

1 Introduction to Statistics

Statistics as a scientific discipline can be divided into two parts:

- *descriptive statistics*, which deals with methods of collecting, processing, presenting and analyzing data, together with their description using special statistical tools and methods,
- *inferential (inductive) statistics*, which is applied when collecting all the data is not possible for any reason (e.g. the population is very large). In that case we draw conclusions about the phenomenon under study on the basis of partial (selected) data. The collection (selection) of this data forms the subject of inductive statistics, whose foundation is probability theory. An essential component of inductive statistics is statistical analysis, whose task is to identify the regularities, relationships and developmental tendencies of mass phenomena. This includes estimation theory and the testing of statistical hypotheses.

Descriptive and inferential statistics are mutually interconnected. The subject of statistics as a scientific discipline is *mass random phenomena* (natural or social) which, under precisely defined material, temporal and spatial conditions, occur, or rather repeat, several times (e.g. births, deaths, the consumption or production of specific products). By examining a mass random phenomenon we can determine its nature, development, regularity, or mutual relationships.

The element on which we observe a specific manifestation of a mass phenomenon (e.g. a person, an object, a product) is called a *statistical unit*.

The set of all statistical units that share the required common properties is called a *statistical population*. The *size of a statistical population* is characterized as the number of statistical units in a given statistical population. A *statistical variable (characteristic)* represents a certain property of a mass phenomenon that is the subject of the statistical investigation (e.g. the age of respondents, the size of a product, the occupation of respondents).

We divide statistical variables into:

- *qualitative variables*, which we use when the properties of the statistical unit are expressed verbally (e.g. customer satisfaction, respondents' occupation),
- *quantitative variables*, which we use when the properties of the statistical unit are expressed numerically, as measurable units (e.g. the weight of a product, a customer's age).

We divide quantitative statistical variables into:

- *continuous*, which can take on any real values from some bounded or unbounded interval (e.g. the width of a product, a respondent's height),
- *discrete*, which can take on only certain specific values within some interval (e.g. the number of products).

We divide qualitative statistical variables into:

- *alternative (dichotomous)*, which can take on only two possible values (e.g. quality - non-quality),
- *polytomous (multi-category)*, which can take on more than two possible values (e.g. damaged clothing - slightly damaged clothing - undamaged clothing).

Statistical investigation can generally be divided into three stages:

1. *Statistical data collection*, whose task is to gather statistical data on the mass phenomena and processes under study.
2. *Statistical processing of the collected data*, whose task is to organize, sort and process the collected statistical data in a clear manner.
3. *Statistical evaluation*, whose task is to analyze and evaluate the obtained results and, on the basis of these results, to formulate conclusions.

2 Descriptive Statistics

At the beginning of a statistical investigation we generally obtain a large amount of data, which we record in the order in which it was collected. The first step in processing this data is sorting it, which achieves clarity of the statistical population. At the same time, certain characteristic features and regularities of the analyzed population may become apparent.

2.1 Sorting a Statistical Population

Sorting a statistical population is the arrangement of statistical units into groups, or classes, that are homogeneous with respect to a certain statistical variable. According to the number of sorting variables we distinguish between *single-stage* sorting (sorting according to a single sorting variable, e.g. a respondent's age, a product's height) and *multi-stage* sorting (sorting according to several sorting variables, e.g. a respondent's age and gender, a product's height and width). We will consider only single-stage sorting.

With regard to the number of different values, we divide the sorting of a statistical population into *simple* and *interval* sorting. In this textbook we will work mainly with quantitative statistical variables.

When sorting a statistical population, two rules must be observed:

- the *rule of completeness*, i. e. every statistical unit must be classified,
- the *rule of uniqueness*, i. e. every statistical unit must be classified into exactly one class.

2.1.1 Simple Sorting

Simple sorting is used when the statistical variable is discrete or continuous and at the same time takes on few different, often repeating, values. Let $x_1, x_2, \dots, x_n$ be the obtained values of the observed statistical variable, where the size of the statistical population is n . Let the total number of different values of the statistical variable be k . It follows that $k \leq n$. Each different value x_j of the statistical variable forms one class. We arrange the values $x_1, x_2, \dots, x_n$ into a non-decreasing sequence, thus obtaining the *order statistics (variation series)* in the form

$$x_{\min} = x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n-1)} \leq x_{(n)} = x_{\max}.$$

For each value x_j , $j = 1, 2, \dots, k$ of the statistical variable we determine the following frequencies:

- The *absolute frequency* n_j of the value x_j of the statistical variable represents the number of statistical units with the same value x_j .

Table 1: Frequency table (frequency distribution table)

j	x_j	n_j	f_j	N_j	F_j
1	x_1	n_1	f_1	n_1	f_1
2	x_2	n_2	f_2	$n_1 + n_2$	$f_1 + f_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	x_k	n_k	f_k	n	1

- The *relative frequency* f_j is the ratio of the absolute frequency n_j of the value x_j of the statistical variable to the size of the statistical population n . It holds that

$$f_j = \frac{n_j}{n}. \quad (1)$$

- The *cumulative absolute frequency* N_j of the value x_j of the statistical variable determines how many statistical units have a value less than or equal to the value x_j . It holds that

$$N_j = n_1 + n_2 + \dots + n_j = \sum_{i=1}^j n_i. \quad (2)$$

- The *cumulative relative frequency* F_j of the value x_j of the statistical variable is the ratio of the cumulative absolute frequency N_j of the value x_j of the statistical variable to the size of the statistical population n . It holds that

$$F_j = \frac{N_j}{n} = \frac{n_1 + n_2 + \dots + n_j}{n} = f_1 + f_2 + \dots + f_j. \quad (3)$$

Relative frequencies are often expressed as percentages. The value $100 f_j \%$ represents the percentage of elements from the statistical population that have the value x_j of the observed statistical variable. The value $100 F_j \%$ indicates the percentage of elements from the statistical population that have a value less than or equal to the value x_j of the observed statistical variable.

The following relations hold for the frequencies

$$\sum_{j=1}^k n_j = N_k = n, \quad \sum_{j=1}^k f_j = F_k = 1.$$

We record the frequencies in a *frequency table* (*frequency distribution table*) (Table 1).

2.1.2 Interval Sorting

Interval sorting is used when the statistical variable is continuous and at the same time takes on a large number of different values that generally repeat rarely or not at all. In the case

of interval sorting, the statistical units $x_1, x_2, \dots, x_n$ in the population are divided into k intervals (classes), where $k \leq n$, and n is the size of the statistical population.

Let us denote I_j as the j -th class interval, where t_{j-1} is its lower bound and t_j is the upper bound of the interval. When determining the intervals it is necessary to specify which of the bounds will belong to the interval, that is, whether we will consider the interval type $(t_{j-1}, t_j]$ or $[t_{j-1}, t_j)$.

Each class interval I_j will be represented by the so-called *class mark* z_j , which we determine as the midpoint of the given class interval, i.e. it holds that

$$z_j = \frac{t_{j-1} + t_j}{2}, \quad j = 1, 2, \dots, k, \quad (4)$$

where k is the number of class intervals.

When performing interval sorting of statistical data it is important to choose the correct number of class intervals (classes) and their width. The number of classes k must not be too small, so as not to lead to an overly simplified view of the properties of the population, which could obscure some of its characteristic features. On the other hand, it should not be too large either, because the processing could become confusing and, again, some regularities characteristic of the given population might be lost. There is no uniform opinion or general rule for determining the number of classes. However, there are several recommended procedures. We present a few of them:

- $k \approx 1 + 3,322 \log n$ (Sturges' rule),
- $k \approx \sqrt{n}$,
- $0,55n^{0,4} \leq k \leq 1,25n^{0,4}$, where n is the number of statistical units in the population.

The *width* h of a class interval can be approximately calculated using the following relation

$$h \doteq \frac{R_V}{k}, \quad (5)$$

where k is the number of class intervals and R_V is the *range*, for which

$$R_V = x_{(n)} - x_{(1)} = x_{\max} - x_{\min}. \quad (6)$$

At the conclusion of interval sorting we again calculate the individual frequencies for each class interval, similarly to the case of simple sorting. We record the results clearly in the form of a frequency table, where the value x_j is replaced by the value of the class mark z_j .

2.2 Graphical Representation

The *graphical representation* of a statistical population is a certain presentation of the processing of statistical data in the observed statistical population. It provides a faster and more illustrative idea of the tendencies, relationships and characteristic features of the analyzed data. Various types of graphs are used in statistics. We present the most commonly used ones: the scatter plot, the line graph, the bar chart and the pie chart.

A *scatter plot* displays the measured values by means of points in a rectangular coordinate system. The coordinates of these points are formed by the corresponding values of the statistical variable on the x -axis and the corresponding absolute, or relative, frequencies on the y -axis. In the case of interval sorting of frequencies, we use the value of the class mark z_j instead of x_j . A special case of the scatter plot is the *correlation plot*, which is used in regression analysis to display the dependence of two variables. In this case, the values of the first (independent) variable are plotted on the x -axis and the values of the second (dependent) variable on the y -axis.

Line graphs are scatter plots in which the individual points are connected by line segments. They are often used to display the course of a time series, but also to display absolute or relative frequencies.

The *frequency polygon* is a line graph of a frequency distribution (Fig. 1(a)).

The *bar chart* is very widespread due to its simplicity and clarity. The bars may be arranged either vertically or horizontally.

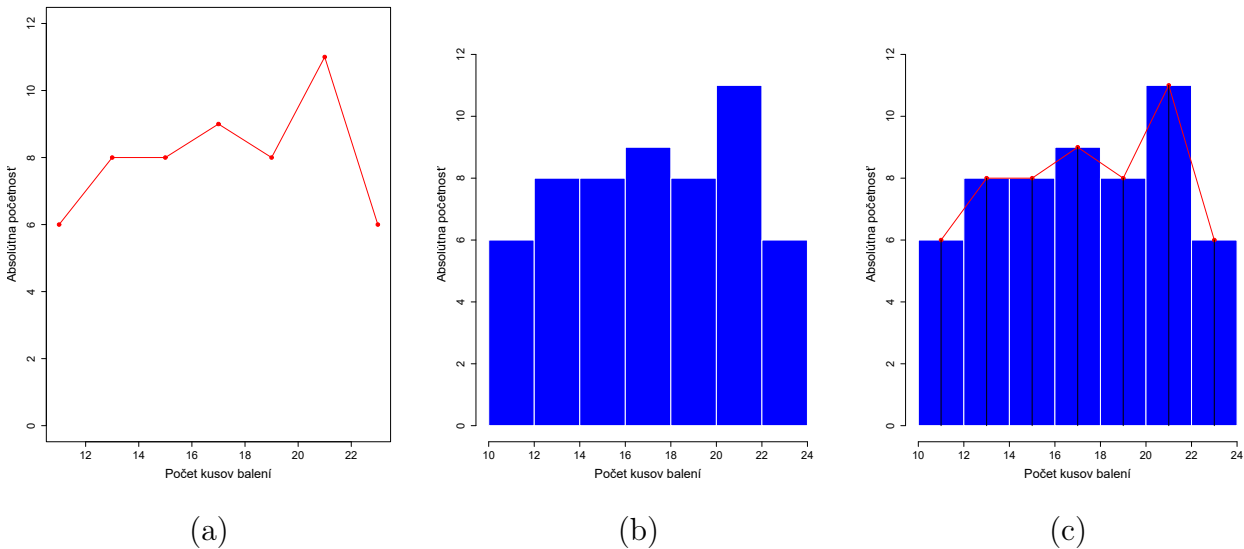


Figure 1: Frequency polygon and histogram

The *histogram* is a special bar chart, used primarily for the graphical representation of a dis-

tribution of absolute, or relative, frequencies (Fig. 1(b)). The bars of a histogram are always arranged vertically. The height of the bars corresponds to the value of the corresponding frequency, and their width corresponds to the width of the class interval. By connecting the midpoints of the top sides of all the bars, we obtain the frequency polygon (Fig. 1(c)). By analyzing a histogram we can determine many properties of the statistical population under study, such as the minimum and maximum value, and from these the variability of the distribution under study and the symmetry of the distribution.

The *pie chart* is a so-called area chart, where the area of the circle represents the entire population under study, divided by means of circular sectors into individual parts (Fig. 2).

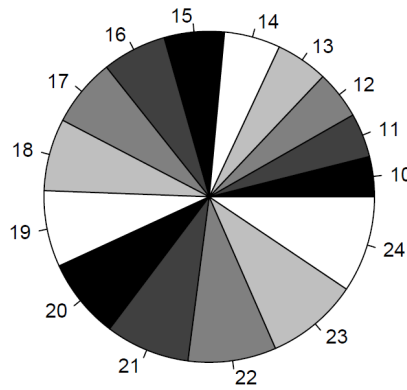


Figure 2: Pie chart

The *box plot* (*boxplot*) is a graph that provides, in a single figure, information about significant values (e.g. quartiles, the central value) as well as extreme values of the statistical population (Fig. 3). The boxplot consists of a basic rectangle (box) and two lines extending from the rectangle, which are terminated by whiskers. The position of the lower and upper sides of the rectangle corresponds to the first quartile $\tilde{x}_{25}$ and the third quartile $\tilde{x}_{75}$, and the position of the horizontal line inside this rectangle corresponds to the median $\tilde{x}_{50}$, i.e. the second quartile. The height of the rectangle is called the *quartile (interquartile) range* R_Q , for which

$$R_Q = \tilde{x}_{75} - \tilde{x}_{25}. \quad (7)$$

Inside the rectangle lie 50% of the measured values. In the simplest case, the whiskers at the ends of the lines extending from the rectangle correspond to the minimum and maximum values in the given population. In general, these lines may have a length equal to at most 1,5 times the height of the rectangle (or rather the quartile range). If some value exceeds this limit, it is considered an *outlier*. If it exceeds 3 times the height of the rectangle, it is considered an *extreme value*. These values are shown using special symbols (e.g. an asterisk, a cross). Thus, the boxplot can also serve as a graphical test for extreme values and outliers.

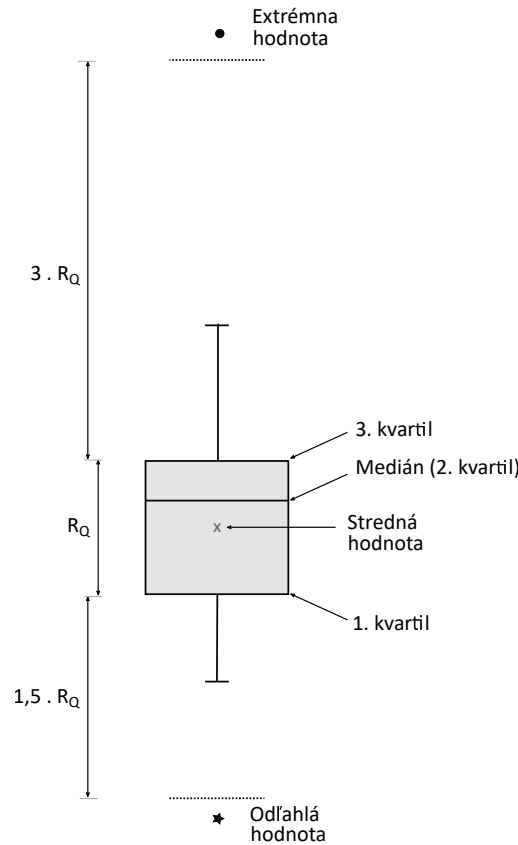


Figure 3: Boxplot

2.3 Numerical Characteristics

Statistical processing using tables and graphs facilitates data analysis and provides useful information about the structure of the statistical population under study, but this information is not always sufficient. Therefore, to describe a statistical population we introduce numerical characteristics. These include characteristics of location, variability, skewness and kurtosis.

2.3.1 Characteristics of Location

Characteristics of location characterize the location (level) of a variable, around which the other values are more or less concentrated. In a graphical representation of a frequency distribution, they indicate the location of the variable on the x -axis. Characteristics of location include central values and quantiles. Characteristics of location that are determined using all the measured values of the statistical variable are called *means* (e.g. arithmetic, geometric, harmonic). Characteristics of location that are determined using selected values of the statistical population include the mode and quantiles. Of the quantiles, the median is used most often.

The *arithmetic mean* $\bar{x}$ is defined as the sum of the measured values $x_1, x_2, \dots, x_n$ divided by

their number n , i. e.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i. \quad (8)$$

Formula (8) is used for an unsorted statistical population. If the values of the statistical variable are sorted into k classes, then the formula

$$\bar{x} = \frac{1}{n} \sum_{j=1}^k x_j n_j, \quad (9)$$

can be used to calculate the arithmetic mean, where n_j is the absolute frequency of the value x_j of the statistical variable. The arithmetic mean calculated using formula (9) is called the *weighted arithmetic mean*. In the case of interval sorting, we use the value of the class mark z_j instead of the values x_j in formula (9). In this case, however, it must be kept in mind that we obtain only an approximate value of the arithmetic mean. The arithmetic mean has a number of important properties. One of them is that the sum of the deviations of the individual values x_i of the statistical variable from the arithmetic mean $\bar{x}$ equals zero, i. e.

$$\sum_{i=1}^n (x_i - \bar{x}) = 0. \quad (10)$$

The arithmetic mean is sensitive to extreme values in the population and is meaningful only when the deviations of the measured values are random and there are no extreme values present in the population.

The *geometric mean* $\bar{x}_G$ is defined as the n -th root of the product of the measured values $x_1, x_2, \dots, x_n$, i. e.

$$\bar{x}_G = \sqrt[n]{\prod_{i=1}^n x_i} = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}. \quad (11)$$

Formula (11) is used for an unsorted statistical population. If the values of the statistical variable are sorted into k classes, then the formula

$$\bar{x}_G = \sqrt[n]{\prod_{j=1}^k x_j^{n_j}} = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot \dots \cdot x_k^{n_k}}, \quad (12)$$

can be used to calculate the geometric mean, where n_j is the absolute frequency of the value x_j of the statistical variable. The geometric mean calculated using formula (12) is called the *weighted geometric mean*. Similarly to the case of the arithmetic mean, if interval sorting is involved, we use the value of the class mark z_j instead of the values x_j in formula (12). The geometric mean is used, for example, to calculate average inflation, average interest, and the average percentage growth between two time periods.

The *harmonic mean* $\bar{x}_H$ is defined as the ratio of the number of measured values $x_1, x_2, \dots, x_n$ to the sum of their reciprocal values, i. e.

$$\bar{x}_H = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}. \quad (13)$$

Analogously to the preceding cases, the formula for the *weighted harmonic mean* has the form

$$\bar{x}_H = \frac{n}{\sum_{j=1}^k \frac{n_j}{x_j}}, \quad (14)$$

where n_j is the absolute frequency of the value x_j of the statistical variable. The harmonic mean is used mainly when calculating average speed and average time.

If the measured values $x_1, x_2, \dots, x_n$ are positive, the following relation holds between the aforementioned means

$$\bar{x}_H \leq \bar{x}_G \leq \bar{x}. \quad (15)$$

The *mode* $\hat{x}$ ($\text{mod}(x)$) is defined as the value of the statistical variable with the largest frequency (the value that occurs most often in the statistical population). A statistical population may have more than one mode. In the case of an interval frequency distribution, when determining the mode it is either sufficient to determine the modal (most frequent) interval, or we calculate the mode $\hat{x}$ within this interval using the relation

$$\hat{x} = a + h \frac{d_1}{d_1 + d_2}, \quad (16)$$

where a is the start of the modal interval, d_1 is the difference between the absolute frequency of the modal interval and the preceding interval, d_2 is the difference between the absolute frequency of the modal interval and the following interval, and h is the length of the class interval.

The *quantile* $\tilde{x}_p$ is a value that divides an ordered set of values of a certain quantitative statistical variable into two parts. One part contains $p\%$ of the values, which are less than or equal to the given quantile x_p (e.g. 5%, 15%, 50%, 95%). The other part, conversely, contains $(1 - p)\%$ of the values, which are greater than or equal to the quantile x_p (i. e. 95%, 85%, 50%, 5%). In the case of an interval frequency distribution of the statistical variable, we can obtain only approximate values of the quantiles. The most commonly used quantiles include the 50% quantile, quartiles, deciles and percentiles.

The *median* $\tilde{x}$ is the 50% quantile, that is, the value of the statistical variable from the statistical population arranged into order statistics that divides this series into two equally

numerous parts. It is, in fact, the middle value of the order statistics. If the size of the statistical population n is an odd number, then for the median

$$\tilde{x} = x_{(\frac{n+1}{2})}. \quad (17)$$

If the size of the statistical population n is an even number, then the median is determined using the relation

$$\tilde{x} = \frac{1}{2} \left(x_{(\frac{n}{2})} + x_{(\frac{n+2}{2})} \right). \quad (18)$$

In the case of an interval frequency distribution of the statistical variable, we calculate the median using the relation

$$\tilde{x} = a + h \frac{\frac{n+1}{2} - N_{j-1}}{n_j}, \quad (19)$$

where a is the start of the class interval containing the median (the median interval), h is the length of the class interval, n_j is the absolute frequency of the interval containing the median, n is the size of the statistical population, and N_{j-1} is the cumulative absolute frequency of the interval preceding the interval containing the median.

Quartiles are three values that divide a statistical population arranged into order statistics into four equally numerous parts, each of which contains 25% of the values. The *first (lower) quartile* $\tilde{x}_{25}$ separates the quarter (25%) of the smallest values of the statistical variable from the rest. The *second quartile* $\tilde{x}_{50}$ is, in fact, the median. The *third (upper) quartile* $\tilde{x}_{75}$ separates three quarters (75%) of the smallest values of the statistical variable from the rest. In the case of an interval frequency distribution of the statistical variable, we calculate the first quartile $\tilde{x}_{25}$ using the relation

$$\tilde{x}_{25} = a + h \frac{\frac{n}{4} - N_{j-1}}{n_j}, \quad (20)$$

where a is the start of the class interval containing the first quartile, h is the length of the class interval, n_j is the absolute frequency of the interval containing the first quartile, n is the size of the statistical population, and N_{j-1} is the cumulative absolute frequency of the interval preceding the interval containing the first quartile. For the third quartile $\tilde{x}_{75}$, in the case of an interval frequency distribution of the statistical variable, we use the relation

$$\tilde{x}_{75} = a + h \frac{\frac{3n}{4} - N_{j-1}}{n_j}, \quad (21)$$

where a is the start of the class interval containing the third quartile, h is the length of the class interval, n_j is the absolute frequency of the interval containing the third quartile, n is the size of the statistical population, and N_{j-1} is the cumulative absolute frequency of the

interval preceding the interval containing the third quartile.

Deciles $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_9$ divide a statistical population arranged into order statistics into ten equally numerous parts, each of which contains 10% of the values.

Percentiles $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_{99}$ divide a statistical population arranged into order statistics into a hundred equally numerous parts, each of which contains 1% of the values.

2.3.2 Characteristics of Variability

Characteristics of variability provide information about the homogeneity (uniformity) of a population, that is, whether the individual measured values in the analyzed population are scattered, or fluctuate. In the ideal case, when all the values in the population are the same (x_i is a constant), we speak of zero variability. The most commonly used (basic) characteristics of variability include the variance (dispersion), standard deviation, mean deviation, range, quartile range and coefficient of variation.

The *variance (dispersion)* s^2 is defined as the arithmetic mean of the squared deviations of the individual measured values $x_1, x_2, \dots, x_n$ from their arithmetic mean $\bar{x}$, i. e.

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2. \quad (22)$$

Formula (22) is used for an unsorted statistical population. If the values of the statistical variable are sorted into k classes, then the formula

$$s^2 = \frac{1}{n} \sum_{j=1}^k (x_j - \bar{x})^2 n_j, \quad (23)$$

can be used to calculate the variance, where n_j is the absolute frequency of the value x_j of the statistical variable. In the case of interval sorting, we use the value of the class mark z_j instead of the values x_j in formula (23).

The *standard deviation* s is the positive square root of the variance s^2 , i. e.

$$s = \sqrt{s^2}. \quad (24)$$

The *mean deviation* $\bar{d}$ is defined as the arithmetic mean of the absolute values of the deviations of all the measured values $x_1, x_2, \dots, x_n$ from the arithmetic mean, i. e.

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|. \quad (25)$$

Formula (25) is used for an unsorted statistical population. If the values of the statistical variable are sorted into k classes, then the formula

$$\bar{d} = \frac{1}{n} \sum_{j=1}^k |x_j - \bar{x}| n_j, \quad (26)$$

can be used to calculate the mean deviation, where n_j is the absolute frequency of the value x_j of the statistical variable. In the case of interval sorting, we use the value of the class mark z_j instead of the values x_j in formula (26).

The *range* R_V is defined as the difference between the largest and smallest value in a given set of measured values $x_1, x_2, \dots, x_n$ (6). Its disadvantage is that it depends on extreme values and also that it provides no information about the variability of the values lying between the largest and smallest value of the population.

The *quartile (interquartile) range* R_Q is defined as the difference between the third (upper) quartile $\tilde{x}_{75}$ and the first (lower) quartile $\tilde{x}_{25}$ (7).

The *coefficient of variation* V is defined as the ratio of the standard deviation s to the arithmetic mean $\bar{x}$, i. e.

$$V = \frac{s}{\bar{x}} 100\%. \quad (27)$$

The coefficient of variation V is one of the relative measures of variability. Unlike the preceding measures of variability, it allows the variability of a statistical variable to be compared across several populations and also across different units of measurement. The larger the value of the coefficient of variation, the larger the proportion the standard deviation represents of the arithmetic mean. This means that the variability of the measured values of the statistical variable is greater. If the coefficient of variation is greater than 40% – 50%, this indicates considerable inhomogeneity of the statistical population.

2.3.3 Characteristics of Skewness and Kurtosis

Characteristics of skewness are based on comparing the degree of concentration of small values of the observed statistical variable with the degree of concentration of large values of that variable.

Characteristics of kurtosis are based on comparing the degree of concentration of values around the central value with the degree of concentration of the other values.

Among the best-known characteristics of skewness is the coefficient of skewness (asymmetry), and among the best-known characteristics of kurtosis is the coefficient of kurtosis (excess). To define them we need to know the concept of the central moment of r -th order.

The *central moment of r -th order* μ_r of a statistical population is defined by the relation

$$\mu_r = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^r, \quad (28)$$

where $r \in \mathbb{N} \cup \{0\}$, $\bar{x}$ is the arithmetic mean of the measured values of the variable. Formula (28) is used for an unsorted statistical population. If the values of the statistical variable are sorted into k classes, we can use the formula

$$\mu_r = \frac{1}{n} \sum_{j=1}^k (x_j - \bar{x})^r n_j, \quad (29)$$

where n_j is the absolute frequency of the value x_j of the statistical variable. It holds that

$$\begin{aligned} \mu_0 &= 1, \\ \mu_1 &= 0, \\ \mu_2 &= s^2, \end{aligned} \quad (30)$$

where s is the standard deviation. Using the central moments of the third and fourth order we define the coefficient of skewness and the coefficient of kurtosis.

The *coefficient of skewness (asymmetry)* γ_3 is defined by the relation

$$\gamma_3 = \frac{\mu_3}{s^3} = \frac{1}{n s^3} \sum_{i=1}^n (x_i - \bar{x})^3, \quad (31)$$

where $\bar{x}$ is the arithmetic mean of the measured values and s is the standard deviation. Formula (31) is used for an unsorted statistical population. If the values of the statistical variable are sorted into k classes, then the formula

$$\gamma_3 = \frac{1}{n s^3} \sum_{j=1}^k (x_j - \bar{x})^3 n_j, \quad (32)$$

can be used to calculate the coefficient of skewness, where $\bar{x}$ is the arithmetic mean of the measured values, s is the standard deviation, and n_j is the absolute frequency of the value x_j of the statistical variable. In the case of interval sorting, we use the value of the class mark z_j instead of the values x_j in formula (32). The coefficient of skewness helps to decide on the agreement of the measured data with the normal distribution model in terms of symmetry about the arithmetic mean $\bar{x}$.

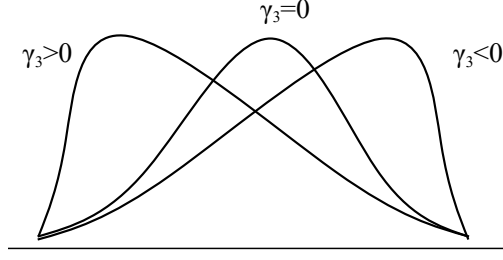


Figure 4: Coefficient of skewness

- If $\gamma_3 = 0$, then the measured values are symmetrically distributed about the arithmetic mean. For the central values, $\bar{x} = \tilde{x} = \hat{x}$ holds and the frequency distribution is symmetric (Fig. 4).
- If $\gamma_3 < 0$, then for the central values it usually holds that $\bar{x} < \tilde{x} < \hat{x}$ and the frequency distribution is *negatively skewed* (Fig. 4).
- If $\gamma_3 > 0$, then for the central values it usually holds that $\bar{x} > \tilde{x} > \hat{x}$ and the frequency distribution is *positively skewed* (Fig. 4).

The *coefficient of kurtosis (excess)* γ_4 is defined by the relation

$$\gamma_4 = \frac{\mu_4}{s^4} - 3 = \frac{1}{n s^4} \sum_{i=1}^n (x_i - \bar{x})^4 - 3, \quad (33)$$

where $\bar{x}$ is the arithmetic mean of the measured values and s is the standard deviation. Formula (33) is used for an unsorted statistical population. If the values of the statistical variable are sorted into k classes, then the formula

$$\gamma_4 = \frac{1}{n s^4} \sum_{j=1}^k (x_j - \bar{x})^4 n_j - 3, \quad (34)$$

can be used to calculate the coefficient of kurtosis, where $\bar{x}$ is the arithmetic mean of the measured values, s is the standard deviation, and n_j is the absolute frequency of the value x_j of the statistical variable. In the case of interval sorting, we use the value of the class mark z_j instead of the values x_j in formula (34). The coefficient of kurtosis expresses the steepness of the frequency distribution of the measured values of the statistical variable. That is, it indicates whether the concentration of the measured values of the statistical variable around the arithmetic mean is, in the frequency distribution, greater or smaller than in a population with a normal distribution.

- If $\gamma_4 = 0$, then the curve of the frequency distribution is exactly as peaked as the Gaussian curve of the normal distribution (Fig. 5).

looms recorded on the i -th working day, $i = 1, 2, \dots, 40$, and by x_j the different values of x_i , $j = 1, 2, \dots, 8$. Then n_j is the absolute frequency of the value x_j of the statistical variable, f_j is the relative frequency of the value x_j of the statistical variable, N_j is the cumulative absolute frequency of the value x_j of the statistical variable, and F_i is the cumulative relative frequency of the value x_j of the statistical variable. The frequency table is then given in Table 2.

Table 2: Frequency table for Example 2.1

i	x_j	n_j	N_j	f_j	F_j	f_j [in %]	F_j [in %]
1	15	1	1	0,025	0,025	2,5	2,5
2	16	1	2	0,025	0,050	2,5	5,0
3	17	9	11	0,225	0,275	22,5	27,5
4	18	13	24	0,325	0,600	32,5	60
5	19	11	35	0,275	0,875	27,5	87,5
6	20	2	37	0,050	0,925	5	92,5
7	21	2	39	0,050	0,975	5	97,5
8	22	1	40	0,025	1,00	2,5	100

b) The next task is to calculate all the numerical characteristics. The arithmetic mean $\bar{x}$ is calculated using relation (9)

$$\bar{x} = \frac{1}{40} (15 \cdot 1 + 16 \cdot 1 + 17 \cdot 9 + 18 \cdot 13 + 19 \cdot 11 + 20 \cdot 2 + 21 \cdot 2 + 22 \cdot 1) = 18,275.$$

This value gives the average number of broken threads per day. The size of the statistical population is 40, which is an even number.

The median is calculated using relation (18)

$$\tilde{x} = \frac{1}{2} \left(x_{(\frac{n}{2})} + x_{(\frac{n+2}{2})} \right) = \frac{1}{2} (x_{(20)} + x_{(21)}) = \frac{18 + 18}{2} = 18.$$

This is the value that divides the order statistics into two equally sized parts.

The mode $\hat{x}$ is defined as the value of the statistical variable with the largest frequency, and in this case $\hat{x} = 18$.

The first quartile $\tilde{x}_{25} = 17$. It is the arithmetic mean of $x_{(10)}$ and $x_{(11)}$ of the order statistics. It separates the quarter of the smallest values from the rest of the values.

The third quartile $\tilde{x}_{75} = 19$. It is the arithmetic mean of $x_{(30)}$ and $x_{(31)}$ of the order statistics. It separates three quarters of the smallest values from the rest of the values.

The variance s^2 is calculated using relation (23)

$$s^2 = \frac{1}{40} [1 \cdot (15 - 18,275)^2 + 1 \cdot (16 - 18,275)^2 + 9 \cdot (17 - 18,275)^2 + 13 \cdot (18 - 18,275)^2 + 11 \cdot (19 - 18,275)^2 + 2 \cdot (20 - 18,275)^2 + 2 \cdot (21 - 18,275)^2 + 1 \cdot (22 - 18,275)^2] = 1,846.$$

The standard deviation s is calculated using relation (24)

$$s = \sqrt{s^2} = \sqrt{1,846} = 1,358.$$

The mean deviation $\bar{d}$ is calculated using relation (26)

$$\bar{d} = \frac{1}{40} (1 \cdot |15 - 18,275| + 1 \cdot |16 - 18,275| + 9 \cdot |17 - 18,275| + 13 \cdot |18 - 18,275| + 11 \cdot |19 - 18,275| + 2 \cdot |20 - 18,275| + 2 \cdot |21 - 18,275| + 1 \cdot |22 - 18,275|) = 1,3.$$

The range R_V is calculated using relation (6)

$$R_V = x_{\max} - x_{\min} = 22 - 15 = 7.$$

The quartile range R_Q is calculated using relation (7)

$$R_Q = \tilde{x}_{75} - \tilde{x}_{25} = 19 - 17 = 2.$$

The coefficient of variation V is calculated using relation (27)

$$V = \frac{s}{\bar{x}} 100\% = \frac{1,358}{18,275} \cdot 100\% = 7,4\%.$$

The coefficient of skewness γ_3 is calculated using relation (32)

$$\begin{aligned} \gamma_3 = \frac{1}{40 \cdot (1,358)^3} [1 \cdot (15 - 18,275)^3 + 1 \cdot (16 - 18,275)^3 + 9 \cdot (17 - 18,275)^3 + \\ + 13 \cdot (18 - 18,275)^3 + 11 \cdot (19 - 18,275)^3 + 2 \cdot (20 - 18,275)^3 + \\ + 2 \cdot (21 - 18,275)^3 + 1 \cdot (22 - 18,275)^3] = 0,407. \end{aligned}$$

Since $\gamma_3 > 0$, the frequency distribution is positively skewed.

The coefficient of kurtosis γ_4 is calculated using relation (34)

$$\begin{aligned} \gamma_4 = \frac{1}{40 \cdot (1,358)^4} [1 \cdot (15 - 18,275)^4 + 1 \cdot (16 - 18,275)^4 + 9 \cdot (17 - 18,275)^4 + 13 \cdot (18 - 18,275)^4 + \\ + 11 \cdot (19 - 18,275)^4 + 2 \cdot (20 - 18,275)^4 + 2 \cdot (21 - 18,275)^4 + 1 \cdot (22 - 18,275)^4] - 3 = 0,591. \end{aligned}$$

Since $\gamma_4 > 0$, the curve of the frequency distribution is more peaked than the Gaussian curve of the normal distribution.

c) The individual graphs are shown in Fig. 6, Fig. 7 and Fig. 8.

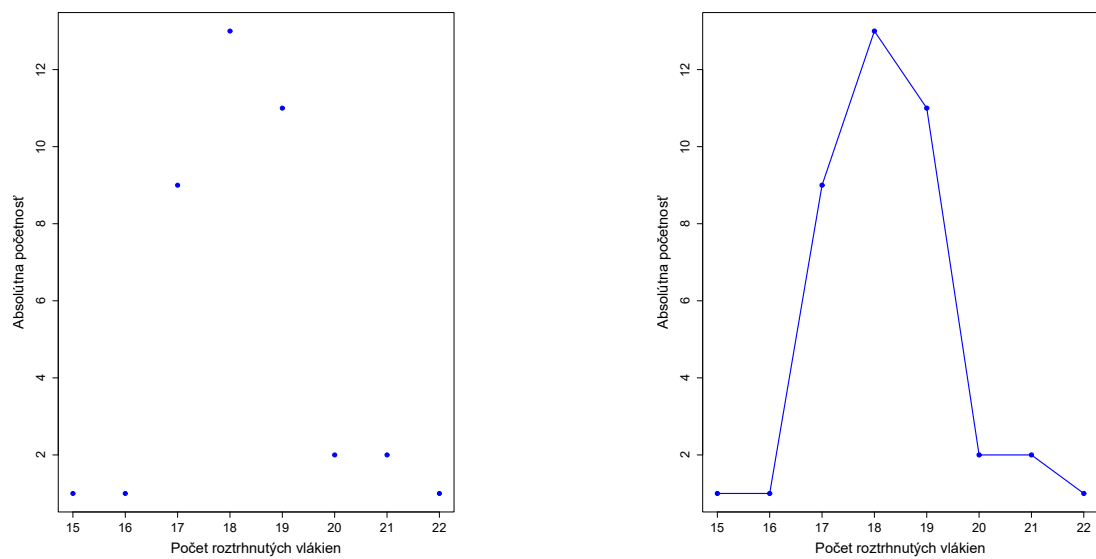


Figure 6: Scatter plot of absolute frequencies and polygon of absolute frequencies from Example 2.1

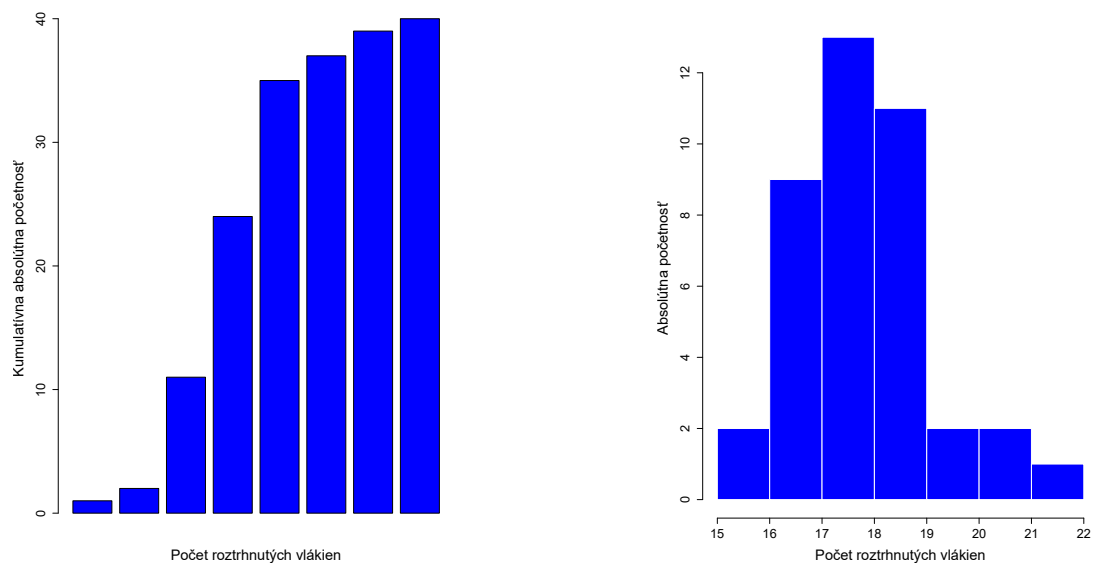


Figure 7: Bar chart of cumulative absolute frequencies and histogram of absolute frequencies from Example 2.1

Example 2.2 For 100 milk samples we recorded the percentage fat content in the milk, obtaining the following values [in %]: 4,25; 4,26; 4,25; 4,07; 4,20; 4,40; 4,05; 4,00; 4,17; 3,96; 4,30; 4,22; 4,25; 4,15; 4,25; 4,25; 3,97; 4,10; 4,47; 4,10; 4,17; 4,05; 4,25; 4,18; 4,05; 4,27; 4,03; 4,05; 4,15; 4,15; 4,40; 4,27; 4,25; 4,12; 4,30; 4,17; 4,20; 4,17; 4,22; 4,32; 4,20; 4,05;

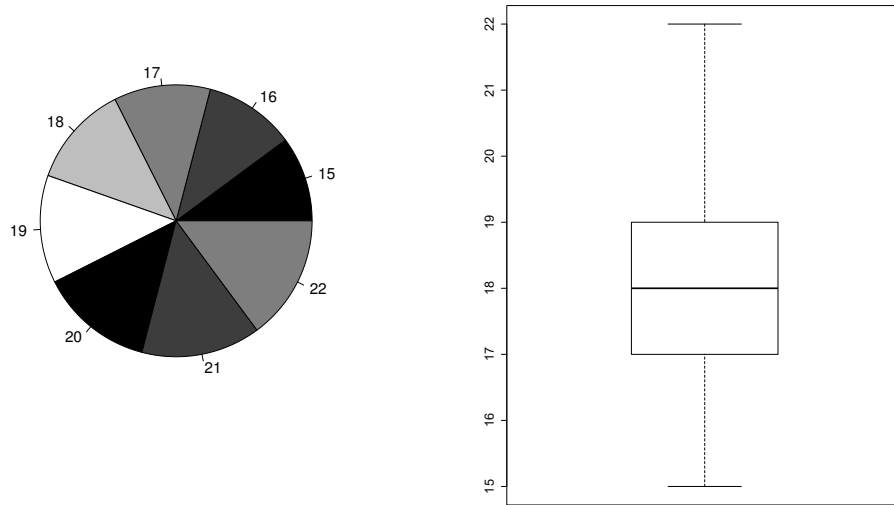


Figure 8: Pie chart and boxplot from Example 2.1

4,20; 4,37; 4,20; 4,25; 4,20; 4,15; 4,33; 4,10; 4,15; 4,25; 4,25; 4,05; 4,32; 4,27; 4,12; 4,27; 4,50; 4,38; 4,27; 4,17; 4,17; 4,12; 4,25; 4,12; 4,30; 4,33; 4,33; 4,21; 4,12; 4,15; 4,22; 4,20; 4,25; 4,50; 4,27; 4,10; 4,15; 4,31; 4,32; 4,10; 4,10; 4,17; 4,32; 4,02; 4,20; 4,17; 4,30; 4,44; 4,17; 4,20; 4,45; 4,07; 4,27; 4,15; 4,12; 4,42; 4,21; 4,35.

- Let us sort the statistical population into class intervals and construct the frequency table.
- Let us calculate all the numerical characteristics.
- Let us construct the polygon of absolute frequencies, the histogram of absolute frequencies and the boxplot.

Solution:

- We choose the number of class intervals k using the rule

$$k \approx \sqrt{n} = \sqrt{100} = 10.$$

The range R_V is calculated using relation (6)

$$R_V = x_{\max} - x_{\min} = 4,50 - 3,96 = 0,54.$$

The width h of the class interval is calculated using relation (5)

$$h = \frac{0,54}{10} = 0,054.$$

However, if we choose the number of class intervals to be 11, the width h of the class interval will be $h = \frac{0,54}{11} = 0,05$, and we will be working with "nicer" values. We choose the interval

type $(t_{j-1}; t_j)$. The value of the class mark z_j of the individual intervals is calculated using relation (4).

Similarly to the previous example, we enter the individual frequencies into the frequency table, calculated on the basis of relations (1), (2), (3). The frequency table is then given in Table 3.

Table 3: Frequency table for Example 2.2

j	$I_j = (t_{j-1}; t_j)$	z_j	n_j	N_j	f_j	F_j	f_j [in %]	F_j [in %]
1	(3, 95; 4, 00)	3,975	3	3	0,03	0,03	3	3
2	(4, 00; 4, 05)	4,025	8	11	0,08	0,11	8	11
3	(4, 05; 4, 10)	4,075	8	19	0,08	0,19	8	19
4	(4, 10; 4, 15)	4,125	14	33	0,14	0,33	14	33
5	(4, 15; 4, 20)	4,175	18	51	0,18	0,51	18	51
6	(4, 20; 4, 25)	4,225	18	69	0,18	0,69	18	69
7	(4, 25; 4, 30)	4,275	12	81	0,12	0,81	12	81
8	(4, 30; 4, 35)	4,325	9	90	0,09	0,90	9	90
9	(4, 35; 4, 40)	4,375	4	94	0,04	0,94	4	94
10	(4, 40; 4, 45)	4,425	3	97	0,03	0,97	3	97
11	(4, 45; 4, 50)	4,475	3	100	0,03	1,00	3	100

b) The next task is to calculate all the numerical characteristics. In the relations that we will use, we replace the value x_j with the value of the class mark z_j .

The arithmetic mean $\bar{x}$ is calculated using relation (9)

$$\bar{x} = \frac{1}{100} (3,975 \cdot 3 + 4,025 \cdot 8 + 4,075 \cdot 8 + 4,125 \cdot 14 + 4,175 \cdot 18 + 4,225 \cdot 18 + 4,275 \cdot 12 + 4,325 \cdot 9 + 4,375 \cdot 4 + 4,425 \cdot 3 + 4,475 \cdot 3) = 4,212.$$

This value gives the average fat content in the milk in a single sample.

The median $\tilde{x}$ is calculated using relation (19)

$$\tilde{x} = 4,15 + 0,05 \frac{\frac{100+1}{2} - 33}{18} = 4,2.$$

Since we have two most frequent intervals, I_5 and I_6 , we may have two modes. The mode is calculated using relation (16)

$$\hat{x}_1 = 4,15 + 0,05 \cdot \frac{4}{4+0} = 4,2,$$

$$\hat{x}_2 = 4,20 + 0,05 \cdot \frac{0}{0-6} = 4,2.$$

We find that in both cases the mode equals the value 4,2. However, this value does not belong to interval I_6 . Therefore only one mode remains, $\hat{x} = 4,2$.

The first quartile $\tilde{x}_{25}$ is calculated using relation (20)

$$\tilde{x}_{25} = 4,10 + 0,05 \cdot \frac{\frac{100}{4} - 19}{14} = 4,121.$$

It is the arithmetic mean of $x_{(25)}$ and $x_{(26)}$ of the order statistics.

The third quartile $\tilde{x}_{75}$ is calculated using relation (21)

$$\tilde{x}_{75} = 4,25 + 0,05 \cdot \frac{\frac{300}{4} - 69}{12} = 4,275.$$

It is the arithmetic mean of $x_{(75)}$ and $x_{(76)}$ of the order statistics.

The variance s^2 is calculated using relation (23)

$$\begin{aligned} s^2 = \frac{1}{100} [& 3 \cdot (3,975 - 4,212)^2 + 8 \cdot (4,025 - 4,212)^2 + 8 \cdot (4,075 - 4,212)^2 + 14 \cdot (4,125 - 4,212)^2 + \\ & + 18 \cdot (4,175 - 4,212)^2 + 18 \cdot (4,225 - 4,212)^2 + 12 \cdot (4,275 - 4,212)^2 + 9 \cdot (4,325 - 4,212)^2 + \\ & + 4 \cdot (4,375 - 4,212)^2 + 3 \cdot (4,425 - 4,212)^2 + 3 \cdot (4,475 - 4,212)^2] = 0,013. \end{aligned}$$

The standard deviation s is calculated using relation (24)

$$s = \sqrt{s^2} = \sqrt{0,013} = 0,114.$$

The mean deviation $\bar{d}$ is calculated using relation (26)

$$\begin{aligned} \bar{d} = \frac{1}{100} [& 3 \cdot |3,975 - 4,212| + 8 \cdot |4,025 - 4,212| + 8 \cdot |4,075 - 4,212| + 14 \cdot |4,125 - 4,212| + \\ & + 18 \cdot |4,175 - 4,212| + 18 \cdot |4,225 - 4,212| + 12 \cdot |4,275 - 4,212| + 9 \cdot |4,325 - 4,212| + \\ & + 4 \cdot |4,375 - 4,212| + 3 \cdot |4,425 - 4,212| + 3 \cdot |4,475 - 4,212|] = 0,091. \end{aligned}$$

The range R_V is calculated using relation (6)

$$R_V = x_{\max} - x_{\min} = 4,5 - 3,96 = 0,54.$$

The quartile range R_Q is calculated using relation (7)

$$R_Q = \tilde{x}_{75} - \tilde{x}_{25} = 4,275 - 4,121 = 0,154.$$

The coefficient of variation V is calculated using relation (27)

$$V = \frac{s}{\bar{x}} 100\% = \frac{0,114}{4,212} \cdot 100\% = 2,7\%.$$

The coefficient of skewness γ_3 is calculated using relation (32)

$$\gamma_3 = \frac{1}{100 \cdot (0,114)^3} [3 \cdot (3,975 - 4,212)^3 + 8 \cdot (4,025 - 4,212)^3 + 8 \cdot (4,075 - 4,212)^3 + \\ + 14 \cdot (4,125 - 4,212)^3 + 18 \cdot (4,175 - 4,212)^3 + 18 \cdot (4,225 - 4,212)^3 + 12 \cdot (4,275 - 4,212)^3 + \\ + 9 \cdot (4,325 - 4,212)^3 + 4 \cdot (4,375 - 4,212)^3 + 3 \cdot (4,425 - 4,212)^3 + 3 \cdot (4,475 - 4,212)^3] = 0,265.$$

Since $\gamma_3 > 0$, the frequency distribution is positively skewed.

The coefficient of kurtosis γ_4 is calculated using relation (34)

$$\gamma_4 = \frac{1}{100 \cdot (0,114)^4} [3 \cdot (3,975 - 4,212)^4 + 8 \cdot (4,025 - 4,212)^4 + 8 \cdot (4,075 - 4,212)^4 + \\ + 14 \cdot (4,125 - 4,212)^4 + 18 \cdot (4,175 - 4,212)^4 + 18 \cdot (4,225 - 4,212)^4 + 12 \cdot (4,275 - 4,212)^4 + \\ + 9 \cdot (4,325 - 4,212)^4 + 4 \cdot (4,375 - 4,212)^4 + 3 \cdot (4,425 - 4,212)^4 + 3 \cdot (4,475 - 4,212)^4] - 3 = -0,134.$$

Since $\gamma_4 < 0$, the curve of the frequency distribution is flatter than the Gaussian curve of the normal distribution.

c) The individual graphs are shown in Fig. 9 and Fig. 10.

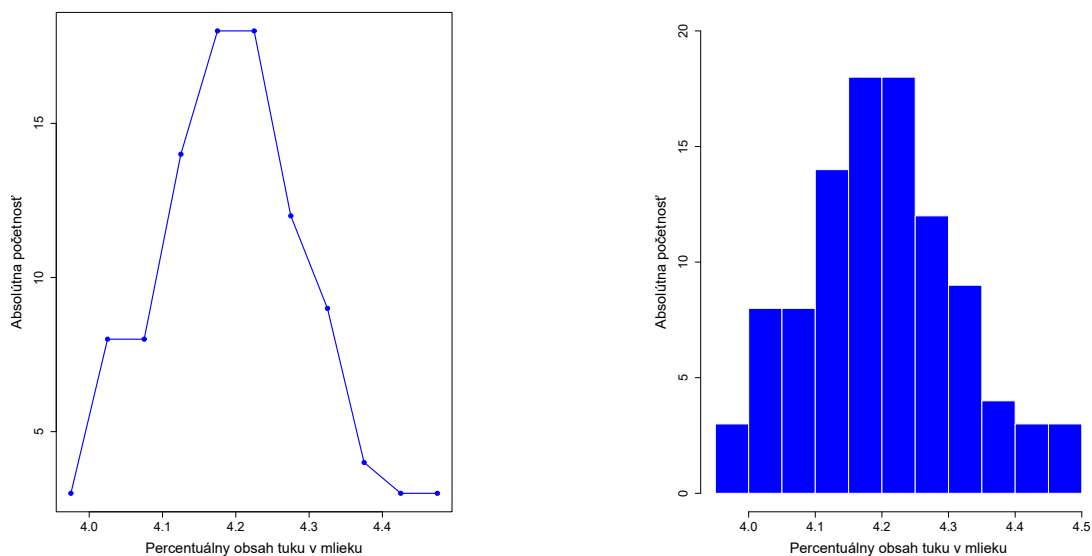


Figure 9: Polygon of absolute frequencies and histogram of absolute frequencies from Example 2.2

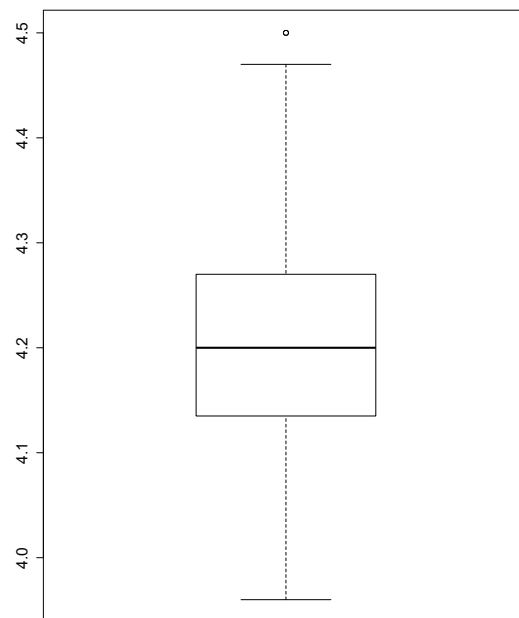


Figure 10: Boxplot from Example 2.2