Basics of Biostatistics

topics

1. descriptive statistics (what data are)
2. hypothesis testing (comparing data)
3. correlation and regression analysis

Recommended readings:

- **Medical Biophysics Practices** – ed. M. Kellermayer, 3rd ed., Semmelweis Publisher: **Appendix: Biostatistics**

Further readings:

- Harvey Motulsky: Intuitive Biostatistics – A Nonmathematical Guide to Statistical Thinking, Oxford University Press

- Nature Collection: Statistics for Biologists

INTRODUCTION

There is a story about a man, who had whiskey and soda on Monday, gin and soda on Tuesday, and rum and soda on Wednesday. Because the result was always the same, he drew the conclusion, that soda made him drunk.

Maybe not in terms of whiskey and soda, but our behavior resembles that of the man in the above story. Many of us **draw inappropriate conclusions easily** and make decisions based on them. One tends to generalize from single cases and select subjectively from the available information for the purpose of self-justification. Standpoints and opinions made this way are often very hard to change.

In first approximation, **statistics** is the field of science that helps to combat this general "disease". **It** helps to arrange one's thoughts critically and **keeps skepticism alive**, which is the base of every intellectual activity.

Because statistical reports are part of the everyday life, and it may seem that everybody knows the methods used to get them as well. This is partially true, **because anyone, who went to school** in Hungary in the last twenty years, **definitely used statistical procedures** quite a few times. When a student **calculates the average of his/her grades** from different subjects in order to know what he/she could expect in the school record at the end of the year, the student, though unaware of it, applies a typical case of making a statistical **estimation**.

The word "statistics" has several different meanings. The meaning used here stems from the Latin word "status". Its original meaning is condition, status, the state of things. Collected **data** make it possible to know and describe the status. Data are individual facts, **qualitative or quantitative** properties. Typical everyday examples of data are, for example, personal data such as name, birthplace, date of birth; names and prices of the products sold in a shop; regarding medical conditions they may be the paleness of the face, blood pressure, or a result of any laboratory diagnostic test.

Usually, data collection has a purpose. One asks a telephone number of someone to be able to call him or her later. The attitude of just collecting data in the hope that it may be useful for something, and trying to find the aim later is usually not appropriate (it is the characteristic of secret agencies only). Collected but unsorted data are usually entirely useless. What could we do with telephone numbers listed in the order of theirs arrival to the switchboard? Often data are sorted according to their importance. The doctor applies this when describing the condition (status) of the patient. Thus, **data must be collected, processed, conclusions need to be drawn, and most of the time decisions should be made**. **Statistics** is a field of science that is able to do all these. The "**bio**" prefix indicates that here the methods of statistics are used to analyze **phenomena of the living world**. Methods of **medical statistics** are even more specialized for problems occurring in medicine.

It is not easy to convince freshmen in medicine that **statistics is very important**. Furthermore, it is **inevitable** for them. Some examples are listed below.

Students have to carry out different measurements during most practices, some theoretical classes and later **during their advanced studies** as well. **Reliable conclusions can be drawn from the measured data** only by statistical methods.

Case-history sheets and laboratory files contain a large number of data. It is extremely important that physicians, dentists and pharmacists are able to **use statistical methods to evaluate data properly, draw conclusions, and judge the reliability of the results**. Statistics can save us from the deceptions present in the overwhelming advertisements of new drugs and procedures.

Understanding medical literature is sometimes difficult, as it often contains statistics. As an example, let us quote from a medical paper: "In the first group of patients the average preoperative refractive disturbance (ametropia) of -3.94 ± 1.3 diopters decreased during a one year follow-up period to -0.47 ± 0.54 value" and later: "Statistical results were analyzed using a two sample *t*-test and regression analysis." The question is, what do those numbers mean and what are these methods?

Last but not least, **statistics provide a unique view**, a way of thinking, which is **very similar to the way of thinking of physicians**. Let us illustrate this with an example. Suppose that your patient complains about a headache, and you want to find a reason. Based on your medical studies you will recall most of the possible reasons of a headache, such as:

1. High blood pressure.
2. Improper eyeglasses.
3. Inner pressure of the eyes is too high.
4. A tumor in the head.
5. Calcification of cervical vertebra.
6. Sensitivity to weather changes.
7. An oxygen-deficient working environment.

Because there could be several true answers, all the possibilities must be checked one by one, and decide whether the suspicion was well formulated. This procedure is principally the same as the hypothesis testing method that will be discussed later.
In the first approximation, there are four types of statistical procedure: **data collection, organization of data, analysis of data and the drawing of conclusions.**

DATA COLLECTION AND MAIN TYPES OF DATA

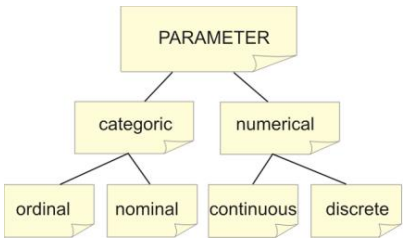


Table 1. *Classification of data.*

As mentioned earlier, data collection is motivated by a goal. There are data that are used only to identify and distinguish certain things. Larger number of data is collected in a hope that following data analysis a previously formulated question can be answered. **The way of collecting data or accessing data is called "experiment"** in general. Some data are already known, we just need to ask someone about it. Other data need to be measured somehow. In this regard the investigation of a natural phenomenon or casting a dice are both experiments. The data (the result of the experiment) can be of several different types. **Qualitative data** can be sorted into categories (**categorical data**). **Quantitative data** are characterized by a number (**numerical data**). Qualitative data are, for example, the names of the diseases, types of pathogens, or the severity of the condition. The size of the rash or the duration of the sickness can be expressed by a number (and a unit), therefore these are numerical (quantitative) data. There are two kinds of qualitative (categorical) data depending whether they can be sorted naturally in some order. An example of **ordinal** (sortable) data is, for example, the severity of the disease: modest, medium, strong. **Nominal** (not sortable) data are for example the blood groups: A, B, AB, 0. There are two sub-groups within the numerical data as well: continuous and discrete. If the result of the measurement can have any value within a certain interval, it is called **continuous** (e.g., weight, height, blood pressure). Other data can only have **discrete** values (e.g., number of children in the family). The above types of data are summarized in table 1. We have to mention that continuous data are only theoretically continuous. In practice we always work with discrete numbers (otherwise one would need to use an infinite number of decimal digits).

absolute frequency
absolute HäuFig.keit
abszolút gyakoriság

relative frequency
relative HäuFig.keit
relatív gyakoriság

ORGANIZING DATA, FREQUENCY AND GRAPHIC REPRESENTATION

In everyday life we often deal with a large number of data that are connected to a given problem. We need to organize and summarize our observations so that **we obtain an overview of the data**.

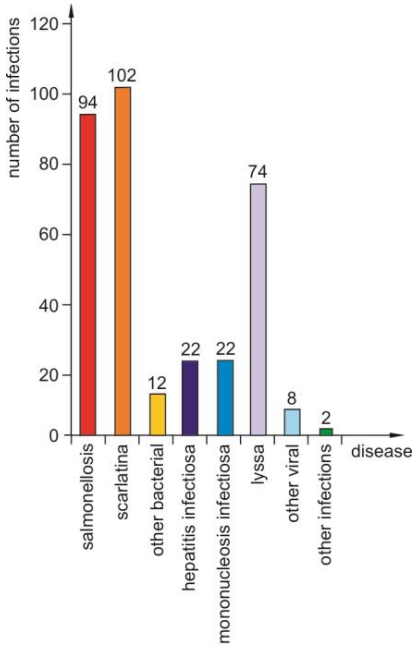


Fig. 1a. *Bar graph. Absolute frequencies of infections as a function of categories.*

INFECTION	DISEASE	Absolute frequency		Relative frequency	
bacterial	Salmonellosis (Food poisoning by Salmonella)	94	208	0.280	0.619
	Scarlatina (Scarlet fever)	102		0.304	
	Other bacterial	12		0.036	
viral	Hepatitis infectiosa (Hepatitis)	22	126	0.065	0.375
	Mononucleosis infectiosa (Mono)	22		0.065	
	Lyssa (Rabies)	74		0.220	
	Other viral	8		0.0238	
other	Other infections	2	2	0.006	0.006
total:		336	336	1.000	1.000

Table 2*. A summary table of infections.*

Table 2 summarizes the infections reported to occur in Budapest in October, 2000. Numbers in the first column of the table (94, 102, and so on) are occurrences of individual infectious diseases (Salmonellosis, Scarlatina, etc.) during the given time period. These numbers are called **absolute frequencies**. In the next column, subtotals (208, 126, 2) of the first column are calculated that correspond to larger groups of bacterial, viral or other types of infections. These are also absolute frequencies.

From the absolute frequencies the **relative frequencies can be calculated**. The relative frequency equals the absolute frequency of the category divided by the total number of cases (336). Relative frequencies are always numbers between 0 and 1 and are listed in the next column of the table. If percentage is preferred, multiply these relative frequencies by 100. Thus, relative frequency is a ratio, as

we can see from the definition. To be appropriate, **when speaking about relative frequency, both the category and what we compare it to must be specified.**
In our example, if the question is how frequent is salmonella infection among the bacterial infections, we have to divide the number of salmonella cases (94) by the total number of bacterial infections (208). The result (0.452) is a relative frequency, but here the comparison was made with the bacterial infections (208) rather than the total number of the infectious cases (336).

There are many ways to represent the absolute and relative frequencies graphically. There are two of these illustrated in Figs. 1a. and b.

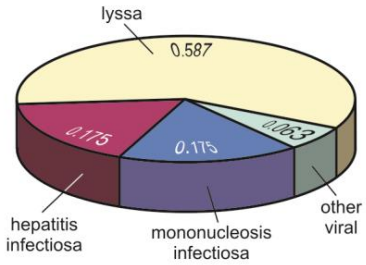


Fig. 1b. *Pie chart. Relative frequencies of the viral infections (categories).*

Check your knowledge:
Is there a contradiction in the following statement?
"In my group the relative frequency of girls is smaller than in the group of my friend, but we still have more girls in our group."

GRAPHICAL REPRESENTATION OF THE RELATIONSHIP BETWEEN DATA

It occurs many times in our experiments that **several characteristics** are determined simultaneously, or **one attribute is measured as a function of a fixed parameter** (e.g., as a function of time, like the regular measurement of body temperature in the hospital). In such cases we are interested in the relationship, the connection between the two sets of data. To get a better overview of data it is practical to **plot them in a graph**. Here the word **connection** is used in a very general sense, and **it does not mean causality**.

When plotting data we have to make two important decisions: the **scaling of the axes** and the choice of the **staring point** (origin). It is not necessary to represent the origin (zero values of both axes) on the graph. For example, if the index of refraction of protein solutions is measured, we know that we cannot get lower value than the index of refraction of distilled water (1.333).

The rule of thumb is that the graph (data points) should fill the available area of display as much as possible (see Fig. 2). It is useless, however, to increase the scale so much that the analysis of the data would give greater accuracy than the measurement itself. Such an apparent increase of accuracy may be confusing.

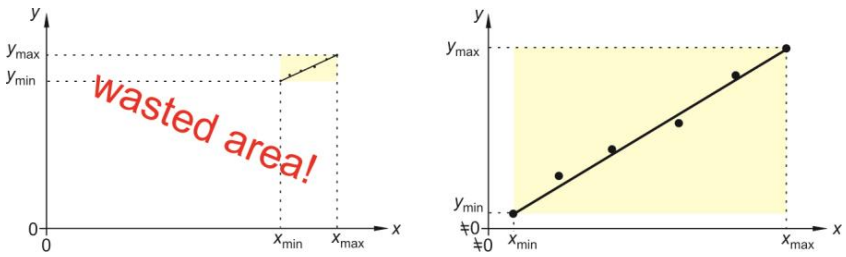


Fig. 2. *Improper and proper arrangement of the graph.*

Measured data always have errors. Hence, when drawing a line through the measured data points, draw a smooth line across the data rather than connecting the scattered points. Try to display equal number of points above and under the curve in such a way that their distances from the curve roughly identical. This way in most cases (except of some rare examples), you will get a smooth, continuous curve fitted to the measured data (see linear regression).

STATISTICAL INFERENCE AND PROBABILITY CALCULUS

The final goal of the statistical methods is to draw conclusions. The scheme of statistical inference is very similar to logical induction. In logics, the syllogism is a form of induction, where certain statements **necessarily indicate** further statements. (Classical example: every man is mortal. John is a man. Therefore, John is mortal.)

Statistical inference is not entirely the same as logical inference, however. Logical **inference gives a statement that is 100 % sure**, whereas **statistical inference yields a statement of given probability** (always less than 100 %). We may be **mistaken** in case of statistical inference. For example, if we state something with

95 % probability it means that in 5 cases out of 100 we were wrong, the inaccuracy is 5 %. The accuracy of our statements can be expressed in numbers.

By reformulating the statement we can decrease the inaccuracy at will, but this may yield more meaningless statements. Let us have an example. The police report after a bank robbery says: "… *eye witnesses saw the suspects getting in a car; it was concluded that the suspects left the scene with their own car or by a taxi, or they may have used a stolen or a rented car."* **If we decrease the inaccuracy of our inference, it has a price by which the usefulness of the conclusion is reduced; it is less valuable.** Our inductions are governed by these two opposite tendencies.

The reason of inaccuracy (in contrast to logic) is that in case of statistical inference **we are not able to take all of the circumstances into account.** A coin tossed is not governed by the mere chance when heads or tails fall. The situation is that we do not know all the necessary data with the right accuracy that will determine unambiguously the final position of the coin, that is, whether the result will be heads or tails. Because we cannot take every circumstance into account, we cannot give a definite answer; hence we say that **this event is random, where the word "random" expresses just the lack of our knowledge**.

In association with gambling it was observed a long time ago that even the **random, mass events follow some rules**. If the experiment of tossing coins is repeated many times, the result will be heads in the half of the cases and tails in the other half. We cannot prove this, but based on the experience, we can say that in case of a large number of (independent) experiments the **relative frequency** of the heads [(number of heads)/(number of heads + tails)], and that of the tails [(number of tails)/(number of heads + tails)] **show stability** (law of large numbers), and both will be around ½. Based on this, we can say that the **probability** (chance) of getting heads in one toss is exactly ½. The probability of getting tails is obviously the same.

Probability calculus gives a mathematical description of laws **of mass events** in the material world **that are not determined unambiguously by the circumstances**. Our experiments, observations and data fall into this category. Consequently, statistics are based on the principles of probability calculus.

POPULATION, VARIABLE, SAMPLE

Everyone makes measurements, experiments or observations. Some examples are listed in the Table 3.

WHO MEASURES WHAT?		
PHYSICIST	PHYSICIAN	STUDENT DURING THE PRACTICE OF MEDICAL PHYSICS (topic and number of practice)
length	body height	red blood cell diameter (3.)
frequency	pulse rate	pulse rate (9., 20.)
temperature	body temperature	–
concentration	blood glucose level	blood plasma protein concentration (5.)
voltage	ECG-signal	ECG-signal (24.)
power density	hearing threshold	hearing threshold (22.)
pressure	blood pressure	blood pressure ( )
impedance	skin impedance (resistance)	skin impedance (21.)

Table 3. *What does a physicist, a physician and a medical student measure?*

Let us emphasize again that the goal of our measurements is to understand something or to answer a question.

Let us choose, as an example, the measurement of pulse rate that can be easily done during the practice of medical biophysics as well. The pulse rate, the frequency of heartbeat, is technically a **continuous parameter**; we use discrete values (integers) just for simplicity, and because we are used to that. The unit of

the pulse rate is 1/minute. We will work only with the number values in the following, but note that all the final results should have units, too. There are many questions which could be answered by this measurement.

Such questions are:

1. *WHAT IS THE VALUE* of pulse rate of medical student Doris Diligent?
2. *WHAT IS THE* normal *VALUE* of the pulse rate?
3. *DOES* the pulse rate *CHANGE* after holding breath for one minute?
4. *IS THERE A DIFFERENCE* between the pulse rate of girls and boys?
5. Etc., etc.

Let us examine the questions one by one. **Without knowing any statistics** one would think that answering the first question is very easy. We just need to measure the pulse number of Doris, and we will **have a result and that's it**. However, if one has heard about statistics already, then **skepticism wakes up** and instead of a definite answer even further questions will be asked. Such questions are: is this the right answer for sure? Did I make any mistake during the measurement? If the investigator knows that "chance" factors also influence the experiment, then an "accurate" measurement cannot be performed no matter how hard it is attempted. Hence a decision is made to **perform the measurement again and again**.

When performing a measurement several times, it is always assumed that **the same thing is measured** again, and the **same result is expected**. In other words, the pulse rate of Doris **is expected not change in the long run in any direction**, **and the results might differ** only due to random variations. We can say that the result have two parts: the main one is the deterministic (constant) and the subsidiary one is the stochastic (random). Naturally, these two parts can not be separated directly.

Multiple measurements can be regarded as if there was a set of all possible observations, called the **population** (fundamental ensemble), and during every measurement we choose an element of this set. In our example the set has an infinite number of elements, but this is not a necessary condition. An element of this set in general is called the **variable**, and x is its usual symbol. The variable may attain different values. The value of the variable is given by the particular measurement.

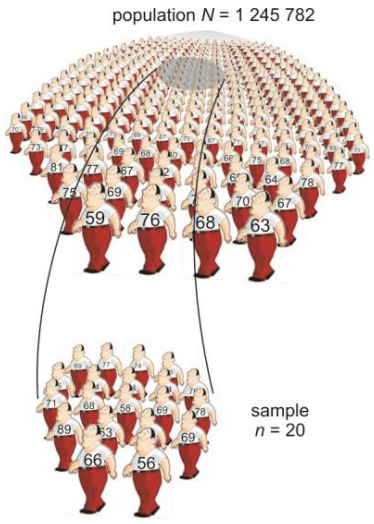


Fig. 3. *Illustration of population, variable (pulse rate, the values of which are indicated on the chest of the figures) and sample.*

variable
Variable
változó

A single measurement is not sufficient to answer the other questions either. In the case of the second question we assume that there is a fundamental deterministic normal pulse rate, and the pulse rate of the individuals is randomly scattered around that value. In this case we can imagine the population as a large but finite number (say, the total number of people in a country, $N = 1\ 245\ 782$, Fig. 3.) of individuals with their known pulse rates in that moment. These individuals together with their pulse rates will form the population. Thus, in this case the population has a finite number of elements.

In the first case the measurement was repeated on the same person (many times), and in the second case the pulse rate was measured (once) on a large number of people. The variable is very similar in both cases, but the population is different. The third and fourth questions will be discussed later.

Although the questions ask something about the population, in most cases we do not and cannot know the whole population. Therefore, we choose a **sample** from the population that contains a total N elements. **Sampling means choosing *n* elements, ideally randomly, from the population.** Sampling happens, for example, when we measure several times. Sampling makes sense only if n can be much smaller than N (see Fig. 3).

sample
Stichprobe
minta

DISTRIBUTION OF THE SAMPLE, FREQUENCY DISTRIBUTION, HISTOGRAM

frequency distribution
HäuFig.keitsverteilung
gyakorisági eloszlás

These concepts will be introduced through an example. Let us choose the second question from the list (*WHAT IS* the normal value of the pulse rate?) and measure the pulse rate of the students of a group. The student group together with the pulse rate data can be considered as a sample ($n = 20$) taken from the population related to the question (see Fig. 3). The collected data (x_i, where i = 1, 2, 3, 4, 5,…,20) are listed in the Table 4 below.

histogram
Histogramm
hisztogram

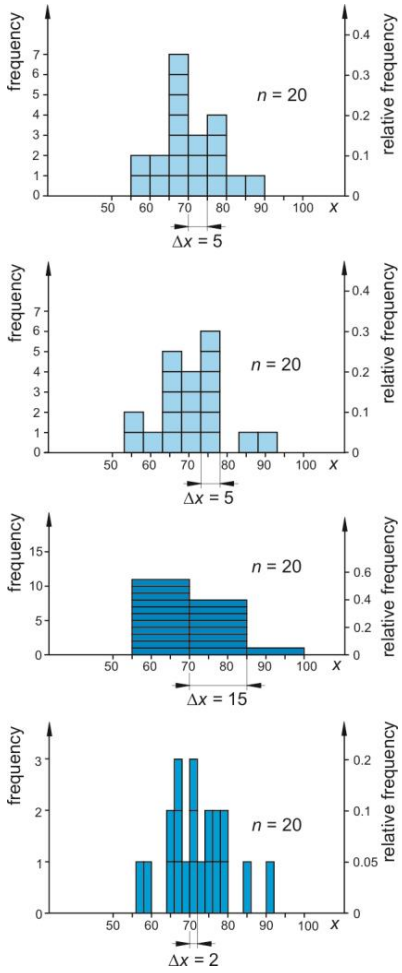
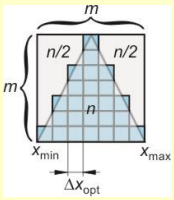


***Fig. 5**. Several possible histograms of the sample. The first was made based on Table 5. Every rectangle represents one observation.*

Comment 1:
The appearance of the histogram is "esthetic" if it is neither sporadic (contains no gaps), nor jam-packed into one or two classes (it has a structure).



If we want to construct the "optimal" histogram into a quadrangle area, then the number of classes (intervals) should roughly be equal to the maximum number of elements in one interval, both denoted by m. Then one element occupies a quadrangle area (instead a rectangle). According to the figure, the optimal number of the classes is:

$$m = \sqrt{2n}\ .$$

The optimal size of the classes (Δx_{opt}) can be obtained by dividing the difference of the maximum (x_{max}) and minimum (x_{min}) of the data by the optimal number of classes:

$$\Delta x_{opt} = \frac{x_{max} - x_{min}}{m}\ .$$

The first two histograms of Fig. 5 were constructed in accordance with these principles.

66	56	89	63	66	69	71	68	58	69
78	66	64	84	74	76	69	77	74	76

***Table 4**. Pulse rate values of the student group (example).*

Such a table may look much better than a simple list of numbers. In order to make sense out of the values and grasp their meaning, we have to organize them. If we plot our data on a coordinate line, the variation around a "normal" value becomes apparent.

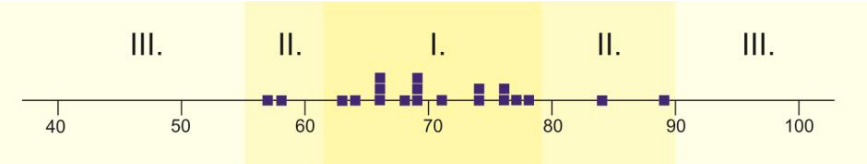


***Fig. 4.** Pulse rate data plotted on a coordinate line.*

Albeit arbitrary, three regions may be distinguished: I. many datapoints, II. few datapoints, III. no datapoints at all.

A reefinement of this picture leads to the concept of the **frequency distribution** that we get by grouping the data into classes. Let us make intervals of the same width (classes) through the axis and count the data (frequency) falling in these **classes**. The relative frequency distribution can be calculated as well. The intervals need not be of the same width, but this way it is easier to handle them.

Because the given set of individual data can be grouped in more than one way, many different frequency distributions can be constructed from the same dataset. Table 5 represents one of them.

CLASS LIMITS	FREQUENCY	RELATIVE FREQUENCY
$55 \le x_i < 60$	2	0.10
$60 \le x_i < 65$	2	0.10
$65 \le x_i < 70$	7	0.35
$70 \le x_i < 75$	3	0.15
$75 \le x_i < 80$	4	0.20
$80 \le x_i < 85$	1	0.05
$85 \le x_i < 90$	1	0.05
total:	$n = 20$	1.00

***Table 5.** One possible frequency distribution of the sample.*

The frequencies and relative frequencies can be represented graphically in a **bar diagram (bar graph)**. The graph consists of a series of rectangles, each with an area proportional to the frequency of data in the corresponding class interval represented on the horizontal axis. The method can be used with uneven class distribution as well. This graphic representation of data is called the **histogram**. Equal class widths are convenient, because in this case the frequency if proportional to the height of the rectangle.

Fig. 5 shows several different histograms constructed from the same pulse rate data. Class widths are equal in the first two histograms, but the class limits differ; in the last two cases the class widths were changed as well. There are no strict rules for constructing histograms, although some esthetic guidelines may apply (see Comment 1).

As we can see in Fig. 5, classes of the variable are represented along the horizontal axis of the histogram and the **absolute and relative frequencies** along the vertical axis. Every small rectangle (or square) corresponds to one measured value, thus the total number of rectangle units equals the total number of measurements (n = 20). This is the total area under the frequency curve. The total area under the relative frequency curve is always 1, or 100 % (because of the division with the total number of measurements n).
Although the shapes of the four histograms (Fig. 5) are rather different, which depends on their construction, some regularity can be seen. We can observe that all

of them have a "hill" roughly in the middle and around the same value, and their "width" is very similar. If the data size is increased and at the same time the class width decreased (there is no limit to continue the process), then the rough steps of the envelope observed imitially gradually smooth into a continuous curve (Fig. 6.).

DISTRIBUTION OF THE POPULATION, THEORETICAL DISTRIBUTION CURVE

Let us have a closer look at the tendency shown in Fig. 6. If the population consists of a finite number of elements (N), then upon increasing the number of the sample elements (n) the sample size will eventually reach the polulation size, hence the sample will contain all the elements of the population ($n = N$). **Thus, the distribution of a sample with N elements yields the distribution of the population.** The only uncertainty arises from the arbitrary choice of the class limits. For populations containing an infinite number of elements we can only say that upon increasing the sample size the sample distribution approaches better and better that of the population. In this case the population is described by a **theoretical distribution**.

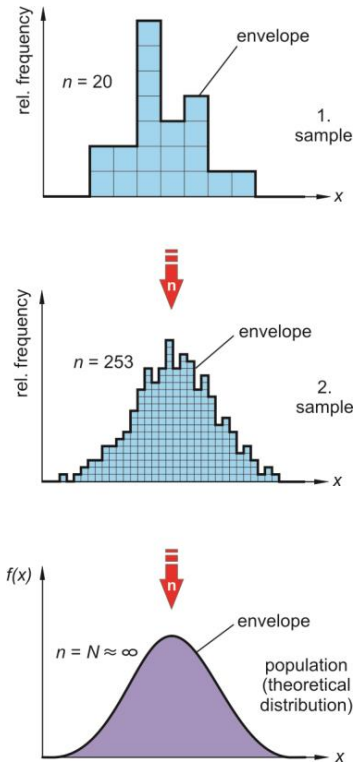


Fig. 6. *Increasing the data size and decreasing the class width gradually smoothes the envelope of the histogram.*

The population distribution determines all the properties of the variable. It provides the probabilities of all the possible values of the variable (nothing more can be said about the variable). Let us have an interval (a, b) on the coordinate axis. The probability that a randomly chosen value falls within the interval (a, b) equals the area that lies under the distribution curve in this interval (from a to b). If in the interval (a, b) the distribution curve has small values, the area under the curve is small, and the corresponding probability of incidence of these values of the variable will be low (Fig. 7/1). However, if in the interval (a, b) the distribution curve has large values, the area and the corresponding probability will be high (Fig. 7/2). If the width of the interval is increased, the area under the curve increases too, which means higher probability of incidence for these values (Fig. 7/3). Similarly to the histograms, the **total area under the curve equals 1**, because the "interval" (a, b) which in this case spans from $-\infty$ to $+\infty$ contains any randomly chosen value for sure. (See earlier remark about the inaccuracy and usefulness of a statement)

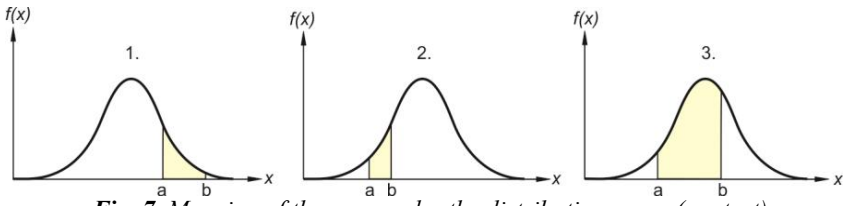


Fig. 7. *Meaning of the area under the distribution curve (see text).*

It is important to note that we always speak about an interval, because there is no area above a single value (the width of such an "area" would be zero). Consequently, in case of continuous variables probability that a randomly chosen value exactly matches a given number is zero. This technically means that all the measured data are different. In practice, however every measured number means an interval as we always use numbers with finite decimal places. The last digit is always rounded. (See earlier: continuous and discrete character of data).

The theoretical distribution describes all the possible data (i.e., the population), whereas the histogram concerns only the elements of a sample taken from the population (i.e., the data of the specific measurement).

PRINCIPAL THEOREM OF STATISTICS

Let us recall how we obtained the theoretical distribution: the number of elements in the sample, thus the number of measured data was increased. The principal theorem of mathematical statistics is that **in case of large samples, the empirical distribution function** (i.e., the envelope of the histogram) **approximates very well the theoretical distribution function**. Consequently, one may hope that **the more frequently data occur within a certain interval in the sample, the more probable is the appearance of these values in the population as well**.

A characteristic parameter of a population is determined by mathematical statistics through the examination of only a certain number (preferably few) of its elements. Sampling means choosing the elements to be examined (the sample) in a way that enables us later to draw reliable conclusions (inferences) about the whole population. This is usually achieved by **random selection of sample elements** (See Comment 2). Notably, the problems and aspects are particularly relevant in medicine.

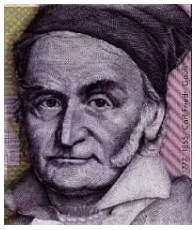
NORMAL OR GAUSSIAN DISTRIBUTION

normal distribution,
Gaussian distribution
Normalverteilung, Gauss Verteilung
normális eloszlás, Gauss-eloszlás

Depending on the examined variable, the **theoretical distribution may have different shapes**. However, **in most of the cases** it is a **symmetric bell-shaped curve with one peak** (we shall give the reason for this later on), which is called normal or **Gaussian distribution**. This type of distribution is illustrated in Fig. 6. and Fig. 7. The mathematical expression of the Gaussian distribution function is:

$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} . \qquad (1)$$

Th expression may seem somewhat complicated, but in fact it is a modification of the $f(x) = e^{-x^2}$ function, decorated with some parameters. The normal or Gaussian distribution is not a single distribution function. Due to its parameters it describes a whole family of them: the shape of the curves is similar, but their position, width and height may vary (see Fig. 8).



Carl Friedrich Gauss (1777-1855), German matematician.

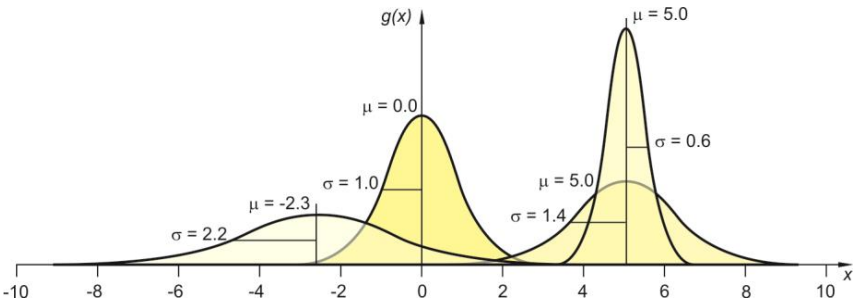


Fig. 8. *Some Gaussian distributions with different position (μ) and width (σ).*

Starting in the centre of the curve and working outward the height of the curve descends gradually at first, then faster and finally slower again, resulting in a bell-shaped curves with tails spanning to the infinity. Although the curve descends at the extremes toward the horizontal axis, it never actually touches it, no matter how far out one goes. The total area under the curve is 1 by definition (see earlier: theoretical distribution).

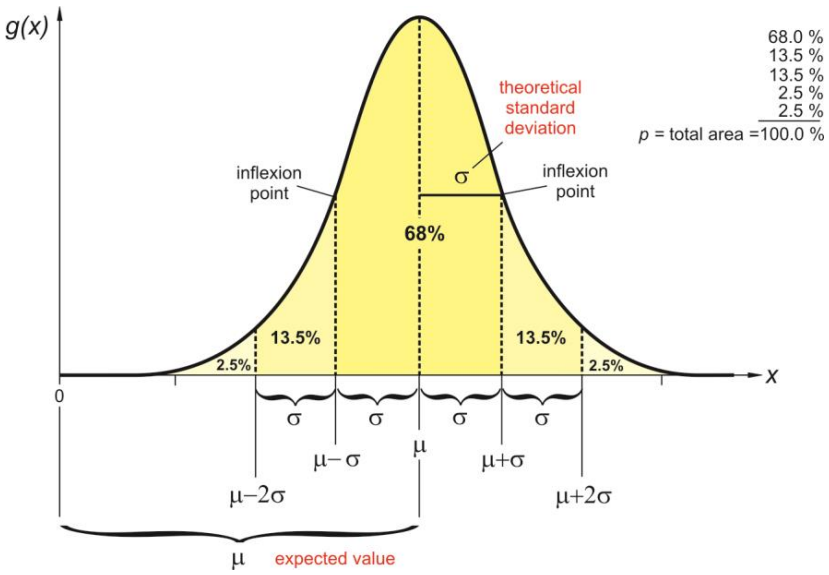


Fig. 9. *The bell-shaped Gaussian distribution and its parameters.*

Comment 2.
The sample has to be representative with respect to the population. It is a fundamental requirement that the distribution of the investigated parameter, apart from random sampling variations, has to be the same as that in the whole population. We have to keep this in mind when designing experiments.
When we organize a survey about the occurrence of thyroid problems, for example, we have to collect data from every region of the country, taking into consideration the density of the population. Over-representation of certain regions may lead us to incorrect conclusions. As an example, iodine-deficient tap water leads to much higher occurrence of the symptomes of hypothyroidism in the northern counties of Hungary than in the southern ones.

expected value
Erwartungswert
várható érték

theoretical standard deviation, SD
theoretische Streuung
elméleti szórás

empirical standard deviation, SD
empirische Streuung
tapasztalati szórás

mean, average
Durchschnitt
átlag

* (Unfortunately, the word "average" is often used to refer to *any* measure of central tendency, therefore it is better to use "mean", and we shall follow this practice.)

The **constants μ and σ** in the previous formula **are the parameters of the distribution**. These parameters specify one curve from the infinite number of possibilities. The p**arameter μ is the so-called expected value** that gives the position of the maximum of the curve on the x axis. The p**arameter σ is the theoretical standard deviation**, which characterizes the width of the distribution. The width of the distribution is roughly 2σ at the half height (more precisely: the so-called inflection points of the curve are at a distance of σ from the μ value) (Fig. 9). Based on this, the customary notation for the normal distribution is $N(\mu,\sigma)$.

Some general statements are valid for the relationship between the bell-shaped normal curve and its parameters. About two thirds (68 %) of the area lie under the curve between $\mu-\sigma$ and $\mu+\sigma$, and 95 % of the area is between $\mu-2\sigma$ and $\mu+2\sigma$. Only two thousandths of the area under the curve falls beyond the $(\mu-3\sigma,\ \mu+3\sigma)$ range, thus most of the area is included within a 6σ long section around the expected value. The h**eight of the curves** is not an independent parameter; it is **inversely proportional to σ** as a result of the fixed (unit) total area under the curve.

Among the infinite number of possible normal distributions there is a special one, for which $\mu=0$ and $\sigma=1$. This distribution is called the **standard normal distribution**, and according to the notation defined above it is $N(0,1)$ (second curve from the left on the Fig. 8).

The outstanding significance of the normal distribution is pointed out by the well known **central limit theorem** of probability calculus. According to this, **the values that are influenced by many little and independent effects follow a normal distribution**. This explains why the majority of variables occurring in nature are normally distributed.

As a "medical" example, the Gaussian distribution of body **height** and **blood pressure** can be mentioned. The height of adult men in Hungary corresponds to the $N(171, 7)$ distribution (measured in cm). The diastolic pressure, measured in Hgmm, of schoolboys follows the $N(58, 8)$ and that of smoking young men follows the $N(84, 10)$ distribution.

Let us have a closer look at the first example of the heights, where $3\sigma=21$ (cm). We can say that the height of the vast majority of adult men (more than 99 %) is between 150 and 192 cm. There are a few 2-meter-tall men, but this is not typical at all. The most common height is 170 cm, but one can meet men of 160 and 180 cm very often too. This shows an important feature of the **living world**: although there are **typical values**, the **diversity**, that is the difference between individuals is very important, too.

in the second example one may notice at a first glance that the blood pressure of smokers is not only higher (84 > 58), but its theoretical standard deviation is larger (10 > 8) as well. However, if one calculates the **relative** (theoretical) **standard deviation**, which is the σ/μ ratio, the situation will be the opposite ($10/84 \approx 0.12 < 8/58 \approx 0.14$). Often the relative standard deviation, which can be expressed in percentage ($(\sigma/\mu)\cdot(100\ \%)$), reveals more than the absolute standard deviation. Standard deviation may be small or large, but what is important is how large it is relative to the expected value. Thus, determination of both μ and σ is a very important task. The "exact" determination of parameters is, however, a tedious work, and in case of an infinite number of elements in the population it is impossible. The parameters will only be estimated.

ESTIMATION OF THE PARAMETERS, STATISTICAL PROPERTIES OF THE SAMPLE

We know that the Gaussian curve is determined unambiguously by its two parameters (μ and σ). Our goal is to give the best possible estimate for these parameters by using the data of a sample.

The **expected value** (μ) **is estimated most often by the mean** ($\bar{x}$), which is the arithmetic mean (average*) of the data (elements of the sample):

Comment 3.
Further options for estimating the expected value μ:
1. The **mode** is the number that occurs with the greatest frequency, namely the value corresponding to the maximum of the frequency distribution. As the frequency distribution is ambiguous (depends on the choice of classes), so is the mode. When there are only few data available, it is especially not a good attribute. (The **mode** of the data from the Table 4 is, according to the upper graph of the Fig. 5, between 65 and 70.)
2. The **median** is the middle one or the average of the two middle ones from the data organized in ascending order. Note that extreme data do not influence the value of the median. Especially when the measured extreme values are unreliable for technical reasons, this is the best estimate of the expected value. (The **median** of the sample given in Table 4. can be read from the Fig. 4. as 69.)
(In case of Gaussian distribution and a sample of large number of elements we have:
mean ≈ mode ≈ median
(Skewed distributions also exist where this statement is not valid.)

Comment 4.
Table of the data:

x_i	$n=20$
x_1	66
x_2	56
x_3	89
x_4	63
x_5	66
x_6	69
x_7	71
x_8	68
x_9	58
x_{10}	69
x_{11}	78
x_{12}	66
x_{13}	64
x_{14}	84
x_{15}	74
x_{16}	76
x_{17}	69
x_{18}	77
x_{19}	74
x_{20}	76
$\sum x_i = 1413$	

The mean of the pulse rate:

$$\bar{x}=\frac{\Sigma x_i}{n}=\frac{1413}{20}\approx 71\ (1/\text{min})$$

(rounded).
The **sum of squares**:

$$Q_x=\Sigma x_i^2-\frac{(\Sigma x_i)^2}{n}=$$

$$=101075-\frac{1413^2}{20}=1246.55$$

The **variance**:

$$s^2=\frac{Q_x}{n-1}=\frac{1246.55}{19}\approx 66\ (1/\text{min})^2$$

(rounded).
The **empirical standard deviation**:

$$s=\sqrt{\frac{Q_x}{n-1}}=\sqrt{\frac{1246.55}{19}}\approx 8\ (1/\text{min})$$

(rounded).
The **degree of freedom**: 19.

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}. \qquad (2)$$

The mean is the **most stable central tendency measure** of the distribution that is responsive to the exact value of each element of the sample, and is least sensitive to the change of the sample. What makes the mean of high importance is that the sum of all deviations from this number equals zero (because the sum of the negative deviations will be equal to the sum of positive deviations):

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum x_i - \sum \bar{x} = \sum x_i - n\bar{x} = 0. \qquad (3)$$

If we imagine the data spread across a board according to their values, then the mean corresponds to the position of the balance point of the distribution.

The theoretical standard deviation σ is estimated from the squares of the deviation of the points from the mean. It is called the **empirical standard deviation** (s), and it is defined as:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}. \qquad (4)$$

To avoid misunderstanding, the variable is often indicated in the subscript s_x. The square of the empirical standard deviation is called the **variance**:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}. \qquad (5)$$

Because the sum of the squared deviations (sometimes called sum of squares) present in the numerator of the above formula and very similar terms (quadratic expressions) will occur in our calculations very often, it is convenient to introduce a special notation (Q) for it. Because calculation of the sum of squares is rather tiresome, we derive an equivalent but more calculation-friendly form of this expression:

$$Q_x = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}. \qquad (6)$$

($n - 1$), the denominator in formulas (4) and (5), the expression, is called the **degree of freedom**. This term of statistical calculus is related to the estimation of parameters and is closely connected but obviously not always equal to the sample size (number of datapoints). In the beginning, the degree of freedom of the sample of n elements is n. If, however, a newly added value is estimated from pre-existing values of the sample, then the number of the actually used previously estimated values must be subtracted from the original number of freedom n. Because in the calculation of the empirical standard deviation the previously estimated mean of the same sample is needed, the degree of freedom is n–1. In more complicated cases there will be a formula given for the determination of the degree of freedom. (Comment 4 contains the calculation of the most important characteristics of the sample from Table 4).

variance
Varianz
variancia

degree of freedom
Freiheitsgrad
szabadságfok

The estimated and „true" values of a parameter are somewhat different. This difference is the error of the estimated parameter (discussed later). There are essentially two types of error: **inaccuracy and distortion**. Inaccuracy is the error which causes a random **deviation from the true value in either the positive or the negative direction**. **Distortion** causes the estimated value of the parameter to be **systematically smaller or larger than the "true" value** of the parameter. Whereas inaccuracy can be estimated, distortion cannot.

Using the definitions of the mean and empirical standard deviation given above (leaving distortions out of consideration) the following statement can be made: as the number of sample elements approaches infinity, the mean approaches the

expected value and the empirical standard deviation approaches the theoretical standard deviation with higher and higher accuracy. Or, with symbols:

$$\text{if} \quad n \to \infty, \quad \text{then} \quad \bar{x} \to \mu \text{ and } s \to \sigma. \tag{7}$$

The **empirical standard deviation *s* is the measure of variability of the data**. It gives the average **deviation of the data from the mean**. Similarly to the Gaussian distribution (Fig. 9), 68 % of the elements of the sample are within the interval ($\bar{x} \pm s$), 95 % are within the interval ($\bar{x} \pm 2s$), and more than 99 % are within the interval ($\bar{x} \pm 3s$).

The interval ($\bar{x} \pm k \cdot s$) calculated from a large number of data ($n \approx 1000$) contains exactly 95 % of the elements of the sample ($k \approx 2$) and it is called **reference range** or **normal range**. This is used mostly in the field of laboratory diagnostics. (In certain medical applications the interpretation of normal range can be different.) For 95 % of the healthy people the diagnostic parameter will fall in the normal range and for 5 % will be outside of it (see Comment 5).

In this case the distortion is not a problem, because if the entire dataset is systematically shifted, the reference range is shifted accordingly. One can sometimes observe this when comparing results from different diagnostic laboratories. Reference ranges can be slightly different for the same variable, because the applied protocols and apparatuses are not the same.

Comment 5.
"*the diagnostic value is in the normal range*"
Let's make this clear with an example. The probability of throwing 6 on a dice is 1/6, which is around 17 %. Hence the probability of not throwing 6 is 5/6, which makes (5/6 ≈ 0.83)

If instead of the regular dice we use an icosahedron (a regular solid shape having 20 faces) with numbered faces, the probability of throwing 20 is 1/20 = 0.05 = 5 %.

of not throwing 20 is %.
agine the strength of the *agnostic value is in the* it is of equivalent certainty to not throwing 20 by an icosahedron.

CONFIDENCE INTERVAL, ACCURACY AND RANDOM ERROR OF THE ESTIMATED PARAMETER

Let us emphasize once again that the error of the estimated parameter can be distortion as well, which is usually not possible to determine. Because of this, from now on the error will imply inaccuracy only, or **random error**. As stated before, if the number of the elements increases, the mean approaches the expected value more and more (see formula (7)), but we still do not know the answer to the question of how much the mean deviates from the expected value characteristic for the population in case of a sample of *n* elements. In other words**, what is the error of the mean?**

standard error
Standardfehler
standard hiba

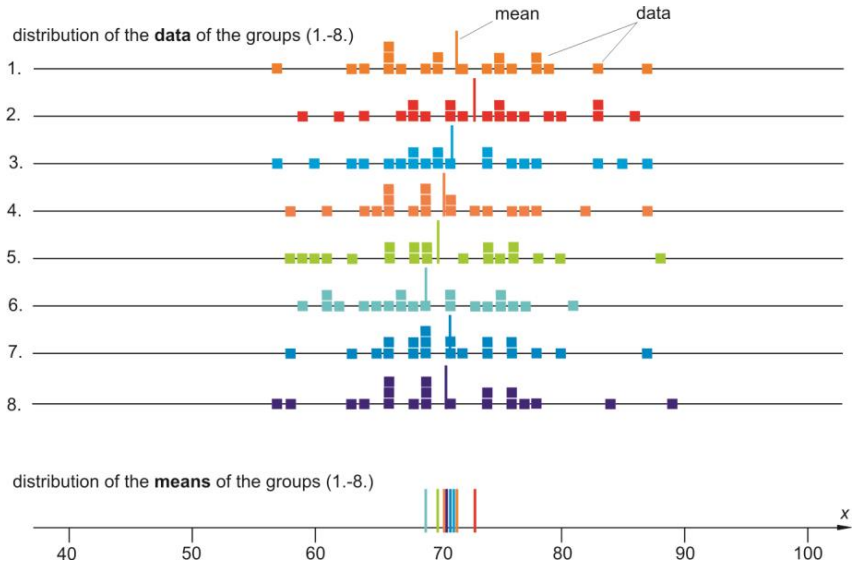


Fig. 10. *Random sampling distribution of means: pulse rate data and corresponding mean of eight groups (samples of 20 students). Note that the means of different groups "spread" much less than the data themselves.*

As discussed previously, the mean as a central parameter is not sensitive to the changes of the sample because all the elements of the sample are involved in the calculation, and, especially in larger samples, a single element plays a minor role in altering it. Thus, the sample means calculated from randomly selected samples (of

Comment 6.
From the pulse rate data we have already calculated the mean (71 (1/minute)) and the empirical standard deviation (8 (1/minute)). The standard error is

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{8}{\sqrt{20}} \approx 2 \text{ (1/min)};$$

and the error limit is (at 95 % confidence level)

$$\bar{x} \pm 2s_{\bar{x}} = 71 \pm 4 \text{ (1/min)}.$$

(rounded).

The result of the measurement can be stated as follows: **"based on our experimental data we can say with 95 % confidence that the expected value of the pulse rate of the examined population is in the 67-75 (1/minute) range"** (Fig. 11).
In order to have the expected value in the chosen range with greater certainty or accuracy, the number of data has to be increased.
How many is enough?
There is no a general rule, but for this situation we can say some considerations. Since the pulse rate is given rounded as an integer value, increasing the accuracy beyond ±1 (1/minute) does not make much sense. The question of certainty is ambiguous, but 99 % certainty or more is rarely needed. According to this, if we choose the error limit, then *n* must be increased until the error decreases to the value where the $3s_{\bar{x}} \le 1$ condition is satisfied.

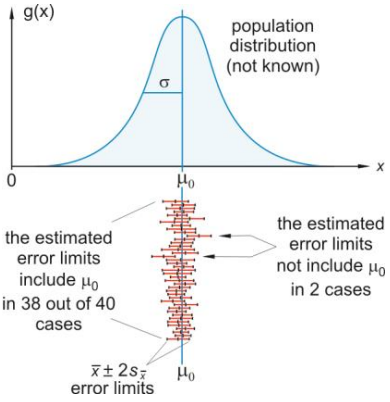


Fig. 11. *Error limit, calculated from the sample contains the expected value of the population with 95 % certainty.*

the same population) are not very different. In other words, the sample means ($\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_i$) "spread" much less around the expected value than the data (Fig. 10.).

This "spread" is expressed as the **standard error** (standard deviation of the sampling distribution of means):

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}. \qquad (8)$$

The final result is usually given in the form of:

$$\bar{x} \pm k \cdot s_{\bar{x}}. \qquad (9)$$

The result is always an interval, a limiting value in negative and positive directions enclosing the **expected value from both sides**. The problem is how to set the value of *k*. Where are these boundaries? This is ambiguous. The wider the range (*k* is large), the more likely that it includes the expected value, and our conclusion is probably right (see earlier: statistical inference). However, a wide range is rather useless in the everyday practice. The narrower the range, the higher the chance that the expected value is outside the chosen range, and the certainty of our conclusion becomes lower. Thus, increasing the estimation certainty and lowering the chance of a mistaken conclusion requires a wide range. By contrast, a narrow range is required for making professionally relevant interpretations.

confidence interval
Konfidenzintervall
konfidencia intervallum

Methods of statistics enable us to declare the extent of accuracy. Accordingly, we calculate a range of values about which we are reasonably confident (certain) that contains the "true" parameters. The interval is referred to as the **confidence interval**, its limits are called **confidence limits**, and the degree of confidence is the **confidence level**.

confidence level
Konfidenzniveau
konfidencia szint

Even though the value of *k* depends on the number of elements of the sample (degree of freedom), for large sample we can say, that if $k = 1$, the confidence level is around 0.68, if $k = 2$, is approximately 0.95 and for $k = 3$ it is greater than 0.99 (see table 6).

Confidence level (approximately)	68%	**95%**	99%
Confidence interval	$\bar{x} \pm s_{\bar{x}}$	$\bar{x} \pm 2s_{\bar{x}}$	$\bar{x} \pm 3s_{\bar{x}}$
		error limit	sure error limit

Table 6. *Confidence levels and the corresponding confidence intervals.*

Having learned all these we may answer the "*WHAT IS THE VALUE* of ...?" type questions (or present the final result of a measurement) according to formula (9) in the form of a confidence interval (see Comment 6, Fig. 11).

It is straightforward from definition (8) that the error decreases with increasing the number of data:

$$\text{if } n \to \infty, \qquad s_{\bar{x}} \to 0, \qquad (10)$$

It is visible that measuring many times has a good reason. For a fixed confidence level we can achieve the narrowing of the confidence interval beyond any limit just by increasing the number of data (see Comment 6).