	10	56,250.0
200.0	75.0	2	30,000.0

1	2	3	4
37.5	150.0	1	5,625.0
75.0	150.0	13	146,250.0
200.0	150.0	10	300,000.0
200.0	350.0	7	490,000.0
400.0	350.0	1	140,000.0
200.0	750.0	3	450,000.0
400.0	750.0	1	300,000.0
Σ			1,946,250.0

Source: own compilation based on data from Table 3.12.

Next, the covariance of the two studied characteristics was calculated:

$$\operatorname{cov}(x,y) = \overline{x \cdot y} - \overline{x} \cdot \overline{y} = 33{,}556.03 - 126.51 \cdot 190.52 = 9{,}453.34$$

By substituting the previous values into Formula (3.10), the result was:

$$r_{xy} = \frac{\operatorname{cov}(x,y)}{S(x) \cdot S(y)} = \frac{9{,}453.34}{84.75 \cdot 175.60} = 0.64$$

The above result shows that there was a moderate positive correlation between city population and city area. The larger the area, the more populous the city. The squared linear correlation coefficient was $0.64^2 = 0.40$, which means that changes in city area explained 40% of the changes in city population.

3.2.4. Correlation ratios

Correlation relations e_{yx} and e_{xy} measure the strength of correlation of two statistical characteristics. They do not measure the direction of the relationship. It is a measure that is not symmetric. This means that before starting the calculations, it is necessary to determine whether you are studying the dependence of a characteristic X on Y, or Y on X, and then choose the appropriate correlation ratio. When the correlation ratio of the dependent characteristic X to the independent characteristic Y *is* analysed, we use the formula:

$$e_{xy} = \sqrt{\frac{S^2(\overline{x}_j)}{S^2(x)}} \tag{3.14}$$

whereas when we are measuring the dependence of a characteristic Y on X, we use the formula:

$$e_{yx} = \sqrt{\frac{S^2(\bar{y}_i)}{S^2(y)}} \tag{3.15}$$

where:

e_{xy} – correlation ratio measuring the dependence of characteristic X on characteristic Y,

e_{yx} – correlation ratio measuring the dependence of characteristic Y on characteristic X,

$S^2(\bar{x}_j)$ – variance of conditional means of characteristic X,

$S^2(\bar{y}_i)$ – variance of conditional means of characteristic Y,

$S^2(x)$ – variance of characteristic X, $S^2(y)$ – variance of characteristic Y.

The variances of the conditional means X and Y are determined from the formulae, respectively:

$$S^2(\bar{x}_j) = \overline{\bar{x}_j^2} - (\bar{x})^2 \tag{3.16}$$

$$S^2(\bar{y}_i) = \overline{\bar{y}_i^2} - (\bar{y})^2 \tag{3.17}$$

where:

$(\bar{x})^2$ – squared mean of characteristic X,

$\overline{\bar{x}_j^2}$ – mean of squared conditional means of X,

$(\bar{y})^2$ – squared mean of Y,

$\overline{\bar{y}_i^2}$ – the mean of squared conditional means of Y.

Correlation ratios are applied to data presented in a correlation table. They can be used when the regression between the observed characteristics is both linear and non-linear. The condition for employing this measure is the quantitative nature of a dependent variable. Ratios take values in the range <0, 1>. They cannot take negative values and therefore do not indicate the direction of the relationship. They are asymmetric i.e. $e_{xy} \neq e_{yx}$. The square of the ratio (e^2_{xy}), expressed as a percentage, tells you what percentage of changes in characteristic X can be attributed to changes in characteristic Y. Similarly, the square of the ratio (e^2_{yx}), carries information on what proportion of changes in characteristic Y can be attributed to changes in characteristic X.

Example 3.7

Determine the correlation between the city size X, and the city population in thousands Y, of 58 Polish cities in 2022 (Table 3.14). As the independent variable X is a qualitative variable, use an appropriate correlation ratio.

Table 3.14. City population in thousands and city size in Poland in 2022, and supporting calculations

City size x_i	City population in thousands y_j				$n_{i\cdot}$		
	50–99	100–199	200–499	500–999			
	$\dot{y}_j$						
	75.0	150.0	350.0	750.0		$\overline{y}_i$	$\overline{y}_i^2 n_{i\cdot}$
Small	10	1	–	–	11	81.82	73,636.36
Average	10	13	–	–	23	117.39	316,956.52
Large	2	10	7	3	22	288.64	1,832,840.91
Very large	–	–	1	1	2	550.00	605,000.00
$n_{\cdot j}$	22	24	8	4	58	×	2,828,433.79
$\dot{y}_j n_{\cdot j}$	1,650	3,600	2,800	3,000	11,050.00		
$\left(\dot{y}_j - \overline{y}\right)^2 n_{\cdot j}$	293,573.13	39,399.52	203,478.00	1,252,083.83	1,788,534.48		

Source: own research based on data from Rocznik Statystyczny Rzeczypospolitej Polskiej 2023 (Statistical Yearbook of the Republic of Poland 2023).

The column $\overline{y}_i$ contains the values of the conditional means of the city population Y. The mean of the squared conditional means was determined as:

$$\overline{\overline{y}_i^2} = \frac{\sum_{i=1}^{k} \overline{y}_i^2 n_{i\cdot}}{n} = \frac{2{,}828{,}433.79}{58} = 48{,}766.10$$

To calculate the variance of the conditional means of the city population $S^2\left(\overline{y}_i\right)$, arithmetic mean of the population is needed:

$$\overline{y} = \frac{\sum_{j=1}^{r} \dot{y}_j n_{\cdot j}}{n} = \frac{11{,}050}{58} = 190.52$$

This means that the average population of a city in Poland in 2022 was 190,52 thousand people.

The variance of the conditional means of the city population was determined from Formula (3.15):

$$S^2(\bar{y}_i) = \overline{\bar{y}_i^2} - (\bar{y})^2 = 48,766.10 - 190.52^2 = 12,469.28$$

The standard deviation of the city population was then calculated:

$$S(y) = \sqrt{\frac{\sum_{j=1}^{r}(\dot{y}_j - \bar{y})^2 n_{\cdot j}}{n}} = \sqrt{\frac{1,788,534.48}{58}} = 175.60$$

By inserting the values into Formula (3.13), the correlation ratio of characteristic Y to characteristic X, i.e., the city population to the city size, was obtained:

$$e_{yx} = \sqrt{\frac{S^2(\bar{y}_i)}{S^2(y)}} = \sqrt{\frac{12,469.28}{175.60^2}} = 0.64$$

The measure shows a moderate strength of the correlation of the observed characteristics. In addition, it can be concluded that the variation in the city size explained 41% ($e_{yx}^2 = 0.41$) of the variation in the city population.

To summarise all four correlation coefficients described above, Table 3.14 compares their range (i.e. what values they can take), whether the measure is symmetrical or not and in which situation it can be used. The absolute value of the calculated coefficient close to 0 indicates no correlation; a value between 0.2 and 0.4 shows a weak correlation; a value close to 0.5 (between 0.4 and 0.6) is defined as a medium correlation; a value between 0.6 and 0.8 shows a strong correlation, while when the coefficient approaches 1 (it is above 0.8), a correlation is very strong.

Table 3.15. Comparison of correlation measures

Name	Range	Symmetry	Characteristics	Application
Tschuprow's T_{xy}	<0, 1> measures strength only	Yes $T_{xy} = Tyx$	mainly qualitative	correlation table
Spearman's rank correlation coefficient R_{xy}	<–1, 1> measures strength and direction	Yes $R_{xy} = R_{yx}$	quantitative qualitative (ordinal)	correlation series
Pearson's linear correlation coefficient r_{xy}	<–1, 1> measures strength and direction	Yes $r_{xy} = r_{yx}$	quantitative	correlation table, correlation series
Correlation ratio e_{xy}	<0, 1> measures strength only	No $e_{xy} \neq e_{yx}$	dependent quantitative	correlation table

Source: own study.

3.3. Linear regression

Regression analysis serves to express the relationship between statistical variables in mathematical language. For this purpose, a mathematical function $f(x)$ is found that best describes the relationship under study (Weinbach, Grinnell 2007; Hozer 1998). This function can take different forms, the simplest of which is the linear function. A regression function describes the relationship between two variables.

The purpose of regression analysis is to predict the value of the dependent variable Y at a given value of the independent characteristic X. However, it should be borne in mind that the predicted values are estimates, as the dependent variable may also be influenced by other factors not taken into account yet. Another important point is that a strong correlation ratio, confirmed by a well-fitting regression function, is not always a causal relationship. In other words, one variable is not always the cause, and the other is not always the effect. The observed relationship can only be determined as causal after logical justification based on the laws governing the field of science from which the phenomena under study have originated.

An example of a regression function is one describing the relationship between expenditure on luxury goods and income in Polish households. Another example is the regression of the mean rental price of a square metre of office space and the floor area of the surveyed properties.

The regression model can be written as:

$$y_i = \hat{y}_i + e_i \tag{3.18}$$

where:

y_i – empirical (real) values of variable Y,

$\hat{y}_i$ – theoretical values of variable Y,

e_i – values of random component.

The theoretical values of $\hat{y}_i$ denote the values obtained from the mathematical function: $\hat{y}_i = f(x_i)$. When differences between the empirical values of y_i and the theoretical values of $\hat{y}_i$ are small, the regression function is a good fit.

The linear regression function takes the form:

$$\hat{y}_i = a_y x_i + b_y \tag{3.19}$$

where:

$\hat{y}_i$ – theoretical values of dependent variable,

x_i – empirical values of independent variable,

a_y – directional coefficient of regression function (slope),

b_y – absolute term of regression function (intercept).

When estimating a linear regression function, its parameters a_y and b_y are determined. The following formulae are then used:

$$a_y = \frac{c(x,y)}{S^2(x)} = r_{yx} \cdot \frac{S(y)}{S(x)} \tag{3.20}$$

$$b_y = \overline{y} - a_y \cdot \overline{x} \tag{3.21}$$

where:

a_y – main parameter of regression function,
b_y – absolute parameter of regression function,
$c(x, y)$ – covariance between the variables Y and X under study,
$S^2(x)$ – variance of variable X,
r_{yx} – Pearson's linear correlation coefficient between variables Y and X,
$S(y)$ – standard deviation of variable Y,
$S(x)$ – standard deviation of variable X,
$\overline{y}$ i $\overline{x}$ – arithmetic means of variables Y and X.

The slope of a linear regression function a_y indicates by how much, on average, the values of variable Y will change (increase when a_y is positive or decrease when a_y is negative) when variable X increases by a unit. The intercept b_y indicates the value of variable Y when the variable X equals zero, but the interpretation of this parameter does not always make economic sense.

Two assumptions are made when determining the linear regression function. First, that the sum of residuals is equal to zero:

$$\sum e_i = \sum (y_i - \hat{y}_i) = 0 \tag{3.22}$$

where:

e_i – values of residual component,
y_i – empirical values of characteristic Y,
$\hat{y}_i$ – theoretical values of characteristic Y.

The second assumption is that the sum of squares of the residuals is minimal, i.e. any linear function with parameters other than the designated a_y and b_y would give a larger sum of the squared residuals:

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \min \tag{3.23}$$

The determined regression function can be a better or worse fit to the empirical data. This is informed by **measures of the goodness-of-fit**, which include the following:

- standard deviation of the residual value (S_e),
- coefficient of determination (R^2),
- Phi coefficient (φ^2).

The standard deviation of the residual value is defined by the formula:

$$S_e = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-2}} \tag{3.24}$$

where:

y_i – empirical values of characteristic Y,

$\hat{y}_i$ – theoretical values of characteristic Y,

n – number of units.

The $\boldsymbol{S_e}$ **fit measure** estimates how much, on average, the empirical values of the Y variable deviate from the theoretical values.

The **coefficient of linear determination**. $\boldsymbol{R}^2 = r_{xy}^2 \cdot 100\%$ denotes the percentage to which changes in one characteristic are explained by changes in another characteristic. The Phi coefficient (φ^2), or linear of determination coefficient, $\boldsymbol{\varphi}^2 = 100 - R^2$ indicates, by analogy with $\boldsymbol{R}^2$, to what degree changes in the dependent characteristic are not explained by changes in the independent characteristic.

When the dependent variable is X, the regression function $f(y)$ of the variable X against the variable Y should be determined:

$$\hat{x}_i = f\left(y\right) = a_x y_i + b_x \tag{3.25}$$

where:

$\hat{x}_i$ – theoretical values of the dependent variable X,

y_i – empirical values of the independent variable Y,

a_x – directional coefficient of the regression function,

b_x – absolute term of the regression function.

The structural parameters of this function are determined from the formulae:

$$a_x = \frac{c\left(x,y\right)}{S^2\left(y\right)} = r_{xy}\frac{S\left(x\right)}{S\left(y\right)} \tag{3.26}$$

$$b_x = \overline{x} - a_x \cdot \overline{y} \tag{3.27}$$

where:

a_x – main parameter of regression function,
b_x – absolute parameter of regression function,
$c(x, y)$ – covariance between variables Y and X,
$S^2(y)$ – variance of the independent variable Y,
r_{xy} – Pearson's coefficient of linear correlation between variables X and Y,
$S(x)$ – standard deviation of variable X,
$S(y)$ – standard deviation of variable Y,
$\overline{x}$ i $\overline{y}$ – arithmetic means of the observed variables.

Example 3.8

The data presented in a correlation series (Table 3.16) includes GDP per capita X (in PPS) and to working time in hours per week Y in 16 selected European countries in 2022. Using a linear regression function, the relationship between paid working time (dependent variable) and GDP per capita (independent variable) is described, and the fit of the regression function to the empirical data is examined. What working time can be expected for a country with a GDP per capita of 130 PPS?

Table 3.16. GDP per capita (in PPS) and weekly working time (in hours) in selected European countries in 2022, and supporting calculations

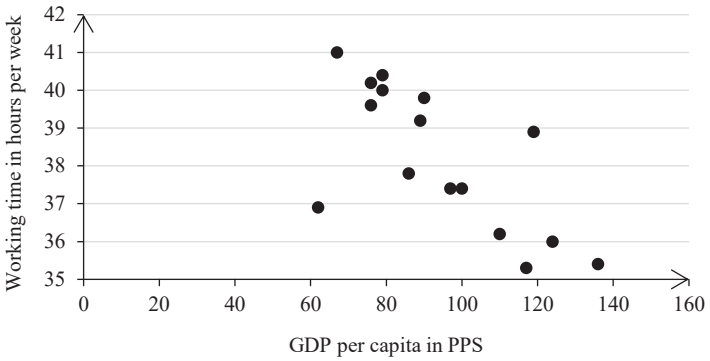
No.	Country	Working time (hours per week) y_i	GDP per capita (PPS) x_i	$\hat{y}_i$	e_i	e_i^2
1	2	3	4	5	6	7
1	Belgium	36.9	62	40.15	–3.25	10.56
2	Czechia	39.8	90	38.47	1.33	1.77
3	Denmark	35.4	136	35.71	–0.31	0.10
4	Germany	35.3	117	36.85	–1.55	2.40
5	Greece	41.0	67	39.85	1.15	1.32
6	Spain	37.8	86	38.71	–0.91	0.83
7	France	37.4	100	37.87	–0.47	0.22
8	Italy	37.4	97	38.05	–0.65	0.42
9	Lithuania	39.2	89	38.53	0.67	0.45
10	Hungary	39.6	76	39.31	0.29	0.08
11	Austria	36.0	124	36.43	–0.43	0.18
12	Poland	40.4	79	39.13	1.27	1.61

1	2	3	4	5	6	7
13	Portugal	40.0	79	39.13	0.87	0.76
14	Romania	40.2	76	39.31	0.89	0.79
15	Finland	36.2	110	37.27	–1.07	1.14
16	Sweden	38.9	119	36.73	2.17	4.71
Σ		611.5	1,507	611.50	0.00	27.36

Source: own compilation based on Eurostat data (lfsa_ewhun2 and tec00114), accessed 20/03/2024.

Figure 3.4 shows the relationship between working time and country wealth, as measured by GDP per capita in 16 selected European countries. The points are arranged in a decreasing order, meaning that the relationship between the variables under study is negative. Thus, as GDP per capita increased, paid working time decreased. As already noted in Subsection 3.1.1, people in wealthy countries work shorter hours.

Figure 3.4. Scatter plot of weekly working time (hours) and GDP per capita (PPS) in Europe in 2022



Source: own elaboration based on Table 3.15.

In order to determine the function parameters of the linear regression, selected measures of structure and correlation analysis are needed. These are presented in Table 3.17.

Table 3.17. Structure and correlation analysis parameters used in regression analysis

Parameter	Working time y_i	GDP per capita x_i
Arithmetic mean	38.22	94.19
Standard deviation	1.84	21.03
Variance	3.39	442.26
Covariance c(x, y)	–27.43	
Pearson's correlation coefficient r_{yx}	–0.71	

Source: own elaboration based on Table 3.15.

The directional coefficient of the regression function of working time against GDP per capita was calculated using the correlation coefficient and standard deviations of both characteristics:

$$a_y = r_{yx}\frac{S(y)}{S(x)} = -0.71 \cdot \frac{1.84}{21.03} = -0.06$$

This parameter indicates that when GDP per capita in selected European countries grows by 1 PPS, their population's working time decreases on average by 0.06 hours per week.

The absolute term of the regression function was calculated as follows:

$$b_y = \overline{y} - a_y\overline{x} = 38.22 - (-0.06) \cdot 94.19 = 43.87$$

In this example, the absolute parameter b_y has no interpretation; however, it can be interpreted in many other cases. It denotes the level of the dependent variable Y, while the dependent variable X takes the value 0. It is obvious that, in European countries, the GDP per capita variable cannot be equal to 0.

The linear regression function took the following form:

$$\hat{y}_i = -0.06 \cdot x_i + 43.87$$

This relationship is negative, as evidenced by the negative value of the main parameter of the regression function $a_y = -0.06$. It is worth noting that, in the regression analysis, the covariance between the variables under study, namely the Pearson correlation coefficient and the main parameter of the regression function, always have the same sign.

Next, the fit of the regression function to the empirical data was examined. The standard deviation of the residual value S_e was determined as well as the coefficient of determination R^2 and the Phi coefficient φ^2. Supporting calculations are presented in Table 3.15. In line with the assumptions of the method, the sum of theoretical values $\hat{y}_i$ equals the sum of the empirical values y_i (611.50), and the sum of the residuals is zero. The mean deviation of the empirical values from the theoretical values was determined as follows:

$$S_e = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-2}} = \sqrt{\frac{27.36}{16-2}} = 1.40$$

The calculated measure of fit indicates that the theoretical values of working time deviate from the empirical values by an average of ±1.40 hours per week. The linear coefficient of determination was calculated:

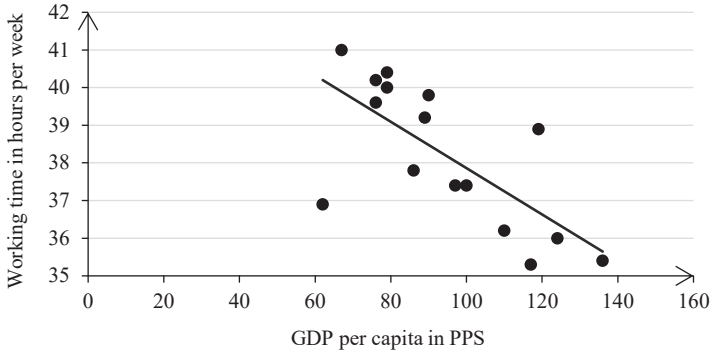
$$R^2 = r_{yx}^2 \cdot 100\% = (-0,71)^2 \cdot 100\ \% = 50\%$$

The measure shows that 50% of changes in working time can be explained by changes in GDP per capita. The Phi coefficient is:

$$\varphi^2 = 100\% - R^2 = 100\% - 50\% = 50\%$$

meaning that 50% of changes in working time in the observed European countries cannot be explained by changes in GDP per capita.

Figure 3.5. Empirical scatterplot and linear regression function of working time (hours per week) and GDP per capita in PPS in selected European countries in 2022



Source: own elaboration based on Table 3.15.

Figure 3.5 shows the relationship between wealth as measured by GDP per capita and working hours. The empirical values of y_i and x_i come from observations. The theoretical values $\hat{y}_i$ were obtained by substituting the empirical values x_i into the regression function. To draw the regression function on a coordinate system (black straight line in bold in Figure 3.5), we plot two points belonging to it and then we draw a straight line through them.

In order to answer the question of how long the expected working time should be in a country with a GDP per capita equal to 130 PPS, it is necessary to use the previously determined regression function. By substituting the independent variable x_i with a value of 130 we obtain: $\hat{y}_i(130) = -0.06 \cdot 130 + 43.87 = 36.07$. This means that, in a country with a GDP per capita of 130 PPS, the expected working time is 36.07 hours per week.

Chapter 4

Dynamics analysis

The dynamics analysis is a part of descriptive statistics that provides theories and methods for examining changes over time in socio-economic phenomena. To study the regularities of dynamics, statistical characteristics are presented in a time series of periods (streams over periods of time) or moments (assets at moments of time). We denote the studied variable (phenomenon) by Y_t, its realisations (empirical values) by y_t, and treat moments or periods as a population (Luszniewicz, Słaby 2008).

In dynamics analysis, we distinguish between the study of short-term changes (methods discussed in Section 4.1) and the study of long-term changes (methods discussed in Section 4.2).

Before discussing the relevant measures of dynamics, it is important to indicate how the average (mean) value in a time series is determined:

a) **arithmetic mean of** a time series of periods

$$\overline{y} = \frac{\sum_{t=1}^{n} y_t}{n} \tag{4.1}$$

b) **chronological mean** for a time series of moments

$$\overline{y}_{Ch} = \frac{\frac{1}{2}y_1 + y_2 + \ldots + y_{n-1} + \frac{1}{2}y_n}{n-1} \tag{4.2}$$

where:

y_t – value of the studied phenomenon in an interval/moment t,

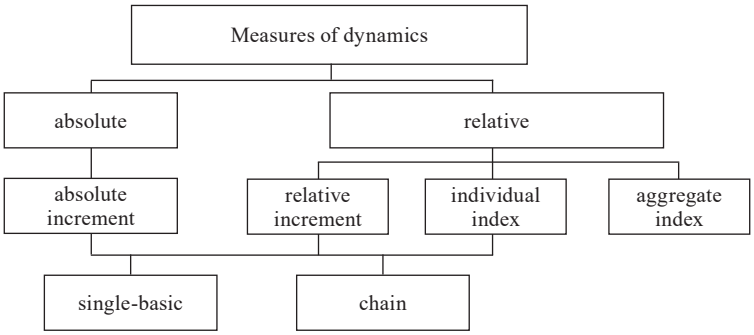
n – number of periods/moments.

The mean determined for the data in a time series represents the mean level of a phenomenon over an interval/moment.

4.1. Short-term change study

Measures of dynamics in the analysis of a short-term change are presented in Figure 4.1 and discussed in Sections 4.1.1 (individual growth and indices) and 4.1.2 (absolute and aggregate indices).

Figure 4.1. Measures of dynamics – examining short-term change



Source: own study.

Measures of dynamics in the study of short-term change are divided into absolute (expressed as a unit measure of the phenomenon) and relative (percentage-based) measures. The three types of measures – namely absolute growth, relative growth, and individual index – can be determined as single-basis or chained measures. The designation of single-basis measures means that the reference point is always a fixed interval/moment. In contrast, for chained measures, the reference interval/moment is variable (it is always the previous interval/moment).

4.1.1. Increments and individual indices

Measures that provide information on the change in the studied phenomenon over two periods/moments include: absolute increments, relative increments and individual indices. All these measures can be designated as single-basis or chained, depending on the reference point adopted. In the first case, the reference (base) is fixed. This is indicated by the notation, for example, in the form: year 2000 = 100, which means that the basis of comparison for all analysed years is the year 2000. In the case

of chained measures, on the other hand, the basis is variable, and it is the preceding interval/moment, which can be written, for example, as: preceding year = 100.

The formulae for the dynamic measures are as follows:

Absolute increments:

- Single-basis:

$$\Delta y_{t/0} = y_t - y_0 \tag{4.3}$$

- Chained:

$$\Delta y_{t/t-1} = y_t - y_{t-1} \tag{4.4}$$

Relative increments:

- Single-basis:

$$\frac{\Delta y_{t/0}}{y_0} = \frac{y_t - y_0}{y_0} \tag{4.5}$$

- Chained:

$$\frac{\Delta y_{t/t-1}}{y_{t-1}} = \frac{y_t - y_{t-1}}{y_{t-1}} \tag{4.6}$$

Individual indexes:

- Single-basis:

$$i_{t/0} = \frac{y_t}{y_0} \tag{4.7}$$

- Chained:

$$i_{t/t-1} = \frac{y_t}{y_{t-1}} \tag{4.8}$$

where:

y_t – the magnitude of the phenomenon in the studied interval/moment,

y_0 – the magnitude of the phenomenon in the baseline interval/moment,

y_{t-1} – the magnitude of the phenomenon in the preceding interval/moment.

In addition to the given measures that are determined for a given pair of periods/moments, we also determine the average levels of change for the entire studied period:

– an average absolute increment:

$$\overline{\Delta y}_{t/t-1} = \frac{y_n - y_1}{n-1} \tag{4.9}$$

– an average rate of change:

$$\overline{T} = \left(\overline{i}_{t/t-1}\right) \cdot 100\% \tag{4.10}$$

where:

y_n – the magnitude of the phenomenon in the last interval/moment,

y_1 – the magnitude of the phenomenon in the first interval/moment.

These measures reveal the average change (increase or decrease) from interval to interval (moment to moment). The former is an absolute measure, expressed as a unit of the studied phenomenon, and the latter (relative) is expressed as a percentage.

Ways to determine $\overline{i}_{t/t-1}$:

$$\overline{i}_{t/t-1} = \sqrt[n-1]{i_{2/1} \cdot i_{3/2} \cdot \ldots \cdot i_{n/n-1}} \tag{4.11}$$

$$\overline{i}_{t/t-1} = \sqrt[n-1]{i_{n/o}} \tag{4.12}$$

$$\overline{i}_{t/t-1} = \sqrt[n-1]{\frac{y_n}{y_1}} \tag{4.13}$$

Formula (4.11) uses chained index values, Formula (4.12) uses the value of the last single-basis index (condition: the basis is the first interval/moment under study) while Formula (4.13) uses the last and first empirical value in the time series.

The value $\overline{i}_{t/t-1}$ can also be used to determine the expected level of the studied phenomenon in future periods/moments (forecast):

$$\hat{y}_{t+k} = y_t \cdot \left(\overline{i}_{t/t-1}\right)^k, \quad \text{for } k = 1, 2, \ldots \tag{4.14}$$

The determination of the aforementioned measures and their interpretation will be shown through an empirical example.

Example 4.1

Table 4.1 shows data on the status of one of the targets of Sustainable Development Goals (SDG7): Renewable Energy Share in Total Final Energy Consumption (%) in Germany from 2014 to 2019 (first two columns). A comprehensive dynamic analysis was carried out. The following columns give the values of the calculated dynamics measures. For single-basis measures, the first year given in the table, i.e., 2014, was taken as the baseline (2014 = 100).

Table 4.1. Renewable Energy Share in Total Final Energy Consumption (%) in Germany, 2014–2019

Years	y_t	Absolute increments		Relative increments		Indexes	
		single-b. $\Delta y_{t/0}$	chained $\Delta y_{t/t-1}$	single-b. $\frac{\Delta y_{t/0}}{y_0}$	chained $\frac{\Delta y_{t/t-1}}{y_{t-1}}$	single-b. $i_{t/0}$	chained $i_{t/t-1}$
2014	14.0	0	–	0	–	1	–
2015	14.6	0.6	0.6	0.0429	0.0429	1.0429	1.0429
2016	14.2	0.2	–0.4	0.0143	–0.0274	1.0143	0.9726
2017	15.2	1.2	1.0	0.0857	0.0704	1.0857	1.0704
2018	16.1	2.1	0.9	0.1500	0.0592	1.1500	1.0592
2019	17.2	3.2	1.1	0.2286	0.0683	1.2286	1.0683

Source: Sustainable Development Report 2023, https://dashboards.sdgindex.org, accessed 10/02/2024.

The analysis began by determining the annual average value of the task throughout the studied years. The time series given is a series of moments i.e., we use the chronological mean (Formula (4.2)):

$$\overline{y}_{Ch} = \frac{\frac{14.0}{2} + 14.6 + 14.2 + 15.2 + 16.1 + \frac{17.2}{2}}{6-1} = \frac{75.7}{5} = 15.14$$

Thus, during the period 2014–2019, the average annual share of renewable energy in total energy consumption in Germany was 15.14%.

Next, we determine the measures of dynamics. The example below shows how to determine and interpret the measures calculated by Formulae (4.3)–(4.8) for 2019:

Absolute increment:

- Single-basis:

$$\Delta y_{t/0} = 17.2 - 14.0 = 3.2$$

Renewable energy share in total energy consumption in 2019 was 3.2 percentage points higher than in 2014.

- Chain:

$$\Delta y_{t/t-1} = 17.2 - 16.1 = 1.1$$

Renewable energy share in total energy consumption in 2019 was 1.1 percentage points higher than in 2018 (previous year).

Relative increment:

- Single-basis:

$$\frac{\Delta y_{t/0}}{y_0} = \frac{17.2 - 14.0}{14.0} = 0.2286$$

Renewable energy share in total energy consumption in 2019 was 22.86% higher than in 2014 (0.2286 · 100%).

- Chain:

$$\frac{\Delta y_{t/t-1}}{y_{t-1}} = \frac{17.2 - 16.1}{16.1} = 0.0683$$

Renewable energy share in total energy consumption in 2019 was 6.83% higher than in 2018 (previous year).

Individual indexes:

- Single-basis:

$$i_{t/0} = \frac{17.2}{14.0} = 1.2286$$

Renewable energy share in total energy consumption in 2019 was 122.86% of the 2014 share (1.2286 · 100%; momentum indicator) or renewable energy share in 2019 was 22.86% higher than in 2014 ((1.2286 – 1) · 100%).

- Chain:

$$i_{t/t-1} = \frac{17.2}{16.1} = 1.0683$$

Renewable energy share in total energy consumption in 2019 was 106.83% of the 2018 share (1.0683 · 100%; momentum indicator) or renewable energy share in 2019 was 6.83% higher than in 2018 ((1.0683 – 1) · 100%).

When analysing the values and the interpretation of the measures, the following relationships between the indices and relative increments can be noted:

$$i_{t/0} = 1 + \frac{\Delta y_{t/0}}{y_0} \tag{4.15}$$

$$i_{t/t-1} = 1 + \frac{\Delta y_{t/t-1}}{y_{t/t-1}} \tag{4.16}$$

which can be written as: index = 1 + relative increment.

We also determine the mean annual change in renewable energy share for the data in the presented time series:

We calculate the mean absolute increment by Formula (4.9):

$$\overline{\Delta y}_{t/t-1} = \frac{17.2 - 14.0}{6 - 1} = 0.64$$

Renewable energy share in total energy consumption in 2014–2019 grew by an average 0.64 percentage points each year.

The mean rate of change is determined by Formula (4.10), which requires calculating the value $\overline{i}_{t/t-1}$. We demonstrate this by calculating it in three ways using Formulae (4.11)–(4.13):

$$\overline{i}_{t/t-1} = \sqrt[6-1]{1.0429 \cdot 0.9726 \cdot 1.0704 \cdot 1.0592 \cdot 1.0683} = \sqrt[5]{1.2286} = 1.0420$$

$$\overline{i}_{t/t-1} = \sqrt[6-1]{1.2286} = \sqrt[5]{1.2286} = 1.0420$$

$$\overline{i}_{t/t-1} = \sqrt[6-1]{\frac{17.2}{14.0}} = \sqrt[5]{1.2286} = 1.0420$$

Regardless of how $\overline{i}_{t/t-1}$ has been determined, we obtain the value of 1.0420. We then establish the mean rate of change:

$$\overline{T} = (1.042 - 1) \cdot 100\% = 4.20\%$$

which means that between 2014 and 2019, the renewable energy share in total energy consumption in Germany grew by an average of 4.20% each year.

Finally, we determine the projected renewable energy share in total energy consumption in 2020 (provided that the observed trend has not changed):

$$\hat{y}_{2020} = y_{2019} \cdot \left(\overline{i}_{t/t-1}\right)^k$$

where $k = 1$ (forecast for the following year):

$$\hat{y}_{2020} = 17.2 \cdot (1.0420)^1 = 17.92$$

Thus, in 2020, in Germany, the projected renewable energy is expected to reach 17.92%, assuming a constant trend.[1]

---

[1] It should be noted that, in the absence of monotonicity in the development of the phenomenon, determining the mean rate of change may be questionable. Dividing the study interval into subperiods with uniform monotonicity may be helpful in this case.

Keeping in mind that the long-term objective for this indicator is a value of 55% (Sustainable Development Report 2023).

Example 4.2

Table 4.2 shows the time series for neonatal mortality (per 1,000 live births) in Greece from 2017 to 2021. This is a target under SDG3.

Table 4.2. Neonatal mortality rate in Greece (per 1,000 live births), 2017–2021

Years	2017	2018	2019	2020	2021
y_t	2.60	2.51	2.42	2.33	2.25

Source: Sustainable Development Report 2023, https://dashboards.sdgindex.org, accessed 10/02/2024.

Questions to be answered:

a) How has the neonatal mortality rate in Greece changed in 2020 compared to 2017?
b) What will be the projected neonatal mortality rate in Greece in 2022?

Answers:

(a) We compare 2020 with 2017 and calculate one absolute measure (absolute growth, a denominated measure) and one relative measure (relative growth or index; relative, dimensionless measures):

- Absolute increment

$$\Delta y_{2020/2017} = 2.33 - 2.60 = -0.27$$

- Index

$$i_{2020/2017} = \frac{2.33}{2.60} = 0.8962$$

The neonatal mortality rate in Greece in 2020, compared to 2017, was lower by 0.27 per 1,000 live births, or by 10.38%.

(b) We calculate the mean relative index:

$$\overline{i}_{t/t-1} = \sqrt[5-1]{\frac{2.25}{2.60}} = \sqrt[4]{0.8654} = 0.9645$$

The forecast for 2022:

$$\hat{y}_{2022} = 2.25 \cdot (0.9645)^1 = 2.17$$

The projected neonatal mortality rate in Greece in 2022 is 2.17 per 1,000 live births.

4.1.2. Aggregate indices (for absolute values)

Individual indices (i) discussed in Section 4.1.1 are used in the analysis of changes over time in a single product, good, or service. If the analysis concerns a group of products (goods or services) as a whole, then we use aggregate indices (I). The following formulae are applied in this case:

- Aggregate value index:

$$I_w = \frac{\sum w_{it}}{\sum w_{i0}} = \frac{\sum q_{it} p_{it}}{\sum q_{i0} p_{i0}} \tag{4.17}$$

- Aggregate price indices:

$$_L I_p = \frac{\sum q_{i0} p_{it}}{\sum q_{i0} p_{i0}} = \frac{\sum w_{i0} \left(i_p\right)_i}{\sum w_{i0}} \tag{4.18}$$

$$_P I_p = \frac{\sum q_{it} p_{it}}{\sum q_{it} p_{i0}} = \frac{\sum w_{it}}{\sum \frac{w_{it}}{\left(i_p\right)_i}} \tag{4.19}$$

$$_F I_p = \sqrt{_L I_p \cdot {}_P I_p} \tag{4.20}$$

- Aggregate quantity indexes:

$$_L I_q = \frac{\sum q_{it} p_{i0}}{\sum q_{i0} p_{i0}} = \frac{\sum w_{i0} \left(i_q\right)_i}{\sum w_{i0}} \tag{4.21}$$

$$_P I_q = \frac{\sum q_{it} p_{it}}{\sum q_{i0} p_{it}} = \frac{\sum w_{it}}{\sum \frac{w_{it}}{\left(i_q\right)_i}} \tag{4.22}$$

$$_F I_q = \sqrt{_L I_q \cdot {}_P I_q} \tag{4.23}$$

where:

w_{it} (w_{i0}) – value of product i in period t (compared 0),

q_{it} (q_{i0}) – quantity of product I in period t (compared 0),

p_{it} (p_{i0}) – price of te product i in period t (compared 0), where $w = q - p$, i_p – individual index of price,
i_q – individual index of quantity.

We consider one value index and three formulas for the price indices and for the quantity indices. When analysing changes in the prices of the studied products in total, we assume that their quantities are constant. This constancy can be adopted at the comparison period level 0 (Laspeyres' formula), at the level of the study period t (Paasche's formula), or at the average level (Fisher's formula; the geometric mean of the previous two indices). Similarly, when examining changes in quantities of products in total, we assume constancy in their prices.

It is also possible to use index equations (relationships) in the analysis:

$$i_w = i_p \cdot i_q \tag{4.24}$$

$$I_w = {}_L I_p \cdot {}_P I_q = {}_P I_p \cdot {}_L I_q = {}_F I_p \cdot {}_F I_q \tag{4.25}$$

The method of analysis using the above formulae is illustrated with examples.

Example 4.3

The average annual consumption per capita and average prices of selected dairy products in the Netherlands in 2014 and 2015 were examined (Table 4.3).

Table 4.3. Average annual consumption per capita and average prices of milk, eggs, and yoghurt in the Netherlands, 2014 and 2015

Product	Unit	Consumption q_0	Price (euro) p_0	Consumption q_t	Price (euro) p_t
		2014		2015	
Milk	l	33.9	0.95	34.5	0.93
Eggs	pcs.	210	0.18	210	0.20
Yoghurt	kg	15.1	0.94	15.4	0.89

Source: Eurostat (prc_dap14; prc_dap15) and https://www.statista.com, accessed 11.02.2024.

Examine:

(a) How have the price, consumption, and value of yoghurt consumed changed over the years?

(b) How has the value of consumption of the surveyed items changed in total?
(c) What was the impact of the change in the value of selected dairy products consumption in total, and what was the impact of the change in the volume of their consumption?

Solution:

(a) We consider one dairy product (yoghurt) and therefore using individual indices:

- Price:

$$i_{2015/2014} = \frac{0.89}{0.94} = 0.9468\,; \quad (0.9468 - 1) \cdot 100\% = -5.32\%$$

- Consumption:

$$i_{2015/2014} = \frac{15.4}{15.1} = 1.0199\,; \quad (1.0199 - 1) \cdot 100\% = 1.99\%$$

- Value:

$$i_{2015/2014} = \frac{13.71}{14.19} = 0.9656\,; \quad (0.9656 - 1) \cdot 100\% = -3.44\%$$

As compared to 2014, in 2015 the price (euro/kg) of yoghurt decreased by 5.32%, its consumption (kg/per capita) increased by 1.99% and the value (euro) of yoghurt consumed decreased by 3.44%.

(b) We use an aggregate index (three products in total) for the value of the dairy products consumed. We calculate the value as the product of quantity and price in 2014 and 2015 separately (Table 4.4):

$$I_w = \frac{\sum w_{it}}{\sum w_{i0}} = \frac{\sum q_{it} p_{it}}{\sum q_{i0} p_{i0}} = \frac{87.79}{84.20} = 1.0427\,; \quad (1.0427 - 1) \cdot 100\% = 4.27\%$$

Table 4.4. Supporting calculations for Example 4.3

Product	w_0 $q_0 p_0$	w_t $q_t p_t$	$q_0 p_t$	$q_t p_0$
Milk	32.21	32.09	31.53	32.78
Eggs	37.80	42.00	42.00	37.80
Yoghurt	14.19	13.71	13.44	14.48
Σ	84.20	87.79	86.97	85.05

Source: own calculations.

We find that the value of consumption of the observed products in total increased by 4.27% in 2015 compared to 2014.

(c) The data (Table 4.3) shows changes in both price and quantity in the consumption of the studied dairy products. We will examine what effect prices and quantity had on the change in values, respectively.

Price impact:

$$_{L}I_{p} = \frac{\sum q_{i0}p_{it}}{\sum q_{i0}p_{i0}} = \frac{86.97}{84.20} = 1.0329\,; \quad (1.0329 - 1) \cdot 100\% = 3.29\%$$

$$_{P}I_{p} = \frac{\sum q_{it}p_{it}}{\sum q_{it}p_{i0}} = \frac{87.79}{85.05} = 1.0322\,; \quad (1.0322 - 1) \cdot 100\% = 3.22\%$$

$$_{F}I_{p} = \sqrt{_{L}I_{p} \cdot {}_{P}I_{p}} = \sqrt{1.0329 \cdot 1.0322} = 1.0325\,; \quad (1.0325 - 1) \cdot 100\% = 3.25\%$$

As a result of the price change, the value of consumption of the observed products increased by 3.29%, assuming constant volumes (i.e. consumption) at 2014 levels.

Due to the change in prices, the value of consumption of the observed items rose by 3.22%, assuming constant volumes (consumption) at the 2015 level.

On average, the price change resulted in an increased value of consumption of the products studied of 3.25%, assuming volume (consumption) remained the same.

Impact of quantity (volume of consumption):

$$_{L}I_{q} = \frac{\sum q_{it}p_{i0}}{\sum q_{i0}p_{i0}} = \frac{85.05}{84.20} = 1.0101\,; \quad (1.0101 - 1) \cdot 100\% = 1.01\%$$

$$_{P}I_{q} = \frac{\sum q_{it}p_{it}}{\sum q_{i0}p_{it}} = \frac{87.79}{86.97} = 1.0095\,; \quad (1.0095 - 1) \cdot 100\% = 0.95\%$$

$$_{F}I_{q} = \sqrt{_{L}I_{q} \cdot {}_{P}I_{q}} = \sqrt{1.0101 \cdot 1.0095} = 1.0098\,; \quad (1.0098 - 1) \cdot 100\% = 0.98\%$$

Due to the change in quantity, the value of the observed dairy products rose by 1.01%, assuming prices stayed at 2014 levels.

As a result of the change in volume, the value of the products increased by 0.95%, assuming prices stayed at 2015 levels.

On average, due to the change in volume, the value of the studied dairy products grew by 0.98%, assuming prices remained constant.

In summary, price changes had a greater impact than changes in volume on the increase in the value of the studied dairy products.

Example 4.4

Annual expenditure and annual household consumption of fruit and vegetables in Spain in 2021 and 2022 were examined (Table 4.5).

Table 4.5. Annual household expenditure and consumption of fruit and vegetables in Spain, 2021–2022

Products	Expenses (thousand euro)		Consumption (thousand kilos)	
	2021 w_0	2022 w_t	2021 q_0	2022 q_t
Fruit	7,257	7,242	4,254	3,731
Vegetables	5,200	4,945	2,681	2,322

Source: https://www.fepex.es, accessed 12/02/2024.

Examine:

(a) How has the total annual household expenditure on fruit and vegetables changed in 2022 compared to 2021?

(b) How have the prices of the goods changed overall?

Solution:

(a) To examine the change in total annual expenditure on fruit and vegetables, we use an aggregate value index (expenditure was given in Table 4.5; we establish the sums of $\sum w_0 = 12{,}457$: and $\sum w_t = 12{,}187$):

$$I_w = \frac{\sum w_{it}}{\sum w_{i0}} = \frac{12{,}187}{12{,}457} = 0.9783\,; \quad (0.9783 - 1) \cdot 100\% = -2.17\%.$$

In comparison to 2021, spending on fruit and vegetables in 2022 decreased by 2.17%.

(b) To estimate aggregate price indices from Formulae (4.18–4.20), we first determine the prices of fruit and vegetables in 2021 and 2022 and then the values needed for the Formulae (Table 4.6).

Table 4.6. Supporting calculations for Example 4.4

Products	p_0 w_0/q_0	p_t w_t/q_t	$q_0\,p_t$	$q_t\,p_0$
Fruit	1.71	1.94	8,257.16	6,364.80
Vegetables	1.94	2.13	5,709.54	4,503.69
Σ			13,966.70	10,868.49

Source: own calculations.

$$_L I_p = \frac{\sum q_{i0} p_{it}}{\sum w_{i0}} = \frac{13{,}966.70}{12{,}457.00} = 1.1212\,; \quad (1.1212 - 1) \cdot 100\% = 12.12\%$$

$$_P I_p = \frac{\sum w_{it}}{\sum q_{it} p_{i0}} = \frac{12{,}187.00}{10{,}868.49} = 1.1213\,; \quad (1.1213 - 1) \cdot 100\% = 12.13\%$$

$$_F I_p = \sqrt{_L I_p \cdot {}_P I_p} = \sqrt{1.1212 \cdot 1.1213} = 1.12125\,; \quad (1.12125 - 1) \cdot 100\% = 12.125\%$$

In comparison to 2021, the prices of fruit and vegetables consumed in households in 2022 rose overall:

- by 12.12%, assuming that volumes (consumption) remained at 2021 levels,
- by 12.13%, assuming that volumes (consumption) remained at 2022 levels,
- by an average of 12.125%, assuming constant volumes (of consumption).

4.2. Long-term change study

When observing a time series over a long period of time, we can see that phenomena become stronger, weaker, or remain the same, with greater or lesser fluctuations. These systematic, unidirectional changes in a time series are referred to as a **trend** (development pattern). Any phenomenon viewed dynamically is the resultant of the action of major (fundamental) and random (side) causes. A trend is the effect of the major causes (Zeliaś 2000). Random causes, on the other hand, disrupt the systematic development of a phenomenon and cause fluctuations.

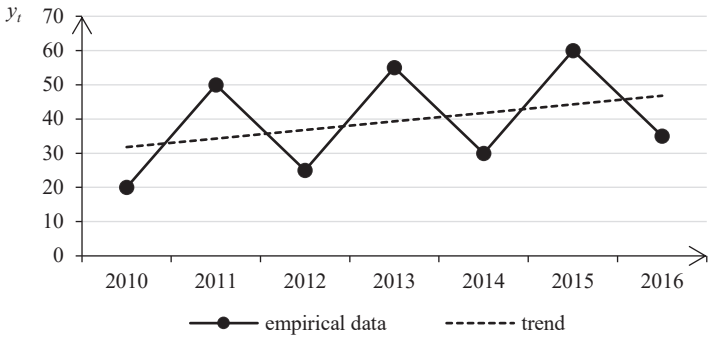
4.2.1. Trend-setting methods

We distinguish between the following trend-setting methods:

a) graphic,
b) mechanical,
c) analytical.

The graphic method consists of the 'freehand' determination of the development trend (a straight line or a curve). Figure 4.2 shows an example of time series data with a plotted line representing an upward trend. This method serves as a preliminary step for further analysis (a rough pattern analysis). By graphically presenting a trend using an Excel spreadsheet, we can simultaneously establish a mathematical function for this trend.

Figure 4.2. Example of trend set with graphical method



Source: own study.

The mechanical method consists of determining moving averages from several adjacent observations (periods/moments) in a time series. We distinguish the following types of moving averages:

a) simple,
b) centred.

Simple moving averages are calculated as 3- or 5-period averages and are used when the phenomenon is not seasonal. Table 4.7 shows data relevant to the implementation of a Target under SDG9: articles published in academic journals (per 1,000 population) in Belgium from 2013 to 2021. Due to the available annual data, we do not account for seasonality. Therefore, the trend set using the mechanical method can be determined as 3- or 5-period moving averages.

We calculate the former (3-period moving averages) by the formulae:

$$\overline{y}_2 = \frac{y_1 + y_2 + y_3}{3}, \quad \overline{y}_3 = \frac{y_2 + y_3 + y_4}{3}, \ldots \tag{4.26}$$

The latter (5-period averages) are calculated as:

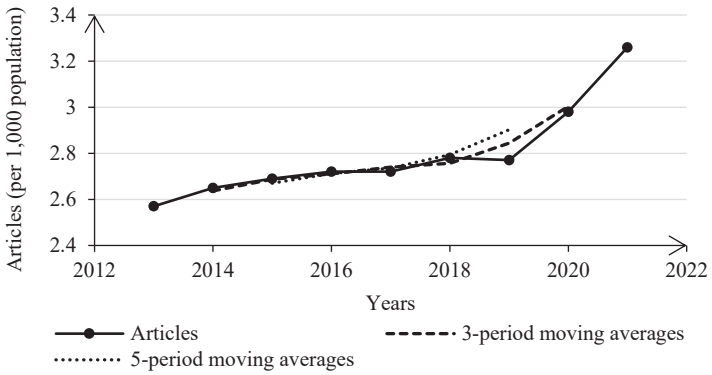
$$\overline{y}_3 = \frac{y_1 + y_2 + y_3 + y_4 + y_5}{5}, \quad \overline{y}_4 = \frac{y_2 + y_3 + y_4 + y_5 + y_6}{5}, \ldots \tag{4.27}$$

Table 4.7. Articles published in academic journals (per 1,000 population) in Belgium, 2013–2021, and designated simple moving averages

Years	Articles	Simple moving averages	
		3-period	5-period
2013	2.57	–	–
2014	2.65	2.64	–
2015	2.69	2.69	2.67
2016	2.72	2.71	2.71
2017	2.72	2.74	2.74
2018	2.78	2.76	2.79
2019	2.77	2.84	2.90
2020	2.98	3.00	–
2021	3.26	–	–

Source: Sustainable Development Report 2023, https://dashboards.sdgindex.org and own calculations, accessed 15/02/2024.

Figure 4.3. Articles published in academic journals (per 1,000 population) in Belgium, 2013–2021, and designated 3- and 5-period simple moving averages



Source: own elaboration based on Table 4.7.

When data in a time series is seasonal (more on this in Section 4.2.2), we determine the trend by means of centred moving averages. These are so-called 2m-period averages, where *m* denotes the number of sub-periods in a year.

Table 4.8 shows data on the number of flats completed in Poland in six-month periods from 2018 to 2021. Due to the data being less than annual, we take their seasonality into consideration. Table 4.8 and Figure 4.4 show that the observed phenomenon takes on smaller values in the first half-years and larger values in the second half-years. Thus, the conclusion is that the analysed time series is characterised by seasonality, and therefore, the trend set by the mechanical method was determined as the *2m-period* centred moving averages.

Here, $m = 2$ (2 half-years per year).

The centred 2*m-period* moving averages were estimated by the Formulae:

$$\overline{y}_2 = \frac{y_1 + 2y_2 + y_3}{4}, \quad \overline{y}_3 = \frac{y_2 + 2y_3 + y_4}{4}, \ldots \tag{4.28}$$

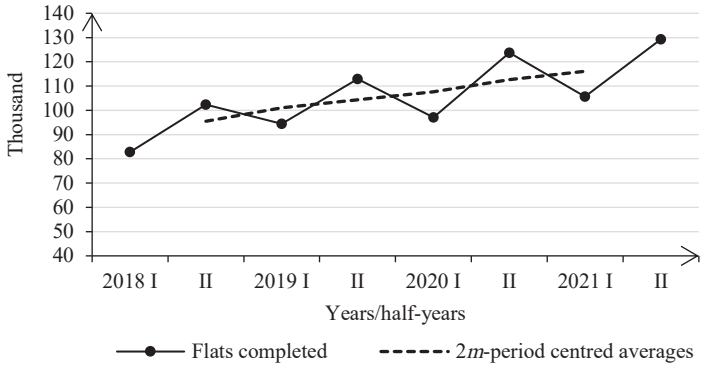
Table 4.8. Flats completed in Poland in the 2018–2021 half-years (in thousands) and the designated centred moving averages

Years/half-years	Flats (thousands)	2*m*-period centred averages
2018 I	82.8	–
II	102.3	95.5
2019 I	94.5	101.1
II	112.9	104.4
2020 I	97.1	107.7
II	123.8	112.6
2021 I	105.6	116.1
II	129.3	–

Source: CSO. Local Data Bank. Accessed 20/11/2021.

The advantage of the mechanical method is the simplicity of the calculations (averages). On the other hand, its shortcomings include the truncation of the time series (evident in Tables 4.7 and 4.8 and in Figures 4.3 and 4.4) and the lack of forecasting capability.

Figure 4.4. Time series data (Table 4.8) and 2*m*-period centred moving averages



Source: own study.

The analytical method consists of determining a mathematical function that describes the dependence of a socio-economic phenomenon observed over time on a time variable *t* (measuring periods or moments of time). The trend function can be linear or non-linear.[2]

We write **the linear trend function** in two ways. depending on the adopted condition for the time variable *t*:

$$\hat{y}_t = a_1 \cdot t + a_0; \quad t = 1, 2, \ ..., \ n \tag{4.29}$$

or:

$$\hat{y}_t = a_1 \cdot t + a_0; \quad \sum t = 0 \tag{4.30}$$

In these notations, we adopt that $\hat{y}_t$ is the theoretical value of the studied phenomenon over time; a_1 and a_0 are parameters of the linear trend function; *t* is the time variable; *n* is the number of periods/moments in the time series. The two notations differ in the way how the time variable *t* is adopted (the specific method is decided by the researcher).[3] These methods are shown as the empirical examples (Table 4.9).

When time series are characterised by seasonality, Formulae (4.29) and (4.30) can also be written in the form:

$$\hat{x}_{ik} = a_1 \cdot t_{ik} + a_0; \ t_{ik} = 1, 2, \ ..., \ n \cdot m \tag{4.29a}$$

---

[2] The non-linear trend functions include power, exponential, and logarithmic functions.

[3] The notation $\sum t = 0$ may be replaced by $\sum (t - \overline{t}) = 0$ (Bak et al. 2021).

or:

$$\hat{x}_{ik} = a_1 \cdot t_{ik} + a_0; \quad \sum t_{ik} = 0 \tag{4.30a}$$

where:

$\hat{x}_{ik}$ – trend value,

a_1 and a_0 – parameters of the linear trend function,

t_{ik} – time variable,

n – number of years,

m – number of sub-periods per year.

Table 4.9. Modes of adopting time variable *t*

$t = 1, 2, ..., n$	
years	t
2016	1
2017	2
2018	3
2019	4
2020	5
2021	6
2022	7

$\Sigma t = 0$ n odd	
years	t
2016	–3
2017	–2
2018	–1
2019	0
2020	1
2021	2
2022	3
Σ	0

$\Sigma t = 0$ n even	
years	t
2015	–3.5
2016	–2.5
2017	–1.5
2018	–0.5
2019	0.5
2020	1.5
2021	2.5
2022	3.5
Σ	0.0

Source: own study.

The parameters of the linear trend function are estimated by the following Formulae:

- For $t = 1, 2, \ldots, n$: Formula (4.29):

$$a_1 = \frac{\overline{y_t \cdot t} - \overline{y}_t \cdot \overline{t}}{\overline{t^2} - \left(\overline{t}\right)^2} \tag{4.31}$$

$$a_0 = \overline{y} - a_1 \cdot \overline{t} \tag{4.32}$$

- For $\sum t = 0$: Formula (4.30):

$$a_1 = \frac{\overline{y \cdot t}}{\overline{t^2}} = \frac{\sum y_t \cdot t}{\sum t^2} \tag{4.33}$$

$$a_0 = \overline{y} = \frac{\sum y_t}{n} \tag{4.34}$$

The value of the slope parameter a_1, determined by both methods (Formulas 4.31 and 4.33), is the same, and so is its interpretation: the average increase/decrease of the phenomenon under study in successive periods/moments. However, the intercept parameter a_0 takes on different values and is interpreted differently.

– In the former case (4.32), it is the value of the trend in the period preceding the study (for $t = 0$).
– In the latter case (4.34), it is the average level of the observed phenomenon.[4]

The methods for determining the parameters of a linear trend are shown in Example 4.5.

To determine the fit of the selected trend function (e.g. linear) to the empirical data, measures of goodness-of-fit are determined:

1. Coefficient of convergence (conformity, indeterminacy):

$$\varphi^2 = \frac{\sum_{t=1}^{n}\left(y_t - \hat{y}_t\right)^2}{\sum_{t=1}^{n}\left(y_t - \overline{y}\right)^2} \tag{4.35}$$

where:

y_t – empirical values,
$\hat{y}_t$ – theoretical values (trend),
$\overline{y}$ – mean of empirical values.

2. Coefficient of determinacy:

$$R^2 = 1 - \varphi^2 \tag{4.36}$$

These coefficients are complementary. The former tells us to what percentage (multiplied by 100%) the trend function does not explain the variability of the phenomenon over time, while the latter tells us to what percentage it explains this variability.

Example 4.5

The time series shows the amount of non-recycled municipal solid waste (kg/capita/day) in France from 2016 to 2021 (Table 4.10).

[4] With the assumption that $\Sigma t = 0$ for a series of moments, the average level is generally determined as a chronological average; therefore, the interpretation here is not exact.

Table 4.10. Non-recycled municipal solid waste (kg/capita/day) in France, 2016–2021

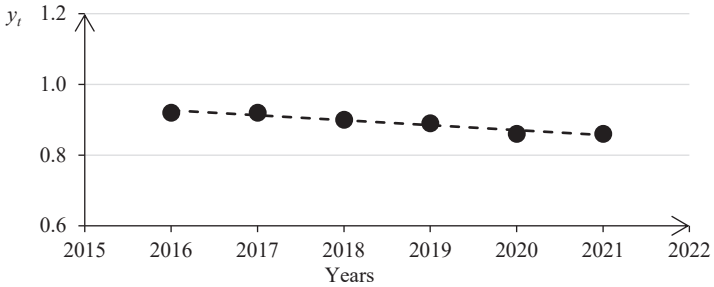
Years	2016	2017	2018	2019	2020	2021
Municipal solid waste	0.92	0.92	0.90	0.89	0.86	0.86

Source: Sustainable Development Report 2023, https://dashboards.sdgindex.org, accessed 15/02/2024.

Task: Determine analytically the linear trend parameters for $t = 1, 2, ..., n$ and for $\sum t = 0$.

The time series data and the trend line are shown in Figure 4.5. It should be noted that the trend of the studied phenomenon is downward.

Figure 4.5. Empirical values and trend line – non-recycled municipal solid waste (kg/capita/day) in France, 2016–2021



Source: own study.

The parameters of the linear trend function were determined for both methods of taking the time variable *t*. The necessary calculations are summarised in Table 4.11.

Table 4.11. Supporting calculations for Example 4.5

Date		Method 1			Method 2		
year	y_t	t	$y_t \cdot t$	t^2	t	$y_t \cdot t$	t^2
2014	0.92	1	0.9	1	–2.5	–2.3	6.25
2015	0.92	2	1.8	4	–1.5	–1.4	2.25
2016	0.90	3	2.7	9	–0.5	–0.5	0.25
2017	0.89	4	3.6	16	0.5	0.4	0.25
2018	0.86	5	4.3	25	1.5	1.3	2.25
2019	0.86	6	5.2	36	2.5	2.2	6.25
Σ	5.35	21	18.5	91	0	–0.2	17.50

Source: own calculations.

Method 1

$$\hat{y}_t = a_1 \cdot t + a_0; \quad t = 1, 2, ..., n \qquad \text{(Formula 4.29)}$$

$$\overline{y_t \cdot t} = \frac{\sum y_t \cdot t}{n} = \frac{18.5}{6} = 3.08, \quad \overline{y} = \frac{\sum y_t}{n} = \frac{5.35}{6} = 0.89$$

$$\overline{t} = \frac{\sum t}{n} = \frac{21}{6} = 3.50, \quad \overline{t^2} = \frac{\sum t^2}{n} = \frac{91}{6} = 15.17$$

$$a_1 = \frac{\overline{y_t \cdot t} - \overline{y}_t \cdot \overline{t}}{\overline{t^2} - (\overline{t})^2} = \frac{3.08 - 0.89 \cdot 3.5}{15.17 - 12.25} = -0.01$$

$$a_0 = \overline{y} - a_1 \cdot \overline{t} = 0.89 - (-0.01) \cdot 3.5 = 0.93$$

$$\hat{y}_t = -0.01 \cdot t + 0.93; \quad t = 1, 2, ..., n.$$

Interpretation of parameters:

a_1: The amount of non-recycled municipal solid waste in France from 2016 to 2021 decreased each year by an average of 0.01 kg/capita/day,

a_0: The trend value (theoretical) in 2015 (t = 0) was 0.93 kg/capita/day.

Method 2

$$\hat{y}_t = a_1 \cdot t + a_0; \quad \sum t = 0 \qquad \text{(Formula 4.30)}$$

$$a_1 = \frac{\sum y_t \cdot t}{\sum t^2} = \frac{-0.20}{17.5} = -0.01, \quad a_0 = \overline{y} = \frac{\sum y_t}{n} = \frac{5.35}{6} = 0.89$$

$$\hat{y}_t = -0.01 \cdot t + 0.89; \quad \sum t = 0$$

Interpretation of parameters:

a_1: The amount of non-recycled municipal solid waste in France in 2016–2021 decreased by an average of 0.01 kg/capita/day (same interpretation as in Method 1).

a_0: The amount of non-recycled municipal solid waste in France from 2016 to 2021 averaged 0.89 kg/capita/day.

The fitting characteristics of the linear trend function to the empirical data were also determined using Formulae (4.35) and (4.36). The calculations are included in Table 4.12 (fit). Theoretical values $\hat{y}_t$ were determined from the trend function (they were the same for the condition $t = 1, 2, ..., n$ and for $\Sigma t = 0$).

The fit measures are:

- The coefficient of convergence:

$$\varphi^2 = \frac{\sum_{t=1}^{n}\left(y_t - \hat{y}_t\right)^2}{\sum_{t=1}^{n}\left(y_t - \bar{y}\right)^2} = \frac{0.0006}{0.0037} = 0.1622$$

which means that the linear trend function does not explain in 16.22% of the variation in the amount of non-recycled municipal solid waste in France from 2016 to 2021.

- The coefficient of determination:

$$R^2 = 1 - \varphi^2 = 1 - 0.1622 = 0.8378$$

which means that the linear trend function explains 83.78% of the variation in the amount of non-recycled municipal solid waste in France from 2016 to 2021 (i.e., the fit to the empirical data is 83.78%).

Table 4.12. Supporting calculations for Example 4.5 (fit)

y_t	Fit		
	$\hat{y}_t$	$\left(y_t - \hat{y}_t\right)^2$	$\left(y_t - \bar{y}\right)^2$
0.92	0.92	0.0000	0.0009
0.92	0.91	0.0001	0.0009
0.90	0.90	0.0000	0.0001
0.89	0.89	0.0000	0.0000
0.86	0.88	0.0004	0.0009
0.86	0.87	0.0001	0.0009
	Σ	0.0006	0.0037

Source: own calculations.

The trend function for data relating to periods/moments shorter than one year (e.g., half-yearly or quarterly data) can be determined:

a) In a **single-step procedure** where a trend is set directly from a time series:

$$\hat{y}_t = a_1 \cdot t + a_0 \qquad \text{(Formula 4.29 or 4.30)}$$

b) In a **two-step procedure** where the conditions are: the time series of periods and $\sum t = 0$:

- First we set the trend for annual data:

$$\hat{Y}_T = A_1 \cdot T + A_0 , \quad \sum T = 0$$

(this is Formula 4.30 written in capital letters for distinction).

- Then we set the *m*-interval trend (*m* is the number of sub-periods in a year, e.g. 2 for half-years or 4 for quarters):

$$\hat{y}_t = a_1 \cdot t + a_0 , \quad \sum t = 0 \qquad \text{(Formula 4.30)}$$

where:

$$a_1 = \frac{A_1}{m^2}$$

$$a_0 = \frac{A_0}{m}$$

The application of the two-step procedure is shown in Example 4.6.

Example 4.6

Using the annual data in the time series, determine a linear six-month trend.

Table 4.13. Number of tourist trips by Germans within Europe, 2014–2019 (million), and supporting calculations

Year	Y_T	T	$Y_T \cdot T$	T^2
2014	236.91	–2.5	–592.28	6.25
2015	247.88	–1.5	–371.82	2.25
2016	255.65	–0.5	–127.83	0.25
2017	243.58	0.5	121.79	0.25
2018	267.88	1.5	401.82	2.25
2019	260.52	2.5	651.30	6.25
Σ	1,512.42	0.0	82.99	17.5

Source: https://ec.europa.eu/eurostat/web/main/data/database (online data code: tour_dem_ttw).

Since the series under consideration is a time series of periods and we apply the condition on the time variable *t* as $\sum t = 0$, we can use a two-step procedure to determine the trend.

We first calculate the annual trend (as we have annual data):

$$\hat{Y}_T = A_1 \cdot T + A_0; \quad \sum T = 0$$

$$A_1 = \frac{\sum Y_T \cdot T}{\sum t^2} = \frac{82.99}{17.5} = 4.74, \quad A_0 = \overline{Y} = \frac{\sum y_t}{n} = \frac{1{,}512.42}{6} = 252.07$$

$$\hat{Y}_T = 4.74 \cdot T + 252.07; \quad \sum t = 0$$

Interpretation of annual trend parameters:

A_1: The number of tourist trips by Germans within Europe increased by an average of 4.74 million each year between 2014 and 2019,

A_0: An average of 252.07 million Germans travelled within Europe each year between 2014 and 2019.

We then determine a six-month trend ($m = 2$):

$$a_1 = \frac{A_1}{m^2} = \frac{4.74}{4} = 1.19, \quad a_0 = \frac{A_0}{m} = \frac{252.07}{2} = 126.04$$

$$\hat{y}_t = 1.19 \cdot t + 126.04, \quad \sum t = 0$$

In this manner, using a time series with annual data, a linear semi-annual trend function was obtained.

Interpretation of the semi-annual (m-interval) trend parameters:

a_1: The number of tourist trips made by Germans within Europe between 2014 and 2019 increased from half-year to half-year by an average of 1.19 million,

a_0: Between 2014 and 2019, Germans made an average of 126.04 million trips within Europe every six months.

4.2.2. Seasonality testing methods

Data in a time series, apart from the fundamental direction of the phenomenon under study (a trend), can be characterised by the following fluctuations:

- random,
- seasonal – fluctuations within a year,
- cyclical – fluctuations in economic phenomena with cycles lasting from a few to more than ten years.

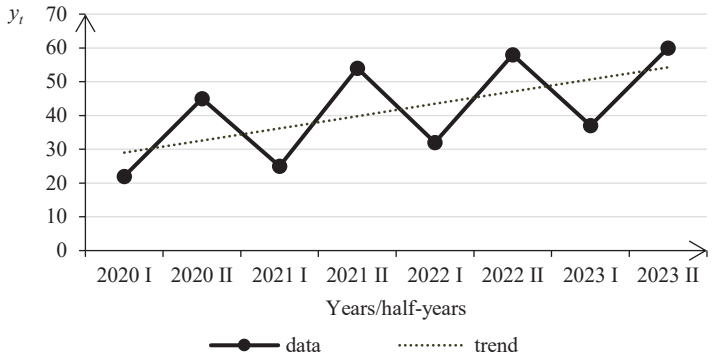
In our considerations, we will deal with seasonal fluctuations, i.e., fluctuations that recur regularly at annual periods. They are the result of permanent causes of a natural, economic and social nature (Zeliaś 2000).

There are two types of seasonal fluctuations: additive and multiplicative.

Additive (absolute) fluctuations are shown in Figure 4.6. Deviations of the magnitude of the observed phenomenon (y_t) from the trend (dashed line) are similar in given sub-periods (e.g., half-years).

It can be concluded that the phenomenon in the first half-years is weaker than the trend value, and in the second half-years it is stronger than the trend value (seasonality).

Figure 4.6. Additive fluctuations – observation of data in the 2020–2023 by half-years



Source: own study.

In this case, the seasonality is measured with the **seasonality components S_k** determined from the formulae:

- If the trend was determined mechanically:

$$S_k = \frac{\sum_{i=1}^{n-1}\left(y_{ik} - \hat{x}_{ik}\right)}{n-1} \tag{4.37}$$

- If the trend is determined analytically:

$$S_k = \frac{\sum_{i=1}^{n}\left(y_{ik} - \hat{x}_{ik}\right)}{n} \tag{4.38}$$

where:

$i = 1, ..., n$ (n is the number of years under observation),

$k = 1, ..., m$ (m is the number of sub-periods in a year, or seasons),

n – the number of identical periods, usually being the number of observation years but, in the case of mechanical trend determination, the series is shortened (hence $n - 1$),
y_{ik} – the empirical value in year i and in sub-interval k,
$\hat{x}_{ik}$ – the trend value in year i and in subperiod k.

After determining m components of seasonality (e.g., for half-years $m = 2$), it is necessary to check that their sum is zero ($\sum S_k = 0$). If not ($\sum S_k \neq 0$), then the components must be adjusted:

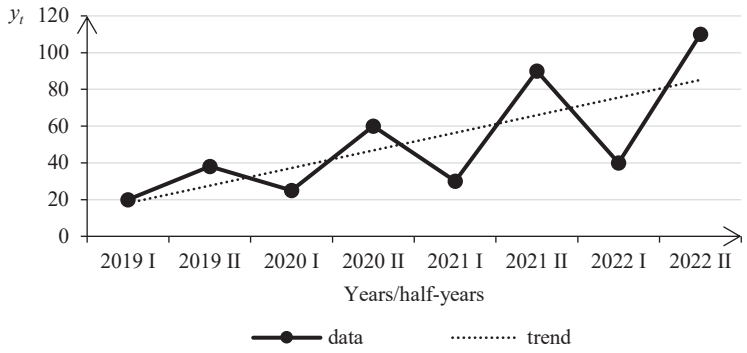
$$ {}_0S_k = S_k - k_S \tag{4.39}$$

where:
${}_0S_k$ – adjusted component,
k_S – adjustment measure:

$$k_s = \frac{\sum_{k=1}^{m} S_k}{m} \tag{4.40}$$

Multiplicative (relative) fluctuations are shown in Figure 4.7. Deviations of the magnitude of the phenomenon under study (y_t) from the trend (dashed line) in given sub-periods (half-years) are not similar (in this particular case, they are increasing). It can be concluded that the phenomenon in the first half-years is weaker than the trend value, and in the second half-years it is stronger than the trend value (seasonality).

Figure 4.7. Multiplicative fluctuations – observation of data in the 2019–2022 by half-years



Source: own study.

In this case, the measures of seasonality are the **seasonality indices W_k**, determined from the following formulas:

- If the trend was determined mechanically:

$$W_k = \frac{\sum_{i=1}^{n-1} y_{ik}}{\sum_{i=1}^{n-1} \hat{x}_{ik}} \tag{4.41}$$

- If the trend is determined analytically:

$$W_k = \frac{\sum_{i=1}^{n} y_{ik}}{\sum_{i=1}^{n} \hat{x}_{ik}} \tag{4.42}$$

where:

$i = 1, ..., n$ (n is the number of observation years),

$k = 1, ..., m$ (m is the number of sub-periods in a year, or seasons),

n – the number of the same periods, usually the number of observation years, but in the case of mechanical trend determination, the series is shortened (hence $n - 1$),

y_{ik} – the empirical value in year i and in sub-interval k,

$\hat{x}_{ik}$ – trend value in year i and in sub-interval k.

After determining m (for half-years $m = 2$) seasonality indices, we check that their sum is m ($\sum W_k = m$).

- If not ($\sum W_k \neq m$) then the indices are adjusted as follows:

$$_0W_k = W_k \cdot k_w \tag{4.43}$$

where:

$_0W_k$ – adjusted index,

k_w – adjustment measure:

$$k_w = \frac{m}{\sum_{k=1}^{m} W_k} \tag{4.44}$$

Example 4.7

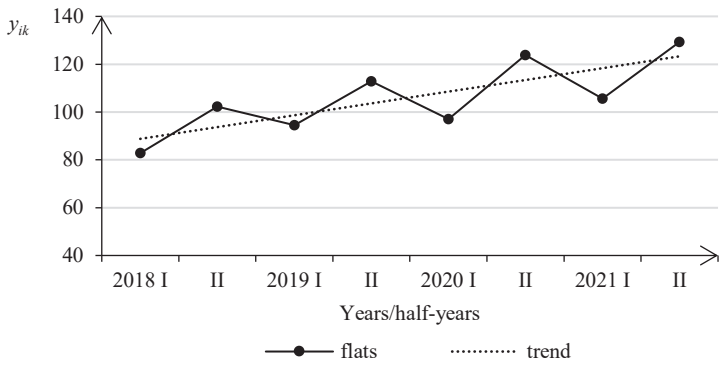
For the given time series data (Table 4.14), determine the appropriate measures of seasonality.

Table 4.14. Flats completed in Poland, 2018–2021 by half-years (thousands), and supporting calculations

Years/ half-years	Flats (thousands) y_{ik}	t_{ik}	$y_{ik} \cdot t_{ik}$	t_{ik}^2	$\hat{x}_{ik}$	$y_{ik} - \hat{x}_{ik}$
2018 I	82.8	–3.5	–289.80	12.25	88.79	–5.99
II	102.3	–2.5	–255.75	6.25	93.72	8.58
2019 I	94.5	–1.5	–141.75	2.25	98.65	–4.15
II	112.9	–0.5	–56.45	0.25	103.58	9.32
2020 I	97.1	0.5	48.55	0.25	108.51	–11.41
II	123.8	1.5	185.70	2.25	113.44	10.36
2021 I	105.6	2.5	264.00	6.25	118.37	–12.77
II	129.3	3.5	452.55	12.25	123.30	6.00
Σ	848.3	0.0	207.05	42.00	–	–

Source: CSO. Local Data Bank. Accessed 20/01/2022.

Figure 4.8. Empirical data of the time series (Table 4.14) and determined trend line



Source: own study.

The data presentation in Figure 4.8 shows that the time series is characterised by additive fluctuations. It is therefore necessary to determine the seasonality components S_k (Formula 4.38) for the first and second half-years. Firstly, the trend line

must be determined. In this case, Formula (4.30a) is adopted (with the assumption of $\sum t_{ik} = 0$). The calculations can be found in Table 4.14 and below:

- Trend line:

$$a_1 = \frac{207.05}{42} = 4.93\,; \quad a_0 = \frac{848.3}{8} = 106.04$$

$$\hat{x}_{ik} = 4.93 \cdot t_{ik} + 106.04\,; \quad \sum t_{ik} = 0$$

- Seasonality Components Calculation:

$$S_1 = \frac{-5.99 - 4.15 - 11.41 - 12.77}{4} = -8.58$$

$$S_2 = \frac{8.58 + 9.32 + 10.36 + 6.00}{4} = 8.57$$

- Adjustment:

$$\sum S_k = -8.58 + 8.57 = -0.01 \neq 0\,; \quad S_k = \frac{-0.01}{2} = -0.005$$

$${}_0S_1 = -8.58 - \left(-0.005\right) = -8.575\,; \quad {}_0S_2 = 8.57 - \left(-0.005\right) = 8.575$$

The sum of the adjusted components is 0 (for clarity, three decimal places have been left in the calculations).

Interpretation of the components of seasonality:

${}_0S_1 = -8.575$: between 2018 and 2021 in Poland, the number of flats completed in the first six months was on average 8.575 thousand (or 8,575 dwellings) below the trend value.

${}_0S_2 = 8.575$: between 2018 and 2021 in Poland, the number of flats completed in the second half of the year was on average 8.575 thousand (or 8,575 dwellings) higher than the trend value.

Example 4.8

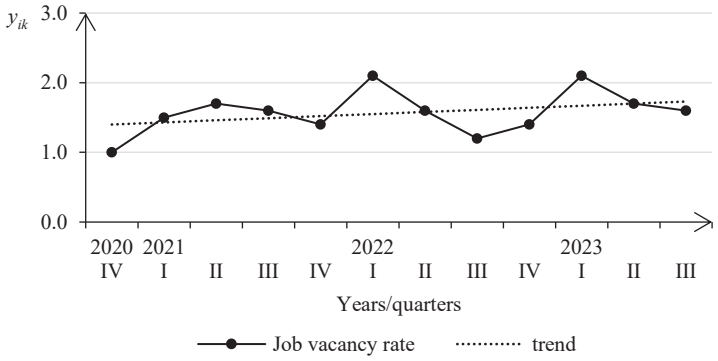
For the given time series data (Table 4.15), determine the appropriate measures of seasonality.

Table 4.15. Job vacancy rate in Croatia, 2020Q4–2023Q3 (quarterly data; %) and supporting calculations

Years/ quarters	Job vacancy rate (%) y_{ik}	t_{ik}	$y_{ik} \cdot t_{ik}$	t_{ik}^2	$\hat{x}_{ik}$
2020 IV	1.00	1	1.00	1	1.40
2021 I	1.50	2	3.00	4	1.43
II	1.70	3	5.10	9	1.46
III	1.60	4	6.40	16	1.49
IV	1.40	5	7.00	25	1.52
2022 I	2.10	6	12.60	36	1.55
II	1.60	7	11.20	49	1.58
III	1.20	8	9.60	64	1.61
IV	1.40	9	12.60	81	1.64
2023 I	2.10	10	21.00	100	1.67
II	1.70	11	18.70	121	1.70
III	1.60	12	19.20	144	1.73
Σ	18.90	78	127.40	650	–
Mean	1.58	6.50	10.62	54.17	–

Source: Eurostat (jvs_q_nace2) and auxiliary calculations, accessed 15/02/2024.

Figure 4.9. Empirical time series data (Table 4.15) and trend line



Source: tab 4.15 and own study.

From the data presentation in Figure 4.9, it is clear that the time series is characterised by fluctuations, but the deviations from the trend are not uniform. Thus, the seasonality indices W_k (Formula (4.42); for four quarters) need to be determined. Firstly, we need to establish the trend line. In this case, Formula (4.29a) is adopted (with the assumption of $t_{ik} = 1,2,\ ...,\ n \cdot m$). The calculations can be found in Table 4.15 and below:

$$a_1 = \frac{10.62 - 1.58 \cdot 6.50}{54.17 - 6.50^2} = 0.03\,; \quad a_0 = 1.58 - 0.03 \cdot 6.50 = 1.37$$

$$\hat{x}_{ik} = 0.03 \cdot t_{ik} + 1.37\,, \quad t_{ik} = 1, 2, ..., n \cdot m$$

- Seasonality Indices Calculation:

$$W_1 = \frac{1.5 + 2.1 + 2.1}{1.43 + 1.55 + 1.67} = 1.2258$$

$$W_2 = \frac{1.7 + 1.6 + 1.7}{1.46 + 1.58 + 1.70} = 1.0549$$

$$W_3 = \frac{1.6 + 1.2 + 1.6}{1.49 + 1.61 + 1.73} = 0.9110$$

$$W_4 = \frac{1.0 + 1.4 + 1.4}{1.40 + 1.52 + 1.64} = 0.8333$$

$$\sum W_k = 1.2258 + 1.0549 + 0.9110 + 0.8333 = 4.0250 \neq m$$

- Adjustment of Seasonality Indices:
  Since the sum of the components is not 4 (or m), they need to be adjusted:

$$k_w = \frac{4}{4.0250} = 0.9938$$

- Adjusted indices:

$$_0W_1 = 1.2258 \cdot 0.9938 = 1.2182$$

$$_0W_2 = 1.0549 \cdot 0.9938 = 1.0483$$

$$_0W_3 = 0.9110 \cdot 0.9938 = 0.9053$$

$$_0W_4 = 0.8333 \cdot 0.9938 = 0.8282$$

The sum of the adjusted indicators is 4.

Interpretation of seasonality indicators:

${}_0W_1 = 1.2182$: between 2020 and 2023 in Croatia, in the first quarters, the job vacancy rate was on average 21.82% higher than the trend value ((1.2182 – 1) · 100% = 21,82%),

${}_0W_2 = 1.0483$: between 2020 and 2023 in Croatia, in the second quarters job vacancy rate was on average 4.83% higher than the trend value ((1.0483 – 1) · 100% = 4,83%),

${}_0W_3 = 0.9053$: between 2020 and 2023 in Croatia, in the third quarters job vacancy rate was on average 9.47% lower than the trend value ((0.9053 – 1) · 100% = –9.47%),

${}_0W_4 = 0.8282$: between 2020 and 2023 in Croatia, in the fourth quarters job vacancy rate was on average 17.18% lower than the trend value ((0.8282 – 1) · 100% = –17.18%).

4.2.3. Forecasting

The decomposition of time series, i.e., the separation of trend and seasonality, makes it possible to predict the development of phenomena in the future, i.e., to produce forecasts (Luszniewicz, Słaby 2008). Socio-economic phenomena, apart from their development trend, may or may not be characterised by seasonality. Therefore, the prediction of the magnitude of the observed phenomena can be made in one of the following ways:

a) forecasting for data without seasonal variations – the forecast is based on a trend function, e.g., the linear function

$$\hat{y}_t = a_1 \cdot t + a_0 \qquad \text{(Formula 4.29 or 4.30)}$$

after substituting an appropriate value of the time variable t,

b) forecasting for data with seasonal variations – the forecast is based on a trend function (e.g., linear) and an appropriate measure of seasonality (component S_k or index W_k):

- for additive fluctuations:

$$\hat{y}_{ik} = a_1 \cdot t_{ik} + a_0 + S_k \qquad (4.45)$$

which can also be written as:

$$\hat{y}_{ik} = \hat{x}_{ik} + S_k \qquad (4.46)$$

where $\hat{x}_{ik}$ denotes the trend value,

- for multiplicative fluctuations:

$$\hat{y}_{ik} = (a_1 \cdot t_{ik} + a_0) \cdot W_k \tag{4.47}$$

which can also be written as:

$$\hat{y}_{ik} = \hat{x}_{ik} \cdot W_k \tag{4.48}$$

Example 4.9

For data on the amount of non-recycled municipal solid waste in France from 2016 to 2021 (kg/capita/day, Example 4.5), a linear trend function was determined for both ways of taking the time variable *t*:

$$\hat{y}_t = -0.01 \cdot t + 0.93; \quad t = 1, 2, ..., n$$

$$\hat{y}_t = -0.01 \cdot t + 0.89; \quad \sum t = 0$$

Determine a forecast of the observed phenomenon in 2022.

As the data is not characterised by seasonality (annual data), the forecast was produced based just on the trend function by substituting an appropriate value of the time variable *t*. This was calculated for both forms of the trend function:

$$\text{using } \hat{y}_t = -0.01 \cdot 7 + 0.93 = 0.86; \quad (t = 1, 2, ..., n)$$

$$\text{using } \hat{y}_t = -0.01 \cdot 3.5 + 0.89 = 0.86; \quad (\sum t = 0)$$

The predicted volume of non-recycled municipal solid waste in France in 2021 is 0.86 kg/capita/day.

Example 4.10

A linear trend function and seasonality components were determined for time-series data on the number of flats completed in Poland from 2018 to 2021 (thousands, Example 4.7):

$$\hat{x}_{ik} = 4.93 \cdot t_{ik} + 106.04; \quad \sum t_{ik} = 0; \quad {}_0S_1 = -8.575; \quad {}_0S_2 = 8.575$$

Calculate the projected magnitude of the phenomenon under study in the first and second half of 2022.

In this case, the forecast should be based both on the trend parameters and on the determined seasonality measures (Formula 4.45):

$$\hat{y}_{\text{I}2022} = 4.93 \cdot 4.5 + 106.04 - 8.575 = 119.65$$

$$\hat{y}_{\text{II}2022} = 4.93 \cdot 5.5 + 106.04 + 8.575 = 141.73$$

The projected number of flats completed in Poland in the first half of 2022 is 119.65 thousand, and 141.73 thousand in the second half.

Example 4.11

A trend function and seasonality indices were determined for the time series data on job vacancy rate in Croatia from 2020Q4 to 2023Q3 (Example 4.8):

$$\hat{x}_{ik} = 0.03 \cdot t_{ik} + 1.37\,, \quad t_{ik} = 1, 2, ..., n,\ n \cdot m$$

$${}_0W_1 = 1.2182\,, \quad {}_0W_2 = 1.0483\,, \quad {}_0W_3 = 0.9053\,, \quad {}_0W_4 = 0.8282$$

What is the forecast for the studied phenomenon in the second and third quarters of 2024?

The forecast is determined based on Formula (4.47):

$$\hat{y}_{\text{III}2024} = \left(0.03 \cdot 15 + 1.37\right) \cdot 1.0483 = 1.91$$

$$\hat{y}_{\text{III}2024} = \left(0.03 \cdot 16 + 1.37\right) \cdot 0.9053 = 1.67$$

The predicted job vacancy rate in Croatia in the second quarter of 2024 is 1.91%, while in the third quarter it is 1.67%.

Annex

Examples of official and non-official statistical data sources

According to the Statistics Explained,[1] an encyclopaedia on European Union statistics, **official statistics** are statistics produced within national statistical systems, i.e., by statistical organisations and units within a country that act on behalf of the national government (e.g., National Statistical Institutes – NSIs). Official statistics are collected within a legal framework and in accordance with basic principles ensuring minimum professional standards, such as independence and objectivity. In the case of the European Union, the legal framework is based on the European Regulation (EC) No. 223/2009 under the name the European Statistics Code of Practice. On the other hand, **non-official statistics** include statistics collected and published by other bodies such as research institutes or private bodies.

National Statistical Institutes produce and publish numerous thematic statistical yearbooks, working papers, reviews, reports, brochures, etc. One strength of the data collected from NSIs could be its geographical accuracy if searching, e.g., for local or microdata from a particular country. NSIs are responsible for **population censuses,** i.e., information on the size, location, and characteristics of a population. The data freely available, however, may have several limitations, for example, being aggregated. In some countries, individual-level data may require obtaining a license and signing an agreement on research conduct methods for handling the data with the NSI. Another special case is data containing sensitive or personal information, which implies that **the General Data Protection Regulation (GDPR) needs** to be considered. The list of national statistical institutes designated to develop, produce, and disseminate national statistics and their English language websites from all European Union countries is listed below.

For comparative analysis among different EU countries, sources other than NSIs are available. Established in 1953, **Eurostat**[2] is the statistical office of the European Union. It produces statistics on the European level in close partnership with NSIs and other

---

[1] https://ec.europa.eu/eurostat/statistics-explained (accessed 21/3/2024).

[2] https://ec.europa.eu/eurostat (accessed 21/3/2024).

national authorities in the EU Member States via a partnership called the **European Statistical System** (ESS). The ESS's objective is to **provide comparable statistics at the EU level and beyond, including** the European Free Trade Association (EFTA) countries. Eurostat is the most reputable and trustworthy source of EU statistics, widely used by different target groups, from the general public, researchers, and politicians, to the industry. In addition to open statistical databases, Eurostat offers numerous thematic publications in English and some in other languages. For example, the Eurostat Regional Yearbook is a yearly flagship publication containing general and regional statistics on different NUTS levels. The NUTS classification, widely used by Eurostat, is defined in *Regulation(EC) No. 1059/2003* as a regional classification introducing a hierarchy of regions and dividing each Member State into regions that are classified as NUTS levels 1, 2, and 3 (from larger to smaller areas). Other flagship publications are the Key Figures on EU series and the Monitoring Report on Sustainable Development. The aim of these flagship publications is to present an overview of Eurostat data on multiple topics of high relevance for society and policy-making. Eurostat's reports (e.g., Quality Report of the European Union Labour Force Survey) offer quality-related aspects of new or experimental data in a statistical area. Online articles present a detailed statistical view on topics via Statistics Explained entries aiming at easy understanding and clarity. As for more methodological publications, Eurostat offers manuals that provide methodologies and standards applied in the ESS. The source of Eurostat data comes from EU members' national accounts, which are pre-specified data sets according to a fixed transmission timetable data. National accounts present the structure of economies as they develop over time. National accounts are created in accordance with the **European system of national and regional accounts**. This system is fully consistent with worldwide guidelines on national accounting, ensuring common concepts, definitions, classifications, and accounting rules that guarantee a consistent, reliable, and comparable quantitative description of an economy. Demographic statistics, on the other hand, are compiled by NSIs from **administrative data sources (i.e.,** population registers, residence registers, social insurance registers, civil status registers, and records on births, deaths, marriages, and divorces). Additionally, population and housing censuses provide very detailed information from the enumeration of the entire population, which improves the quality of demographic data. The EU legislation on demographic statistics is **output oriented, which gives the Member States the possibility** to choose the most appropriate data sources that they have available nationally. The quality criteria for relevance, accuracy, timeliness, and punctuality, accessibility, and clarity, comparability, and coherence, however, must be met.

Global statistics are being published, e.g., by the **United Nations Statistics Division**,[3] notably in the form of the UN Statistical Yearbook since 1948. The UN Yearbook provides a wide range of internationally available statistics on social, economic, and environmental conditions and activities at the national, regional, and world levels. There are several databases freely available via the UN webpage. (For selected examples of the available datasets, see Table A2).

Table A1. List of the National Statistical Institutes of the European Union

Country code	National Statistical Institutes	English website (accessed 21/3/2024)
BE	Statistics Belgium	https://statbel.fgov.be/en
BG	National Statistical Institute	https://www.nsi.bg/en
CZ	Czech Statistical Office	https://www.czso.cz/csu/czso/home
DK	Statistics Denmark	https://www.dst.dk/en
DE	Federal Statistical Office of Germany	https://www.destatis.de/EN
EE	Statistics Estonia	https://www.stat.ee/en
IE	Central Statistics Office	https://www.cso.ie/en
EL	Hellenic Statistical Authority	https://www.statistics.gr/en
ES	National Statistics Institute	https://www.ine.es/en
FR	National Institute of Statistics and Economic Studies	https://www.insee.fr/en
HR	Croatian Bureau of Statistics	https://dzs.gov.hr/en
IT	Italian National Institute of Statistics	https://www.istat.it/en
CY	Statistical Service of Cyprus	https://www.cystat.gov.cy/en
LV	Central Statistical Bureau of Latvia	https://stat.gov.lv/en
LT	State Data Agency (Statistics Lithuania)	https://vda.lrv.lt/en
LU	National Institute of Statistics and Economic Studies	https://statistiques.public.lu/en/statistique-publique/statec.html
HU	Hungarian Central Statistical Office	https://www.ksh.hu/?lang=en
MT	National Statistics Office	https://nso.gov.mt
NL	Statistics Netherlands	https://www.cbs.nl/en-gb
AT	Statistics Austria	https://www.statistik.at/en
PL	Statistics Poland	https://stat.gov.pl/en
PT	Statistics Portugal	https://www.ine.pt
RO	National Institute of Statistics	https://insse.ro/cms/en
SI	Statistical Office of the Republic of Slovenia	https://www.stat.si/statweb/en
SK	Statistical Office of the Slovak Republic	https://slovak.statistics.sk
FI	Statistics Finland	https://www.stat.fi/index_en.html
SE	Statistics Sweden	https://www.scb.se/en

[3] https://unstats.un.org (accessed 21/3/2024).

Table A2. Examples of open access European and global databases (accessed 21/03/2024)

Eurostat database https://ec.europa.eu/eurostat/web/main/data/database	Detailed datasets on EU and EEA / EFTA countries are categorised as follows: General and regional statistics; Economy and finance; Population and social conditions; Industry, trade, and services; Agriculture, forestry, and fisheries; International trade; Transport; Environment and energy; Science, technology, and digital society. In addition to these, datasets on EU policies and cross-cutting issues are freely available
World Bank Open Data https://data.worldbank.org	At the World Bank, the Development Data Group manages statistical and data work. The majority of the data comes from the statistical systems of member countries. The World Bank supports the developing countries by improving the capacity, efficiency, and effectiveness of national statistical systems
UNdata http://data.un.org	UNdata is a free online database offering international statistics through a single-entry point. Users can search and download a variety of statistical information produced by the United Nations and linked international agencies. It contains over 60 million data points and covers themes such as agriculture, communication, education, energy, environment, finance, gender, health, the labour market, national accounts, population and migration, science and technology, tourism, transport, and trade, and many more
UNStats COVID-19 response https://covid-19-data.unstatshub.org	This online database offers data relevant to the COVID-19 response, readily available as geospatial data web services, suitable for the production of maps and other data visualizations and analyses, and easy to download in multiple formats
SDG Global Database https://unstats.un.org/sdgs/dataportal	SDG Global Database provides access to data on more than 210 Sustainable Development Goals (SDG) global indicators, organised by indicator, country, region, or time period

Literature

Bąk I., Markowicz I., Mojsiewicz M., Wawrzyniak K. (2021). *Formulas and Tables. Statistical and Econometric Methods*, CeDeWu, Warszawa.

Bąk I., Markowicz I., Mojsiewicz M., Wawrzyniak K. (2021). *Statystyka opisowa. Przykłady i zadania* (3rd ed.). CeDeWu, Warszawa.

Bieszk-Stolorz B., Markowicz I. (2019). *Analiza trwania w badaniach ekonomicznych. Modele nieparametryczne i semiparametryczne*, CeDeWu, Warszawa.

Catlett J. (1991). *On changing continuous attributes into ordered discrete attributes*. In: Y. Kodratoff (ed.), *Machine Learning – EWSL-91: European Working Session on Learning* (pp. 164–178), Springer-Verlag, Berlin-Heidelberg.

Chmielewski M.R., Grzymala-Busse J.W. (1996). *Global discretization of continuous attributes as pre-processing for machine learning*, International Journal of Approximate Reasoning, 15(4), 319–331. DOI: 10.1016/S0888-613X(96)00074-6.

Dehnel G. (2010). *Rozwój mikroprzedsiębiorczości w Polsce w świetle estymacji dla małych domen*, Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu, Poznań.

Domański Cz., Pekasiewicz D., Baszczyńska A., Witaszczyk A. (2014). *Testy statystyczne w procesie podejmowania decyzji*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Dougherty J., Kohavi R., Sahami M. (1995). *Supervised and unsupervised discretization of continuous features*. In: *Proceedings of the 12th International Conference on Machine Learning* (pp. 164–178), Morgan Kaufmann.

Dudek A. (2013). *Metody analizy danych symbolicznych w badaniach ekonomicznych*, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław.

Dudek H., Krawiec M., Landmesser J. (2011). *Podstawy analizy statystycznej w badaniach rynku*, Wydawnictwo SGGW, Warszawa.

Freedman D., Diaconis P. (1981). *On histogram as a density estimator: L_2 theory*, Probability Theory and Related Fields, 57, 453–476.

Gatnar E. (2000). *Problemy dyskretyzacji zmiennych w nieparametrycznej analizie dyskryminacyjnej*, Prace Naukowe AE we Wrocławiu, 7(874), 190–198.

Gatnar E., Walesiak M. (2004). *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław.

Gatnar E., Walesiak M. (2011). *Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R*, Wydawnictwo C.H. Beck, Warszawa.

Gdakowicz A., Hozer-Koćmiel M., Markowicz I. (2022). *Zastosowanie metod opisu statystycznego do badania zjawisk społeczno-ekonomicznych*. In: I. Markowicz (ed.), CeDeWu.

Gdakowicz A., Putek-Szeląg E. (2020). *The Demand and Supply Analysis and Comparison of Dwellings in Szczecin* (pp. 169–184). *Springer*, https://EconPapers.repec.org/RePEc:spr:eurchp:978-3-030-48531-3_12.

Gdakowicz A., Putek-Szeląg E. (2020). *The use of statistical methods for determining attribute weights and the influence of attributes on property value*, Real Estate Management and Valuation, 28(4), 33–47. DOI: https://doi.org/10.1515/remav-2020-0030.

Gini C. (1912). *Variabilitá e mutabilita*. Reprinted in: *Memorie di metodologia statistica* (eds. E. Pizetti, T. Salvemini) Rome: Libreria Eredi Virgilio Veschi, 1955.

Gołata E., Dehnel G. (2021). *Credibility of disability estimates from the 2011 population census in Poland*, Statistics in Transition New Series, 22(2), 41–65, DOI: 10.21307/stattrans-2021-016.

Heiler S., Michels P. (1994). *Deskriptive und Explorative Datenanalyse*, R. Oldenbourg Verlag München-Wien, Oldenbourg.

Hozer J. (1993). *Mikroekonometria. Analizy, diagnozy, prognozy*, PWE, Warszawa.

Hozer J. (2004). *Matematyczno-ekonomiczne modele funkcjonowania gospodarki*, Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, Szczecin.

Hozer J. (ed.) (1998). *Statystyka. Opis statystyczny*, Katedra Ekonometrii i Statystyki, Uniwersytet Szczeciński, Szczecin.

Hozer-Koćmiel M. (2013). *Time wealth and income wealth*, APE Actual Problems of Economics, 2, 164–171.

Hozer-Koćmiel M. (2020). *Nieodpłatna praca w gospodarstwie domowym. Studium empiryczne*, Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, Szczecin.

Ivanenko V.I., Labkovsky V.A. (2015). *On Regularities of Mass Phenomena*, Sankhyā: The Indian Journal of Statistics, Series A, 77(2), 237–248.

Jajuga K. (1993). *Statystyczna analiza wielowymiarowa*, PWN, Warszawa.

Kocimowski K., Kwiatek J. (1977). *Wykresy i mapy statystyczne*, GUS, Warszawa.

King A.P., Eckersley R.J. (2019). *Statistics for Biomedical Engineers and Scientists. How to Visualize and Analyze Data*. Academic Press.

Kończak G., Trzpiot G. (2018). *Statystyka opisowa i matematyczna z arkuszem kalkulacyjnym Excel* (2nd ed.). Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach, Katowice.

Krzysztofiak M., Luszniewicz A. (1981). *Statystyka*, PWE, Warszawa.

Kukuła K. (1998). *Elementy statystyki w zadaniach*, PWN, Warszawa.

Luszniewicz A., Słaby T. (2008). *Statystyka z pakietem komputerowym STATISTICA PL. Teoria i zastosowania*, Wydawnictwo C.H. Beck, Warszawa.

Markowicz I. (2012). *Statystyczna analiza żywotności firm*. Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, Szczecin.

Markowicz I., Baran P. (2021) *Mirror data asymmetry in international trade by commodity group: the case of intra-Community trade*, Oeconomia Copernicana, 12(4), 889–905. DOI: 10.24136/oc.2021.029.

Młodak A. (2006). *Analiza taksonomiczna w statystyce regionalnej*, Difin, Warszawa.

Nowak E. (ed.) (2001). *Metody statystyczne w analizie działalności przedsiębiorstwa*, PWE, Warszawa.

Ostasiewicz W. (2011). *Badania statystyczne*, Wolters Kluwer, Warszawa.

Panek T. (ed.) (2007). *Statystyka społeczna*, PWE, Warszawa.

Paradysz J. (2012). *Statystyka regionalna: stan, problemy i kierunki rozwoju*, Przegląd Statystyczny, 59(2), 191–204.

Pociecha J. (ed.) (2009). *Współczesne problemy statystyki, ekonometrii i matematyki stosowanej*, Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków.

Scott D.W. (1979). *On optimal and data-based histograms*, Biometrica, 66(3), 605–610.

Siegel S., Castellan Jr. N.J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGrawHill.

Sobczyk M. (2007). *Statystyka* (5th ed., revised). PWN, Warszawa.

Sokołowski A. (2004). *O niewłaściwym stosowaniu metod statystycznych*, StatSoft Polska.

Stanisz A. (2006). *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny*, T. 1, StatSoft Polska, Kraków.

Steven S.S. (1959). *Measurement, Psychophysics and Utility*. In: C.W. Churchman, P. Ratoosh (eds.), *Measurement: Definitions and Theories* (pp. 18–63). New York: Wiley.

Szreder M. (2009). *Statystyka w państwie demokratycznym*, Wiadomości Statystyczne, 54(6), 6–12. DOI: 10.59139/ws.2009.06.2.

Westfall P.H. (2014). *Kurtosis as Peakedness, 1905–2014. R.I.P.*, The American Statistician, 68(3), 191–195, DOI: 10.1080/00031305.2014.917055.

Wilke C.O. (2020). *Podstawy wizualizacji danych. Zasady tworzenia atrakcyjnych wykresów*, Helion SA, Gliwice.

Wilkinson L. (2005). *The Grammar of Graphics*, Springer-Verlag, New York.

Wilkinson L., Friendly M. (2009). *The History of the Cluster Heat Map*, The American Statistician, 63(2), 179–184, DOI: 10.1198/tas.2009.0033.

Wiśniewski J.W. (2013). *Correlation and regression of economic qualitative features*, LAP LAMBERT Academic Publishing, Saarbrücken.

Yule G.U., Kendall M.G. (1966). *Wstęp do teorii statystyki*, PWN, Warszawa.

Yang Y., Webb G.I., Wu X. (2010). *Discretization Methods*. In: O. Maimon, L. Rokach (eds.), *Data Mining and Knowledge Discovery Handbook* (2nd ed., pp. 101–116). Springer, New York Dordrecht Heidelberg London.

Zając K. (1994). *Zarys metod statystycznych*, PWE, Warszawa.

Zeliaś A. (2000). *Metody statystyczne*, PWE, Warszawa.

Other sources

Archive: Business economy – size class analysis, (n.d.). *Eurostat*. Retrieved February 19, 2024, from: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Archive:Business_economy_-_size_class_analysis&direction=next&oldid=75115.

CSO (n.d.). Local Characteristics Bank, Retrieved from: https://bdl.stat.gov.pl.

GUS (2020). *Activity of non-financial enterprises in 2019*. Warszawa.

Eurostat (n.d.). *Eurostat database*. Retrieved from: https://ec.europa.eu/eurostat/data/database.

FEPEX (n.d.). Retrieved February 12, 2024, from: https://www.fepex.es.

Few S. (2007). *Save the Pies for Dessert*. Retrieved June 26, 2021 from: http://www.perceptualedge.com/articles/08-21-07.pdf.

GUS (2014). *Graphical Presentation of Statistical Characteristics.* Warszawa. GUS, *Pojęcia stosowane w statystyce publicznej*. Retrieved April 7, 2024, from: https://stat.gov.pl/metainformacje/slownik-pojec/pojecia-stosowane-w-statystyce-publicznej/2915,pojecie.html?contrast=default.

Urząd Komunikacji Elektronicznej (2020). *Raport o stanie rynku telekomunikacyjnego w Polsce w 2019 r.* Warszawa.

GUS (2021). *Demographic Yearbook 2021.* Warszawa.

GUS (2020, 2021, 2023). *Statistical Yearbook of the Republic of Poland.* Warszawa.

Sustainable Development Report (2023). Retrieved February 10, 2024, from: https://dashboards.sdgindex.org, accessed 10/02/2024.

GUS (2014). *Warunki mieszkaniowe gospodarstw domowych i rodzin. Narodowy Spis Powszechny Ludności i Mieszkań 2011.* Warszawa.

World Inequality Report (2022). Retrieved February 19, 2022, from: https://wir2022.wid.world.

Abstract

Methods of statistical description. An analysis of socio-economic phenomena in European countries

In modern times, it is impossible to imagine any society, field of science, or economy without access to data on socio-economic phenomena. IT development facilitates both the collection of data and their availability. However, conclusions can only be drawn through the analysis of these data with the use of quantitative methods. Statistical methods are the backbone of data analysis. Statistics is recognised as a method of extracting information from data.

This book provides an overview of the methods applied in the search for statistical regularities in socio-economic phenomena in European countries. The authors point to methods for analysing one- and two-dimensional statistical data. They explain how survey methods should be chosen depending on the type of statistical series, characteristics and the purpose of the analysis. All descriptive parameters and functions are determined for the sample source data. Moreover, the manner of their interpretation is specified.

The book is the result of cooperation between authors representing the University of Szczecin (Institute of Economics and Finance; Poland) and the University of Helsinki (Ruralia Institute; Finland).

Keywords: statistical analysis, study of regularity of phenomena, European countries

Streszczenie

Metody opisu statystycznego. Analiza zjawisk społeczno-gospodarczych w krajach europejskich

Współcześnie nie można wyobrazić sobie społeczeństwa, nauki, gospodarki bez dostępu do danych o zjawiskach społeczno-ekonomicznych. Rozwój informatyczny sprzyja możliwości zarówno zbierania danych jak i ich udostępniania. Jednak wnioski wyciągnąć można analizując te dane przy wykorzystaniu metod ilościowych. Podstawą są metody statystyczne. Statystyka uznawana jest za metodę wydobywania informacji z danych.

Niniejsza książka stanowi przegląd metod stosowanych w poszukiwaniu prawidłowości statystycznych w zjawiskach społeczno-gospodarczych w krajach europejskich. Wskazano sposoby analizy jedno- i dwuwymiarowych danych statystycznych. Wyjaśniono jak należy wybierać metody badania w zależności od rodzaju szeregów statystycznych, cech i celu prowadzonej analizy. Wszystkie parametry opisowe i funkcje wyznaczono dla przykładowych danych źródłowych i wskazano sposób ich interpretacji.

Książka jest rezultatem współpracy autorek reprezentujących Uniwersytet Szczeciński (Instytut Ekonomii i Finansów) oraz University of Helsinki (Ruralia Institute).

Słowa kluczowe: analiza statystyczna, badanie prawidłowości zjawisk, kraje europejskie

In modern times, it is impossible to imagine any society, field of science, or economy without access to data on socio-economic phenomena. IT development facilitates both the collection of data and its availability. However, conclusions can only be drawn through analysing this data with the use of quantitative methods. The statistical methods are the backbone of data analysis. Statistics is recognised as a method of extracting information from data. The purpose of data analysis is to provide information that reduces risk when making economic decisions. Although we now have access to a variety of data (often in huge databases, the big data) and computer programmes that facilitate the investigation of these data, to draw the right conclusions we still need specific knowledge. When analysing diverse phenomena, it is necessary to choose the right research methodology. Statistical data analysis means searching for statistical regularities in the structure of phenomena, their correlation and changes over time. The layout of the book is consistent with the known types of statistical regularities and the research methodology used in their identification.

The main authors of the book are scientists working for the Department of Econometrics and Statistics of the Institute of Economics and Finance at the University of Szczecin. They have been researching socio-economic phenomena using quantitative methods for many years.

This book provides an overview of the methods applied in the search for statistical regularities in socio-economic phenomena in European countries. The authors point to methods for analysing one- and two-dimensional statistical data. They explain how survey methods should be chosen depending on the type of statistical series, characteristics and the purpose of the analysis. All descriptive parameters and functions are determined for the sample source data. Also, the manner of their interpretation is specified.

The book is the result of cooperation between authors representing the University of Szczecin (Institute of Economics and Finance; Poland) and the University of Helsinki (Ruralia Institute; Finland).

WYDAWNICTWO NAUKOWE US

71-101 Szczecin, ul. Mickiewicza 64
tel. 91 444 20 06, 91 444 20 09
e-mail: wydawnictwo@usz.edu.pl
www.wn.usz.edu.pl

ISBN 978-83-7972-916-6 (online)
ISBN 978-83-7972-915-9 (print)

9 788379 729159