BIOSTATISTICS

PRINCIPLES OF BIOSTATISTICS

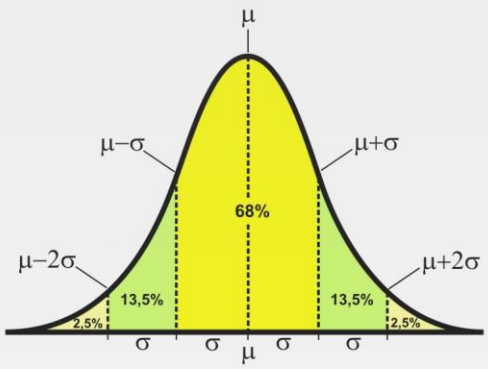


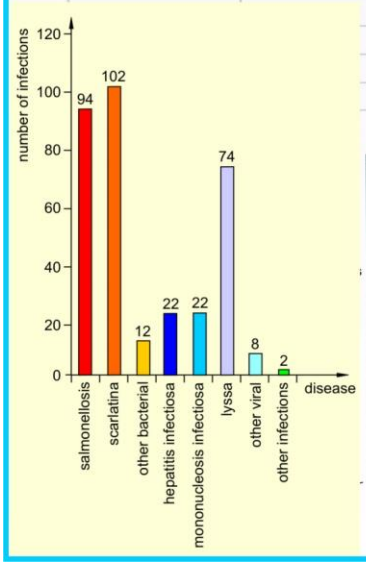
TABLE OF CRITICAL VALUES FOR T

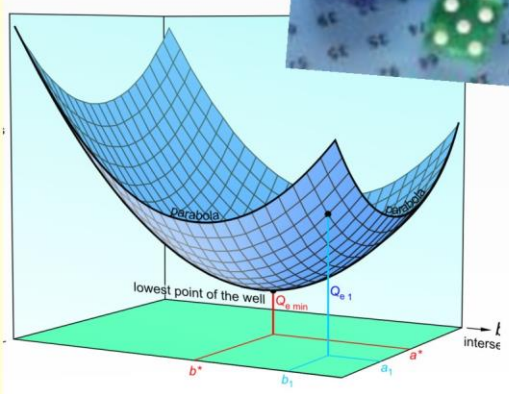
One-Tailed Significance

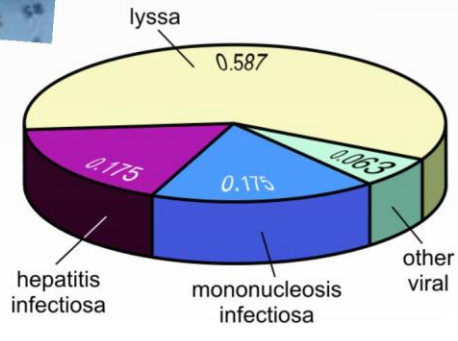
0.1	0.05	0.025	0.005	0.0025	0.0005	0.0002

Two-Tailed Significance

0.2	0.1	0.05	0.01	0.005	0.001	0.0005
1.89	2.92	4.30	9.92	14.09	31.60	44.70
1.64	2.35	3.18	5.84	7.45	12.92	16.33
1.53	2.13	2.78	4.60	5.60	8.61	10.31
1.48	2.02	2.57	4.03	4.77	6.87	7.98
1.44	1.94	2.45	3.71	4.32	5.96	6.79
1.41	1.89	2.36	3.50	4.03	5.41	6.08
1.40	1.86	2.31	3.36	3.83	5.04	5.62
1.38	1.83	2.26	3.25	3.69	4.78	5.29
1.37	1.81	2.23	3.17	3.58	4.59	5.05
1.36	1.80	2.20	3.11	3.50	4.44	4.86
1.36	1.78	2.18	3.05	3.43	4.32	4.72
1.35	1.77	2.16	3.01	3.37	4.22	4.60
1.35	1.76	2.14	2.9	3.33	4.14	4.50
1.34	1.75	2.13	2.9	3.29	4.07	4.42
1.34	1.75	2.12	2.9	3.25	4.01	4.35
1.33	1.74	2.11	2.9	3.22	3.97	4.29
1.33	1.73	2.10	2.8	3.19	3.92	4.23
		2.09	2.8	3.17	3.88	4.19
		2.09	2.8	3.16	3.85	4.15
		2.08	2.8	3.14	3.82	4.11
		2.07	2.7	3.12	3.79	4.08
		2.07	2.7	3.11	3.77	4.05







SUMMARY

DATA: qualitative or quantitative properties.
POPULATION, FUNDAMENTAL ENSEMBLE: group of those individuals (elements) or objects (together with observations) about which the investigator wishes to draw conclusions. The number of elements in a population is N, which may be infinite.
SAMPLE: an appropriate part of the population chosen for the examination in order to draw conclusions about the population. The number of elements in a sample is n, and usually $n << N$.
VARIABLE: a single general element x of the population.
ABSOLUTE FREQUENCY: the number of data in one class if the data are grouped in classes.
RELATIVE FREQUENCY: the ratio of the number of data in one class to the total number of elements in the sample.
FREQUENCY DISTRIBUTION: a list (table) or graphic representation of frequencies or relative frequencies in the classes.
HISTOGRAM: a graphic representation of the frequency distribution, in which a rectanglular area ("bar") with a width of the class represents the frequency within that class.
THEORETICAL DISTRIBUTION: distribution of the population.
NORMAL OR GAUSSIAN DISTRIBUTION, $N(\mu, \sigma)$: a type of theoretical distribution. It has a symmetrical chape and the following parameters: expected value (μ); theoretical standard deviation (σ). Experimental values influenced by a large number of random, independent but small effects follow normal distribution.
GAUSSIAN CURVE: Histogram envelope of a population with Gaussian distribution if $N = \infty$ and $\Delta x \rightarrow 0$, that is, if the population contains an infinite number of elements and the class width approaches 0.
EXPECTED VALUE (OF GAUSSIAN DISTRIBUTION): one of the parameters (μ) of the distribution, a value that corresponds to the maximum of the Gaussian curve. In other words, it is the value above or below which half the cases fall, and it is also the point that divides the area under the curve in half.
THEORETICAL STANDARD DEVIATION: measure of the width of the Gaussian curve (σ). Width of the Gaussian curve at half height is approximately 2σ.
MEAN, AVERAGE: arithmetical average ($\bar{x}$), the most accurate measure of central tendency used for the estimation of the expected value.
EMPIRICAL STANDARD DEVIATION (s): estimate of the theoretical standard deviation ($s = \sqrt{\Sigma(x_i - \bar{x})^2 / (n-1)}$). The degrree of scatter in the data can be described by measures of variability. The most often used one is the empirical standard deviation, which characterizes the average deviation of the data from the mean.
VARIANCE: square of the empirical standard deviation (s^2).
STANDARD NORMAL DISTRIBUTION, $N(0,1)$: normal distribution with parameters $\mu = 0$ and $\sigma = 1$.
STANDARD ERROR: standard deviation of the sampling distribution of means ($s_{\bar{x}} = s / \sqrt{n}$).
CONFIDENCE INTERVAL: interval given by the mean and standard error that contains the expected value with a given probability. The ($\bar{x} \pm 2s_{\bar{x}}$) error limit is used most often at 95% confidence level.
CONFIDENCE LEVEL: a number in % expressing the accuracy that corresponds to the confidence interval.
LINEAR REGRESSION: method of statistical inference about the linear relationship between two variables.
LEAST SQUARES METHOD: if certain conditions are fulfilled, the best fit line is the one for which the summed squares of the distances of the data points from this line is the smallest.
CORRELATION COEFFICIENT: a number (in the range between -1 and 1) representing the strength of the relationship between variables. ($r = 0 \rightarrow$ no relationship, $r = \pm 1 \rightarrow$ relationship of maximal strength)
HYPOTHESIS TESTING: methods used for deciding about the acceptance or rejection of a hypothesis.
NULL HYPOTHESIS (H_0): an initial statement in hypothesis testing. A specific hypothesis to be tested. A **negative** response to the asked question.
ALTERNATIVE HYPOTHESIS (H_1): a hypothesis that is valid if the null hypothesis is rejected. An **affirmative** answer to the asked question.
REGION OF REJECTION: if the null hypothesis is true, then this region contains the sample with a **very low probability**. If the sample falls within this region, then the null hypothesis is rejected.
REGION OF ACCEPTANCE: opposite, or better said complement of the region of rejection.
TYPE I ERROR: **rejecting** the null hypothesis when it is in fact **true**.
TYPE II ERROR: **accepting** the null hypothesis when it is in fact **false**.
STUDENT'S t -TESTS: statistical tests used for examining the hypothesis about the expected values of one or more variables of the normal distribution.
CHI-SQUARE TEST: statistical tests used for examining the independence of categorical variables.
DEGREE OF FREEDOM: the number of **independent** members or sample elements. Evidently, every element cannot be independent if there is a relationship between them. By subtracting the number of related elements from the total sample size we obtain the degree of freedom, that is, the number of independently chosen elements.

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

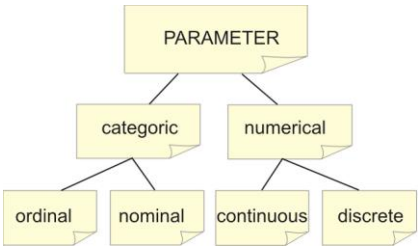


Table 1. *Classification of data.*

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**.

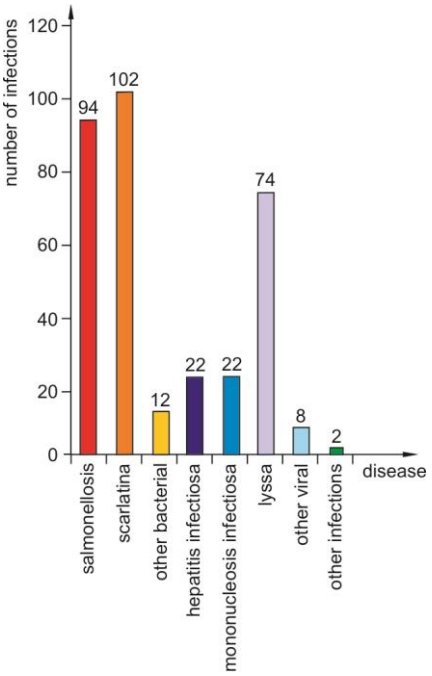


Fig. 1a. *Bar graph. Absolute frequencies of infections as a function of categories.*

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.

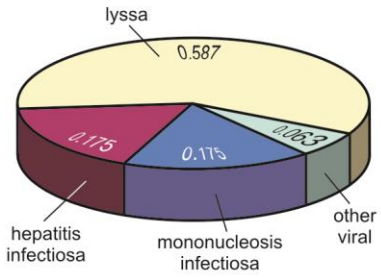


Fig. 1b. *Pie chart. Relative frequencies of the viral infections (categories).*

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**.

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."

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.

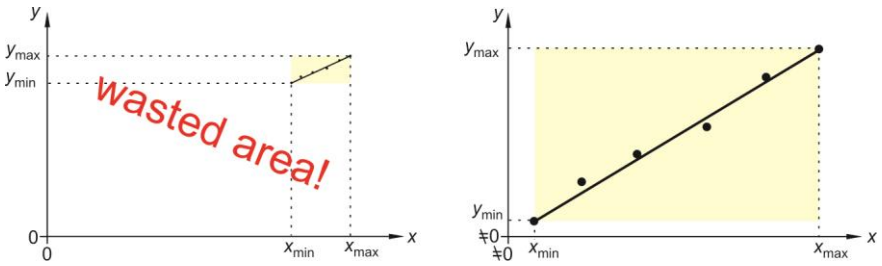


Fig. 2. *Improper and proper arrangement of the graph.*

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.

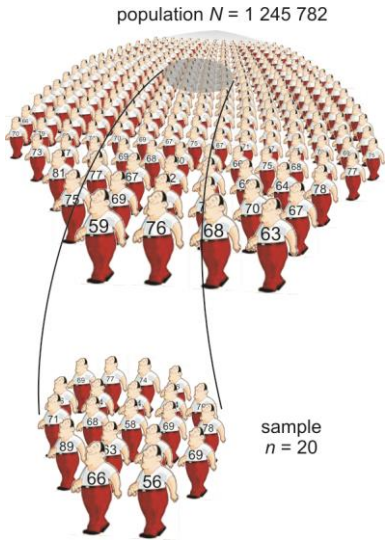


Fig. 3. *Illustration of population, variable (pulse rate, the values of which are indicated on the chest of the figures) and sample.*

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

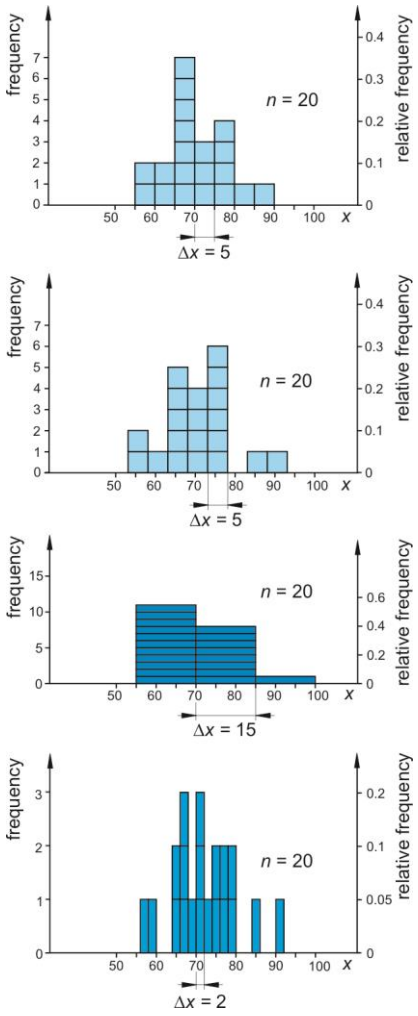
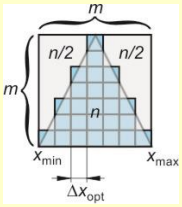


***Fig. 5**. Several possible histograms of the sample. The first was made based on Table 5. Every rectangle represents one observation.*

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.

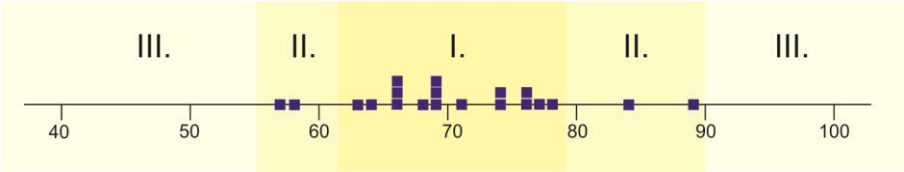


***Fig. 4.** Pulse rate data plotted on a coordinate line.*

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**.

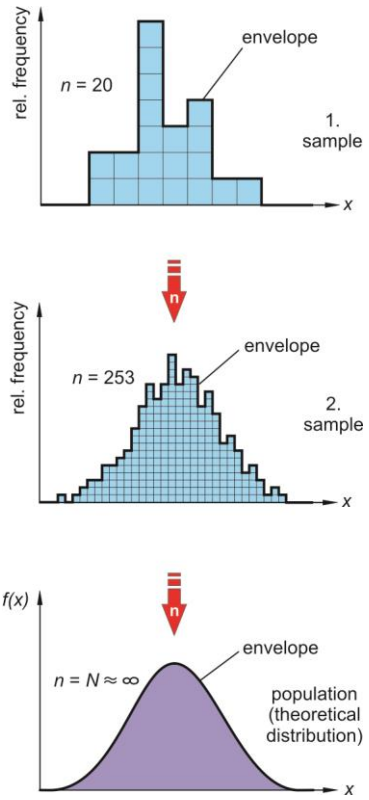


Fig. 6. *Increasing the data size and decreasing the class width gradually smoothes the envelope of the histogram.*

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)

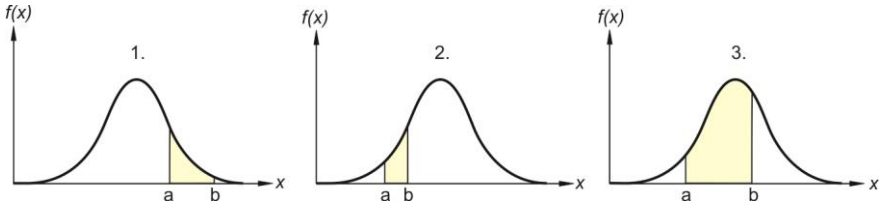


Fig. 7. *Meaning of the area under the distribution curve (see text).*

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.

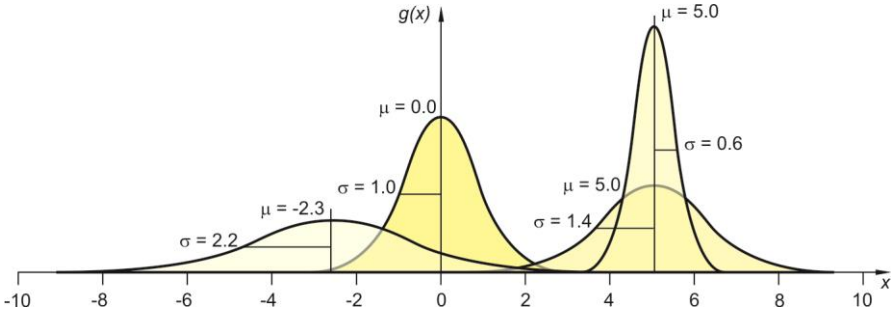


Fig. 8. *Some Gaussian distributions with different position (μ) and width (σ).*

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).

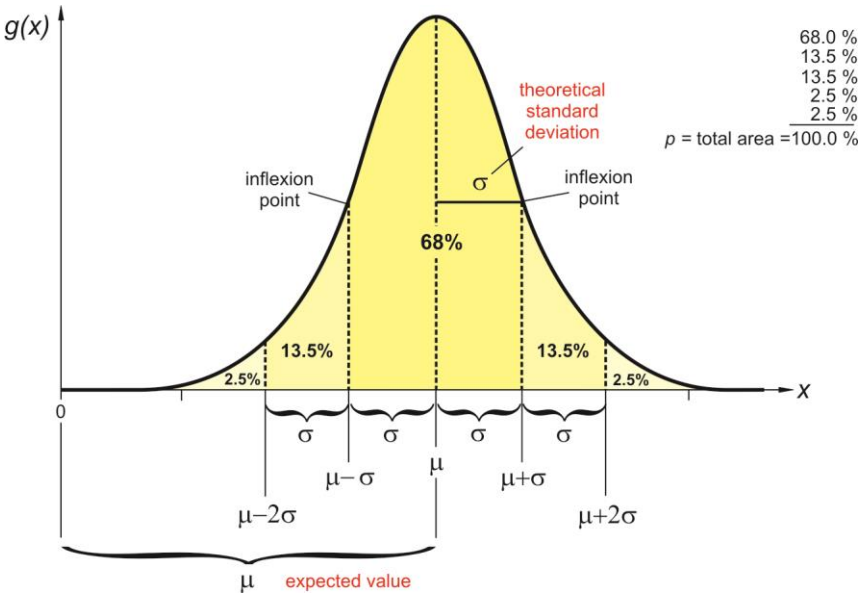


Fig. 9. *The bell-shaped Gaussian distribution and its parameters.*

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.)

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.

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

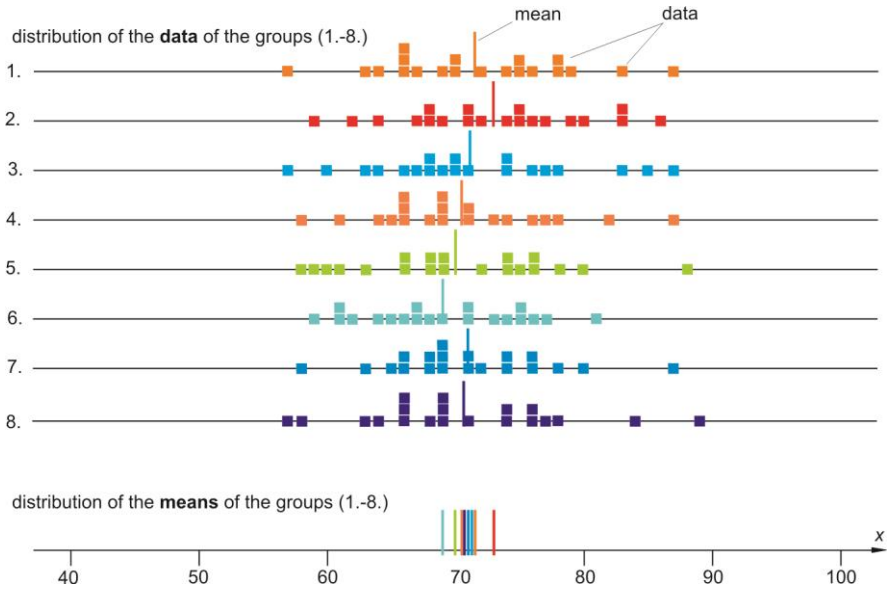


Fig. 10. *Random sampling distribution of means: pulse rate data and corresponding mean of eight groups (samples of 20 students). Note that the means of different groups "spread" much less than the data themselves.*

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.

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

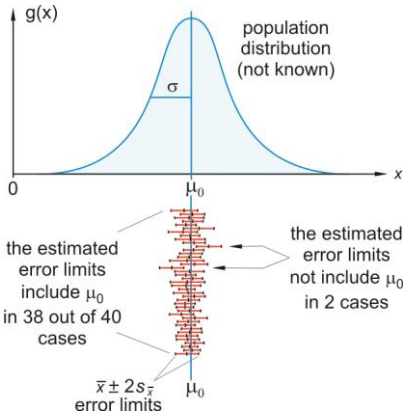


Fig. 11. *Error limit, calculated from the sample contains the expected value of the population with 95 % certainty.*

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).

GRAPHICAL DATA PROCESSING

We have already studied some aspects of the graphical representation of data. Here we will show how to use proper scaling of the graph axis in data processing.

The advantages of the linear curve (e.g., that the datapoints can be "connected" with a ruler) are so great that we always make an attempt to obtain straight lines, even when the relationship between the variables is obviously nonlinear. In these cases we make such transformations on the data that result in a linear function. Let us consider two examples, the exponential function and the power function.

The exponential function:

$$y = a \cdot e^{bx} \tag{11}$$

will attain the following form after a logarithmic transformation:

$$\lg y = (b \cdot \lg e) \cdot x + (\lg a) \tag{12}$$

Here the relationship between **lg *y*** and ***x*** is linear. The slope of the line is (b lg e), and the intercept is (lg a). If special, **lin-log scaled graph paper** is used, we do not have to calculate the logarithm of the y values, because the scaling does the job "automatically" (see Fig. 12, upper graph).

The power function:

$$y = a \cdot x^{b} \tag{13}$$

after a logarithmic transformation becomes:

$$\lg y = (b) \cdot \lg x + (\lg a) \tag{14}$$

Here we find a linear relationship between **lg *y*** and **lg *x***. The slope of the line is (b), and the intercept is (lg a). Calculation of the logarithms can be avoided again by using a **log-log graph paper**. Values of x and y are plotted according to the logarithmic ticks and tick marks of the axes (see Fig. 12, lower graph).

(Comment: plotting of the data and transformation of the scales is very easy by using dedicated computer programs. We can change the format of the axes from linear to logarithmic or change the range of the scale with a single command.)

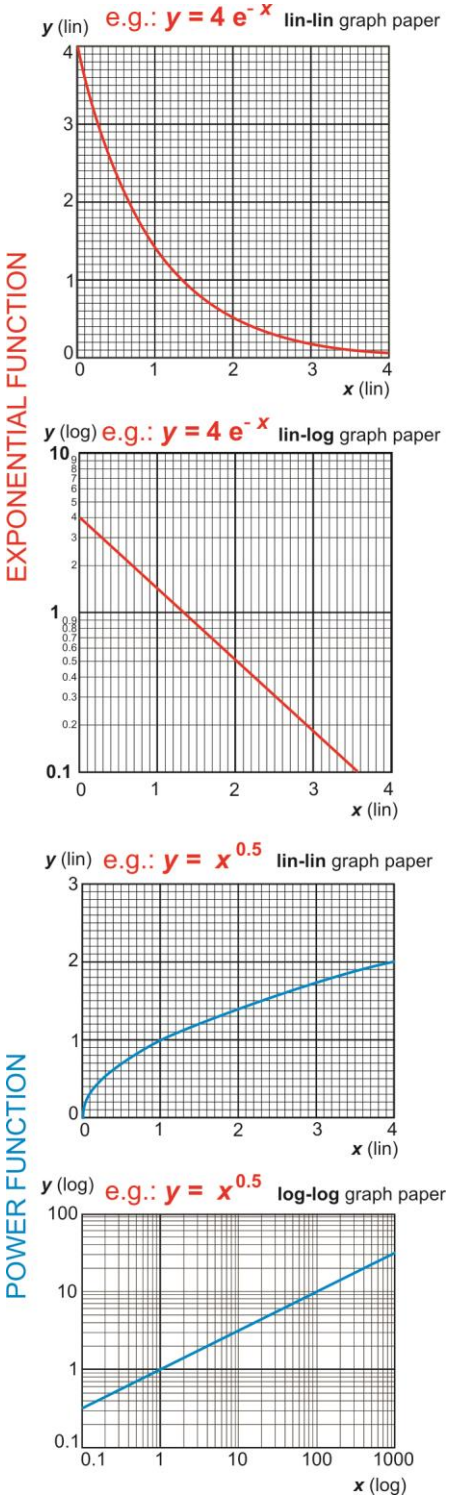


Fig. 12. *Linearization of the exponential and power functions by using lin-log and log-log graph papers, respectively.*

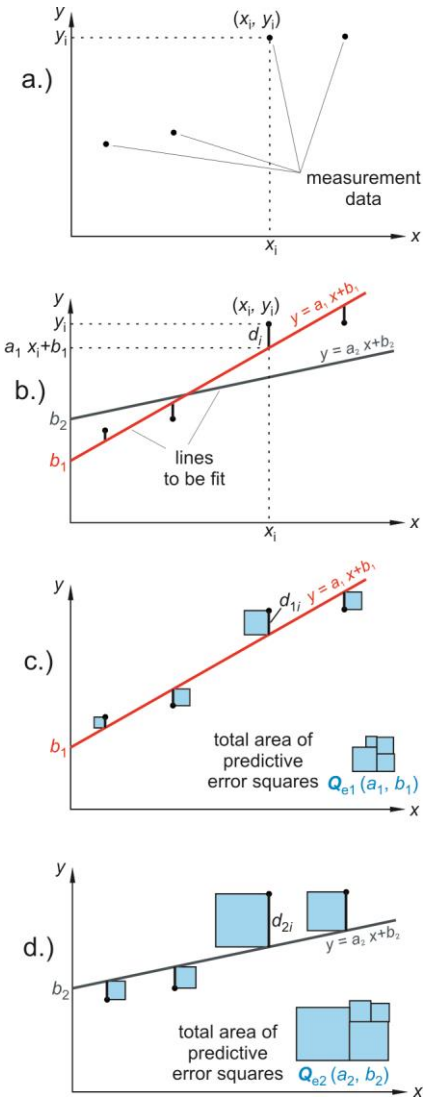


Fig. 13. *Finding the line of the best fit to the datapoints.*

LINEAR REGRESSION

The simplest curve, the straight line, is very easy to obtain subjectively with the aid of a paper, pencil and a ruler. However, we can hardly eliminate our doubts about the uncertainties arising from the errors of data points with the qualitative principles of drawing. Therefore, the problem of **finding the straight line of "best fit" to the data points still persists.**

The equation of a straight line is:

$$y = a \cdot x + b\,, \tag{15}$$

where a is the slope and b is the intercept. The y-intercept is the y value at $x = 0$, or the point of intersection of the line with the y axis. The straight line is determined explicitely by **these two parameters**. The task is to determine the actual values a^* and b^* of the parameters yielding the best fit to our data points. As a first step we will examine **what is meant by the line of best fit**.

Suppose that we have four data points (see Fig. 13a). Furthermore, let us assume that values of x_i are "exact" (without error), preset values, and only the y_i values have an error.

Let us draw an arbitrary line, determined by the parameters a_1 and b_1 ($y = a_1 \cdot x + b_1$), across the data points and calculate the vertical distance of the datapoints from this line (see Fig. 13b, vertical lines).

The distance of a selected (x_i, y_i) datapoint from the x axis is given by the y_i coordinate. At the same time the distance of the (x_i, y) line point from the x axis is obtained by substitution into the equation of the line as ($a\,x_i + b$). The difference is the **vertical distance of the point and the straight line** $d_{1i} = \left(y_i - \left(a_1 \cdot x_i + b_1\right)\right)$. This distance is positive if the point is above the line and negative if it is below. Let us calculate the squares of these distances computed for other points (predictive error squares).

Similarly to the sum of squares that we have defined in relation to the empirical standard deviation by formula (6), let us sum all the squares of distances calculated for the data points (see Fig. 13c, small blue squares) and denote the sum by Q_{e1}. Now, let us draw another line — determined by parameters a_2 and b_2 — and calculate the vertical distances (d_{2i}) of the data points from this line as well. As a result, we get a sum of squares again Q_{e2} (see Fig. 13d). Observe, that the points scatter around the line more when Q_e is larger ($Q_{e2} > Q_{e1}$).

As the data points (x_i, y_i) are always the same, Q_e is determined only by the variation of the parameters a and b. With this assignment we have defined a function with two independent variables a and b and a dependent one Q_e:

$$Q_e(a,b) = \sum_{i=1}^{n} \left[y_i - \left(ax_i + b\right)\right]^2 . \tag{16}$$

Due to the squared distances the function $Q_e(a, b)$ is of second degree in both variables. This means that it can be represented by a quadratic surface similar to a "well" or "pit" having parabolic traces (curves of intersections of the surface with planes parallel to the coordinate planes, where a = const, b = const).

If some other conditions are fulfilled (e.g., the standard deviations of the data points are independent), then the line of the best fit is the one for which the sum of squares $Q_e(a, b)$ has an absolute minimum.

We are looking for the coordinates (a^*, b^*) for which the function $Q_e(a, b)$ has the lowest value. In other words: we have to find the coordinates (a^*, b^*) of the lowest point of the well (the vertex of the paraboloid). The corresponding fitted line is called the **regression line**. The method of finding this line is known as the **least squares method**, or **linear regression** in general.

linear regression
lineare Regression
lineáris regresszió

The word regression means "return" or "reversion", expressing the inference to the connection (correlation) between the variables from the measured data. (Connection does not necessarily mean causality.)

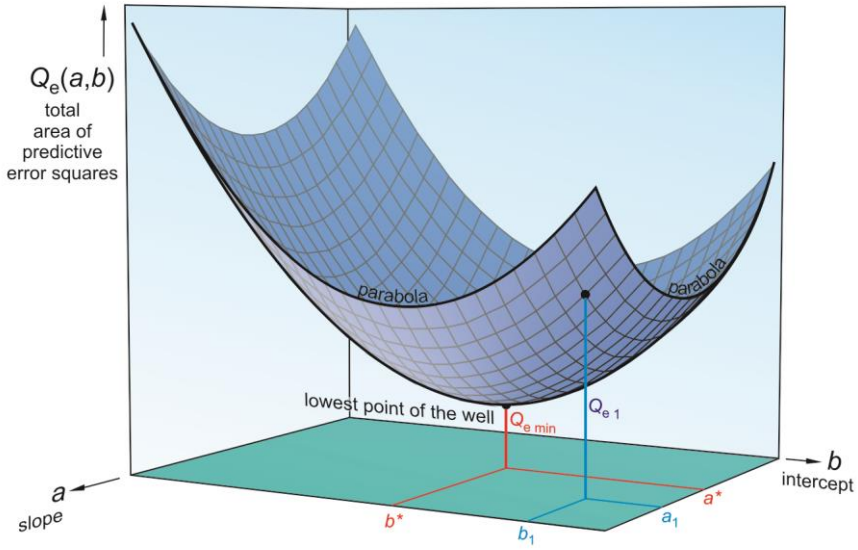


Fig. 14. *The predictive error as a function of the parameters* ***a*** *and* ***b****. The sum of error squareshas a minimum at the bottom of the well.*

After finding the minimum one obtains:

$$a^* = \frac{Q_{xy}}{Q_{xx}} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}, \text{ or } a^* = \frac{s_{xy}^2}{s_x^2}, \tag{17}$$

$$b^* = \bar{y} - a^*\bar{x} = \frac{\sum_{i=1}^{n} y_i}{n} - a^*\frac{\sum_{i=1}^{n} x_i}{n}, \tag{18}$$

where Q_{xx} and Q_{xy} corresponds to the notation introduced in (6),

— $s_{xy}^2 = Q_{xy}/(n-1)$ is the so-called **covariance**,

— s_x^2 is the **variance** of x, and

— $\bar{x}$, $\bar{y}$ are the **means**, respectively.

The line of the best fit can be calculated for any (x_i, y_i) pairs of data from the above equations, even if the points lie obviously along some curve, and not a straight line.

Since it is rather difficult to decide subjectively how well the fitted line approximates the experimental points, the determination of the **correlation coefficient** is desirable:

$$r = \frac{Q_{xy}}{\sqrt{Q_{xx}\cdot Q_{yy}}} = \frac{s_{xy}^2}{s_x s_y}, \tag{19}$$

where the notations are the same as in formula (17).

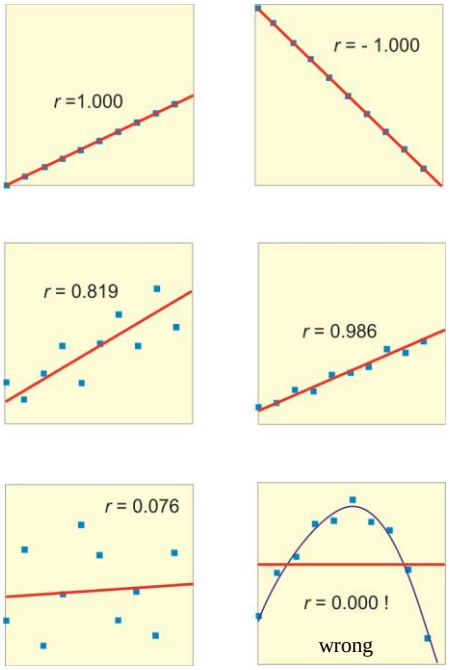


Fig. 15. *Some examples for the values of the correlation coefficient.*

correlation coefficient
Korrelationskoeffizient
korrelációs együttható

Comment 7.
In the following table the accommodation power of the eye is listed as a function of age:

age (years)	20	25	35	45
accommodation power (dpt)	11	8.5	7	3.5

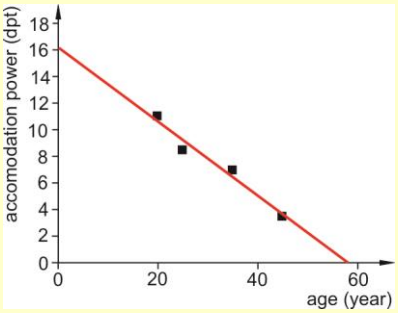


Fig. 16. *Accommodation power of the eye versus age. Fitting a regression line to the data points.*

After plotting the data and doing a linear regression the parameters of the fitted line are:

$a^* = -0.28; \quad b^* = 16.2.$

The correlation coefficient is:

$r = -0.98.$

Because there is no known model that would connect the two variables, parameters can be used for estimation by interpolation only. For example, we can estimate the accommodation power at the age of 40 as:

$-0.28 \cdot 40 + 16.2 \approx 5$ (dpt)

This value was missing from the table. Notably, however, we know that at very young age the accommodation power is definitely lower than 16 dpt. Therefore, the obtained relationship has limited predictive power outside the sample range.

The correlation coefficient characterizes **the strength of the correlation** (connection) between the variables, and it has values between +1 and –1. Positive values of the correlation coefficient belong to a regression line with positive slope, and negative ones to a line with negative slope. If the data points are very close to the regression line, value of $|r|$ will be close to 1 (e.g., $r = 0.9860$). If the line goes through every point, then $|r| = 1$, otherwise the more $|r|$ approaches zero, the greater the deviation of the data points from the regression line (see Fig. 15 and Comment 7 for the example).

We did not discuss yet in **what situation and to what purpose** can the regression analysis be used. An important **aspect** that we need to take into account is whether there is **a model** describing a causality **relation** between the variables (x, y) with parameters of physical meaning, or the line is fitted just to represent the datapoints with the necessary accuracy. Note that even if a model is not known, causality between the variables may exist.

An example for the first case could be the relationship between the index of refraction of solutions and their concentration (see 4. REFRACTOMETER), or the relationship between the optical density of solutions and solute concentration (see 6. LIGHT ABSORPTION). In these cases the parameters of the regression line can be used to **extrapolate** values (to estimate variables outside the measured range). However, the range of validity of the formulas (physical laws) should be taken into account.

In the second example parameters can be used only for **interpolation** (estimation of variables inside the measured range) and even if the correlation coefficient was close to 1, we may not infer a causality relation (see Comment 7).

STATISTICAL INFERENCE, HYPOTHESIS TESTING

hypothesis testing
Hypothesenprüfung
hipotézisvizsgálat

The goal of the calculations that we did so far was to approach, as best as possible, the parameters of the distribution of a variable from the **sample**. This kind of **quantitative inference** belongs to the field of **estimations**.

However, often we need a **qualitative inference**. That means, **we have to give a "yes" or "no" answer** to a question. We have already formulated this sort of questions earlier concerning the pulse rate. These are of the type: "Does it change...?" or "Is there a difference...?".

A decision is always made **based on the sample**. As answering of the formulated question always involves accepting or denying an initial assumption, called the hypothesis, this approach is called **hypothesis testing** (see Comment 8).

The main steps and the features of this testing procedure are demonstrated with the following example. **"Hypothesis testing" is done in court during criminal trials** as well. Although this example is certainly an oversimplification, it will help us to convey the main steps of the hypothesis testing process. The jury needs to decide whether the accused is guilty or not (the judge considers the extent of the sentence only). Thus, there is the following yes/no question posed: **is the accused guilty?** In most jurisdictions the **presumption of innocence** is applied, therefore the accused is considered innocent until his guiltiness is proven (i.e., the jury needs to prove guiltiness and not innocence). Thus, the **"not guilty"** statement is the presumption made by the court. In other words, it is the **initial hypothesis**.

The prosecuting attorney (representing the prosecution) has to substantiate the charge with evidences. The defense attorney (representing the defense) tries to weaken the reliability of the evidences. At the end, the jury evaluates and considers the "strength" of the evidences and makes a verdict. The verdict (or decision) means **accepting or rejecting the presumption, the initial hypothesis**, namely the **"not guilty"** statement. Regardless of the verdict, the decision of the jury may be right or wrong. Thus, a total of four different outcomes of the process may happen.

The **decision** of the court is **correct** (right)
- if the jury **accepts the "not guilty" hypothesis** and the accused is in fact **not guilty**; or,
- if the jury **rejects the "not guilty" hypothesis** (the verdict was guilty) and the accused is in fact **guilty**.

The **decision** of the court is **incorrect** (wrong),
- if the jury **accepts the "not guilty" hypothesis**, but the accused is in fact **guilty**;
- or, if the jury **rejects the "not guilty" hypothesis** (the verdict is guilty), but the accused is in truth **innocent** (see Comment 9).

Statistical **hypothesis testing** differs from the judicial procedure (irrespective of the simplifications) in a sense that the **considerations are based on numerical arguments**, therefore the decision is less influenced by subjective elements.

We will solve an example (see Problem 1) according to the procedure outlined above:

The specific **question** to answer is: *should the sales of a medication be banned because the active ingredient content has changed,* or, more simply, *does the active ingredient content of the tablets differ from that specified* (c.f., "is the accused guilty")*?*

The statement to be tested, or the **presumption,** or the **initial hypothesis** is: *The further sales of the medication shoud not be banned, because the active ingredient content* ***does not differ*** *from the specified value* ("presumption of innocence", not guilty).

Evidences: *The set of data of active ingredient content in mg unit.*

The subsequent steps (**evaluation, consideration and conclusion**) will need a more detailed description. If we had all the pieces of information, that is, if we

Comment 8.
Types of the hypothesis that are investigated most often:

1. *Hypothesis about a parameter of the distribution*. For example, we know, that a variable has normal distribution and we want to test the hypothesis that the expected value of the distribution equals a number μ_0. This type of test is needed as well to decide if a variable was changed or not.

2. *Hypothesis about the parameters of two (or more) distributions*. For example, we know that two independent variables have both normal distributions. We want to check the hypothesis that the expected values of the variables are equal. This way we can answer questions such as do women live longer than men in a given population (or in other words, is there a difference between their expected lifetimes).

3. *Independence test*. The tested hypothesis is formulated as if two or more variables are independent (if there is a connection between them).

4. *Homogeneity test*. The question is, if the distributions of two (or more) variables are identical.

Comment 9.
Possible decisions of the jury:

The fact	The sentence	
	acquitted	sentenced
not guilty	**right**	**wrong**
guilty	**wrong**	**right**

Problem 1:
One of the conditions for the continuous sales of a medication is maintaining a 6-mg content of the active ingredient. In a quality control experiment the measured data of some arbitrarily chosen tablets were (in mg): 6.05, 5.95, 5.75, 5.9, 5.95, 6.05. Should the sales of the medication be banned based on the different content of the active ingredient?

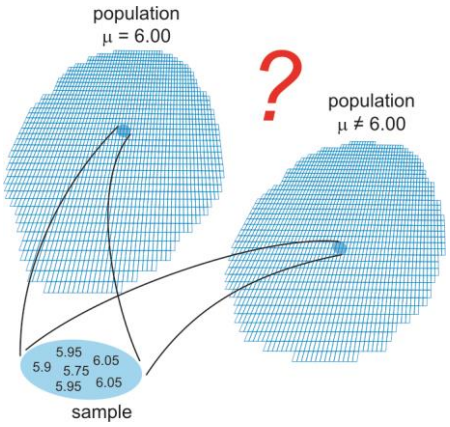


Fig. 17. *Possibilities of the origin of the chosen sample. Values are in mg units.*

knew the population (active ingredient content of each (!) tablet), or, equivalently, its distribution, then we would simply need to compare the expected value of the known distribution of the population (μ) with the specified 6 mg value. In this case, further consideration is not required. We simply conclude that if $\mu = 6$, then we accept, but if $\mu \neq 6$, then we reject the initial hypothesis. With this step we just quantified the initial hypothesis, because $\mu = 6$ is equivalent to the statement: "*The further sales of the medication* ***shoud not be banned****, because the active ingredient content* ***does not differ*** *from the specified value*". We should mention that in practice it is also important that the tablets contain identical amounts of the active ingredient, therefore the standard deviation should be examined as well.

As this ideal case never really happens, the situation is more complicated. We know that the well-chosen confidence interval calculated from a sample includes the expected value only with a given certainty. Thus, **first we choose** the "necessary" certainty (i.e., the **confidence level)**, and **then we check if the corresponding confidence interval includes the value 6**. If **yes**, we **accept**, but if **not, reject** the initial hypothesis. In the end the decision is **based on the sample** at a previously chosen and **fixed confidence level**.

In a different approach, we may assume that there exists a population with an expected value $\mu = 6$, and the question is whether the arbitrarily chosen sample is from this population, or from another one with a different expected value, $\mu \neq 6$ (see Fig. 17)?

Let us choose the confidence level to be 95 %. This means that from 100 arbitrarily chosen similar samples it may happen only 5 times that the corresponding confidence interval does not include 6. We know that this interval can be calculated from the formula $\bar{x} = k \cdot s_{\bar{x}}$ and that for large samples $k \approx 2$. In our case (as the sample is small) k is not yet determined, but we will solve this problem soon. If we use $k = 2$, based on the data (see Problem 1.) the confidence interval is between 6.03 and 5.85. As this includes 6, we **accept the initial hypothesis**.

The result means that the experiment did not provide enough evidence for banning the further sales of the medication. Therefore, the "*further sales* (provided that this was the only criterion*) cannot be denied*", because the active ingredient content does not differ from the specified value - it is not outside the confidence interval.

A similar procedure is used during the discussion of the statistical tests, which follows soon, but for a better understanding we have to explain an important mathematical tool first.

TRANSFORMATION OF THE VARIABLE AND THE DISTRIBUTION OF THE NEW VARIABLE

Transformation of data was already mentioned in the GRAPHICAL REPRESENTATION section, where our goal was to "connect" measured datapoints with a straight line. Here, we will discuss the question of transformation in detail.

When we obtain the data after performing some specified mathematical operations, the result is basically a **new variable**, because from different initial data we get a different result. This type of conversion is called in general a transformation. Among the simplest transformations we can mention the addition of a constant or the multiplication with a constant, but calculating the mean or the empirical standard deviation are transformations as well (see Comment 10).

Let us take an element x (the variable) of a normally-distributed population $N(\mu, \sigma)$, and carry out the transformation $x^* = (x - \mu)/\sigma$. In other words, this operation is carried out on every element of the population.

The question is what the distribution of the new variable x^* is like. As a first step, all the elements are shifted by the expected value, which yields the $N(0, \sigma)$ normal distribution. As a second step, the differences from 0 are divided by σ (shrinking), which yields the $N(0, 1)$ standard normal distribution (see Fig. 18). This transformation has an important practical advantage, as any normally-

Comment 10.
Transformations:

Transformation of a variable in fact means that its value is expressed on a different scale; to every original value a new value is assigned on that new scale.

The conversion of a physical quantity from one system of units to another is also a transformation. E.g.: energy [eV]·1.6·10^{-19} = energy [J].

The goal and sense of the transformation is that such statistical procedures can be applied on the transformed variable that are not possible on the original. However, the conclusions will be valid for the initial variable as well.

By using a **normalizing transformation,** a normally-distributed variable can be obtained from one that is not of normal distribution.

Categorizing transformations convert variables of continuous distribution into ordinal or nominal variables. This is useful in cases when the examined phenomenon changes qualitatively in parallel with the change of a continuous quantitative parameter.
This type of variable is, for example, the age, for which we can apply the following transformation in order to get a two-value nominal, so-called binary variable from a continuous variable:
age < 18 years → child (0)
age ≥ 18 years → adult (1).

Rank transformations convert values or ordinal variables. The elements of the sample are sorted ascending, and the ranks (rank numbers in the list) are used instead of original values. Several non-parametric statistical procedures are based on this rank transformation. (See chapter of EXAMPLES FROM THE FIELD OF MEDICAL STATISTICS)

distributed variable can be transformed into one with standard normal distribution, which makes standardized data processing possible.

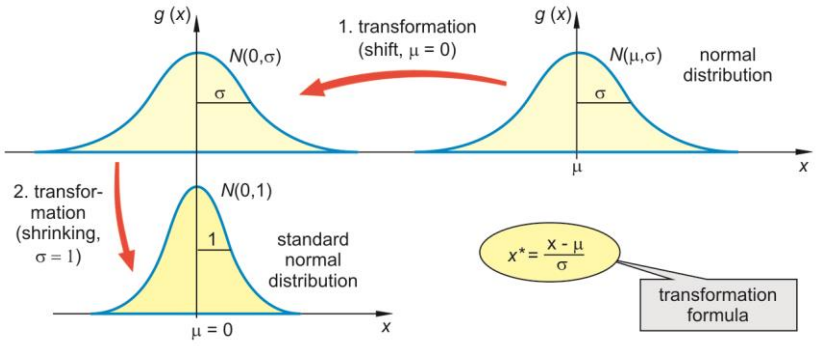


Fig. 18. *Transformation of a normal distribution of general position and width into standard normal distribution ($N(\mu,\sigma) \rightarrow N(0,1)$)*

Now, we would like to use the "same" transformation in a case when σ is not known. Therefore, we will use its estimated value, the empirical standard deviation (s) calculated from the sample of n elements. The variable obtained from this transformation has a similar distribution but not exactly the same as that of N (0, 1). The new distribution is called **Student's *t*-distribution.** For n–1 degrees of freedom $t = (x–\mu)/s$ (see Fig. 19). It is not surprising that this distribution depends on n, as in the transformation we have a parameter s that depends on n. Further properties of the distribution will be discussed in the next section.

Student's distribution of *t*
Student oder *t*-Verteilung
Student-, vagy *t*-eloszlás

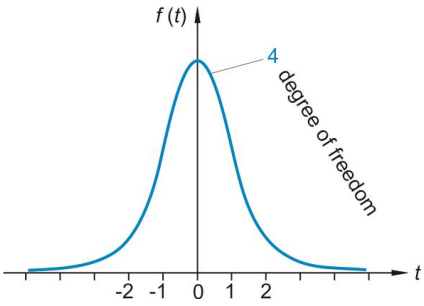


Fig. 19. *Student's t distribution for 4 degrees of freedom. The curve resembles the N(0,1) distribution, but depending on the degree of freedom it differs more or less.*

Different transformations lead to different distributions. Let us see another example: if we have n variables distributed as N (0, 1) and we sum the squares of these variables, then the distribution of the new variable is called a **χ^2-distribution** of n degrees of freedom (see Fig. 20). Naturally, this variable cannot have a negative value.

χ^2-distribution
χ^2-Verteilung
χ^2-eloszlás

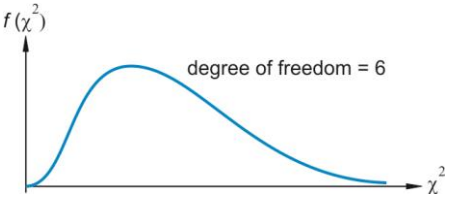


Fig. 20. *The χ^2-distribution of 6 degrees of freedom*

STATISTICAL TESTS

Although the hypothesis test can be performed as it was shown before (although we did not yet specify the value of k), we will rather use **statistical tests** for their simplicity.

statistical test
statistischer Test
statisztikai próba

There are **many different types** of statistical tests depending on the **hypothesis** to be tested, the **conditions of the application** and **the way of realization** of the method, but all of them have **the same basic logic**.

Our initial assumption is that **parameters estimated from a sample have some kind of distribution**; hence if we choose another sample, the estimated parameters will be different. The exact shape of the distribution depends

- first on the distribution of the initial variable,
- second on the parameter or statistical characteristic (it can be the correlation coefficient r as well) considered, and
- third on the number of elements in the sample, or more exactly the degree of freedom.

For the sake of simplicity, instead of many possible distributions we will use only a relatively small number of their standardized versions. To achieve this, we will always transform the examined estimated parameter or statistical characteristics corresponding to the given standardized distribution to the desired shape (see the previous section, like $N(\mu, \sigma) \rightarrow N(0, 1)$).

The simplest and most often used statistical tests are the ***t*-tests** and the **χ^2-tests**. A standardized theoretical distribution belongs to both tests, the t- and the χ^2-distributions, respectively. These are in fact families of distributions as the degree of freedom – as a free parameter – affects the shape of the particular distribution.

William S. Gosset (1876-1937), a famous English statistician wrote under the pseudonym of "Student". In the course of his work on small-sample quality control for the Guiness Brewery in England, Sudent realized, that what we have called t was not distributed precisely as normally distributed x^*, and provided a way to the solution. As a result, we know the proper distribution of this statistics. In honor of Gosset's contribution, the resulting family of distributions is known as Student's distribution or Student's t-distribution.

Let's get back to the previous example (Problem 1) and see what the hypothesis

t-test
t-Test
t-próba

χ^2-test
χ^2-Test
χ^2-próba

null hypothesis
Nullhypothese
nullhipotézis

alternative hypothesis
Alternativhypothese
alternatív hipotézis

two-tailed test
zweiseitiger Test
kétoldalú próba

one-tailed test
einseitiger Test
egyoldalú próba

testing means in the "language" of the tests. The **question** to be answered does not change: *should the sales of a medication be banned because the active ingredient content has changed*, or, more simply, *does the active ingredient content of the tablets differ from that specified* (c.f., "is the accused guilty")?

The initial hypothesis that needs to be tested and about which the decision will be made is called the **null hypothesis** (the reason is discussed later): *the further sales of the medication* ***cannot be banned****, because the active ingredient content* ***does not differ*** *from the specified value* ("presumption of innocence", not guilty).

Let us quantify the null hypothesis: the **sample** of given mean ($\bar{x} = 5.94$) **was chosen from the population of expected value** $\mu_0 = 6$, and not from the one with $\mu' \neq 6$. As we **do not know the expected value μ of the population**, only the mean of the sample, as a matter of fact $5.94 \approx 6$ and we could say that the difference is "not real" but a result of random sampling variation. Thus, $\mu = \mu_0$ and $\bar{x} \approx \mu_0$, or, in other words $\mu - \mu_0 = 0$ and $\bar{x} - \mu_0 \approx 0$. The null hypothesis, which is usually denoted by H_0 can be formulated as $\mu - \mu_0 = 0$, but this cannot be tested directly. Therefore, only the $\bar{x} - \mu_0 \approx 0$ statement remains, although this statement might not be satisfied because of random sampling.

An alternative hypothesis H_1 must also be specified. H_1 is valid if the null hypothesis is rejected. Our first thought might be that defining H_1 is completely useless, because formulating a statement opposite to H_0 is straightforward. That is, if H_0: ($\mu - \mu_0 = 0$), then H_1: ($\mu - \mu_0 \neq 0$). This statement is true in most of the cases, hence **H_0 is rejected both if** $\mu - \mu_0 < 0$ **and if** $\mu - \mu_0 > 0$. This is called the **two-tailed test** (and a non-directional alternative hypothesis).

On some occasions however, we are interested only in one direction of the opposite statement. For example, if we examine the effectiveness of medications such as antihypertensive or antipyretic drugs, then only a reduction of the expected value corresponds to effectiveness. We assume that the elevation of blood pressure or body temperature occurs only by chance. The null hypothesis remains the same H_0: ($\mu - \mu_0 = 0$), but the alternative hypothesis is formulated as H_1: ($\mu - \mu_0 < 0$), hence **H_0 is rejected only if** $\mu - \mu_0 < 0$. This is called the **one-tailed test** (and a directional alternative hypothesis).

If the error (result of the random sampling) **is large enough**, then the formulation of the null hypothesis as $\bar{x} - \mu_0 \approx 0$ is nearly **true** and acceptable. However, **if the error is small**, then we cannot be certain, and the **difference might be "real"**. In order **to make a decision the measurable difference $\bar{x} - \mu_0$ and the standard error** (characteristic of the random variations, standard deviation of the sampling distribution of means, $s_{\bar{x}}$), **need to be compared**. In the previous section we have seen that the variable x of $N(\mu, \sigma)$ distribution can be transformed by formula $t = (x-\mu)/s$ into a Student's t-distribution of $n-1$ degrees of freedom. Now we suppose that if we use instead of the variable and its theoretical standard deviation the mean and its empirical standard deviation in the transformation formula, we will get the same result, thus

$$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}. \qquad (19)$$

In this case the use of the yet-to-be described t-test seems to be the best for this purpose.

ABOUT THE *t*-TESTS IN GENERAL

Fig. 21 shows the Student's t distribution again, but now for different values (2, 4 and ∞) of the degrees of freedom. Note especially that the expected value of the distribution is always 0 ($t = 0$) and its shape resembles the standard normal distribution ($N(1, 0)$) as mentioned earlier. Finally, for infinitely large samples (where degrees of freedom approach infinity) the t distribution and the normal curve are identical. You can see that the value **$t = 0$ corresponds** to the recently

formulated **null hypothesis** ($\bar{x} - \mu_0 \approx 0$), but **the t_s value that is calculated from the sample** characterizes whether the data stand for rejecting or accepting the null hypothesis, thus it is **the measure of the "strength" of the data** (evidences). By analogy of the judicial example, the goal of the prosecution is to increase *t*, and that of the defense is to decrease it.)

Distributions of several statistical characteristics can be converted by specific transformations into the Student's *t* distribution. **Change, deviation, difference and correlation** of variables are always measured by a parameter. This parameter is estimated by a statistical property of the sample, and its standardized form always yields a t_s value.

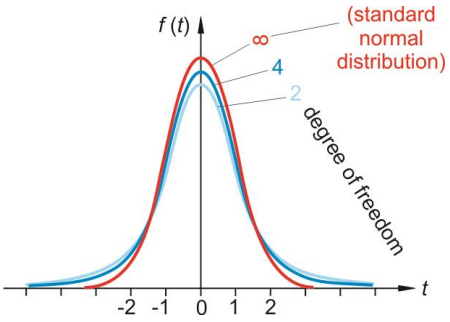


Fig. 21. *The Student's t-distribution for three different degrees of freedom.*

The calculated **t_s value should be zero in principle** if there **is no change, no deviation, no difference or no correlation**. This is the reason why the "no" answer to the original yes/no question is called the null hypothesis. The null hypothesis has a unique role in hypothesis testing. Irrespective of the statement to be confirmed by the study, during the test the validity of the null hypothesis is always assumed. In the end this statement is either accepted or rejected, and the answer to the initial question is negative or positive, respectively.

It is easy to find the reason for the above logic. In case of a **positive answer** to the initial question, the number of possible answers can be infinitely large, but only a single value of the parameter, the zero corresponds to the **negative answer**. Fixing the parameter makes the distribution of the calculated statistical parameter unambiguous, thus one possible distribution (the Student's *t*-distribution) corresponds to the null hypothesis, and many distributions to the opposite statement. However, we should emphasize, that the calculated t_s value can be zero only theoretically. In reality, after performing all the calculations we usually get a nonzero t_s value. We have to make our decisions based on the relationship of the t_s value and the Student's t distribution asumed by the hypothesis.

We would have an easy job in decision making if the Student's *t* distribution spanned only a given range, let's say from t_{begin} to t_{end}. It would be enough to check whether the t_s value is in that interval or not. Inside the interval the null hypothesis would be accepted, and outside it would be rejected. However, the Student's *t*-distribution, just like the Gaussian distribution, spans from $-\infty$ to $+\infty$, therefore a finite range that enables us to make an unambiguous decision does not exist.

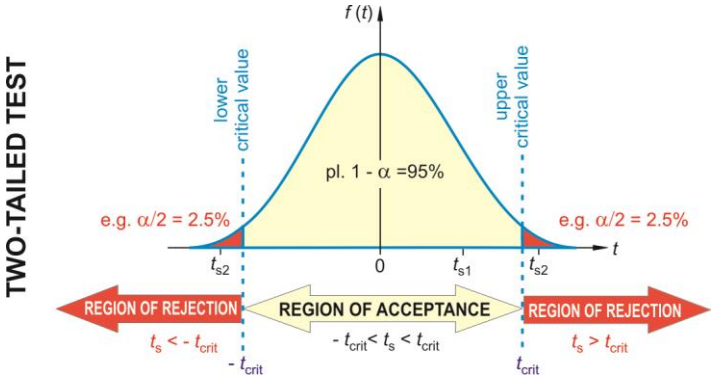


Fig. 22. *Regions of acceptance and rejection in case of a two-tailed t-test.*

Because we must define an interval in order to make a decision, let us cut off the "tails" of the Student's *t* distribution at values far from 0, starting at a critical value t_{crit}. (Exactly how we do this will be discussed later.) We may then pose the question whether the calculated **t_s value is within or outside this interval** (see Figs. 22 and 23). **If it falls within the interval** (e.g. t_{s1}), then **the null hypothesis is accepted. If it is beyond the boundaries** (e.g. t_{s2}), then **the null hypothesis is rejected**, and we say that **the calculated value of t_s is significantly different from 0**, or shortly just: it is **significant** (see Comment 11). The part that was cut off is called **region of rejection** and the remaining interval is the **region of acceptance**.

Comment 11.
The term "significant difference" never means absolute certainty, just as the term "not significant" does not mean that there is positively no difference. There may be real but very small differences, which are smaller than the error of the measurement, the experimental method or the equipment in use. Importantly, the significance analysis can never reveal the reason of the difference.

Regardless of what was our decision, acceptance or rejection of the null

significant
signifikant
szignifikáns

critical region, region of rejection
kritischer Bereich (Ablehnungsbereich)
kritikus tartomány

hypothesis, there is always a chance that our decision is not right.

The error of **rejecting a null hypothesis when it is really true** is known as **type I error**. In the judicial analogy it corresponds to the situation, when the verdict is "guilty", but in fact the accused is innocent. Type I error is made if the calculated t_s value in fact belongs to the Student's t distribution (around 0), but because the tails of the distribution were cut off it falls into the region of rejection and the null hypothesis is rejected. **The probability of type I error can be given exactly**, and it is equal to **the area that was cut off** (see Fig. 22, Fig. 23 and Fig. 9).

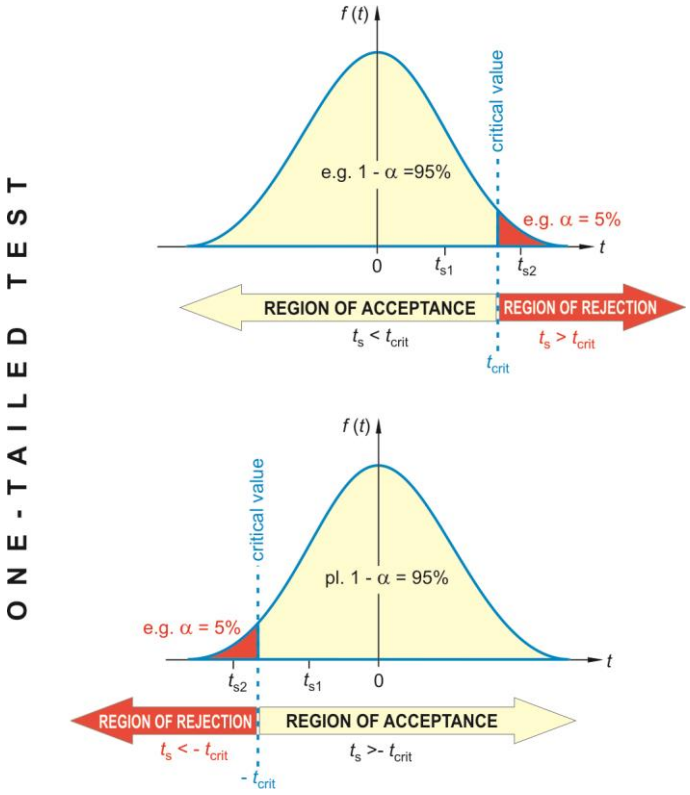


***Fig. 23**. Regions of acceptance and rejection in case of one-tailed t-tests.*

significance level
Signifikanzniveau
szignifikancia szint

The probability of the type I error is called **the probability level of the statistical decision** (p-value), or, rarely, it is called the level of error. In practice we choose explicitly in advance a probability α of the values that are cut off (which corresponds to the area) rather than the distance from zero. This probability is referred as the **level of significance**, because it indicates unambiguously which t values are significant and which are not.

acceptance region
Annahmenbereich
elfogadási tartomány

type I error
Fehler 1. Art
elsőfajú hiba

Another kind of error is made if **a false null hypothesis** (about which we do not really know whether it is false) **is accepted**. This error is called the **type II error**. In the judicial analogy type II error is made if a guilty accused is acquitted. This situation happens if the calculated t_s value "belongs to" an actual sampling distribution centered on a different expected value $t^* \neq 0$, and not to the hypothesized Student's t-distribution centered on 0, although we think (wrongly) that it "belongs to" the latter.

type II error
Fehler 2. Art
másodfajú hiba

The probability (p =) β of this type of error could in principle be measured by calculating the corresponding area of the t^* distribution of expected value. Since this distribution is not known, therefore **the probability of the type II error cannot be determined** (see Fig. 24). There are some methods for the estimation of this error, however.

It is important to emphasize that **α makes sense only when the hypothesis is rejected** and β when the hypothesis is accepted. At the same time it is obvious that by decreasing the probability of one type of error, that of the other increases. Because only the probability of the type I error can be fixed, a useful compromise should be made.

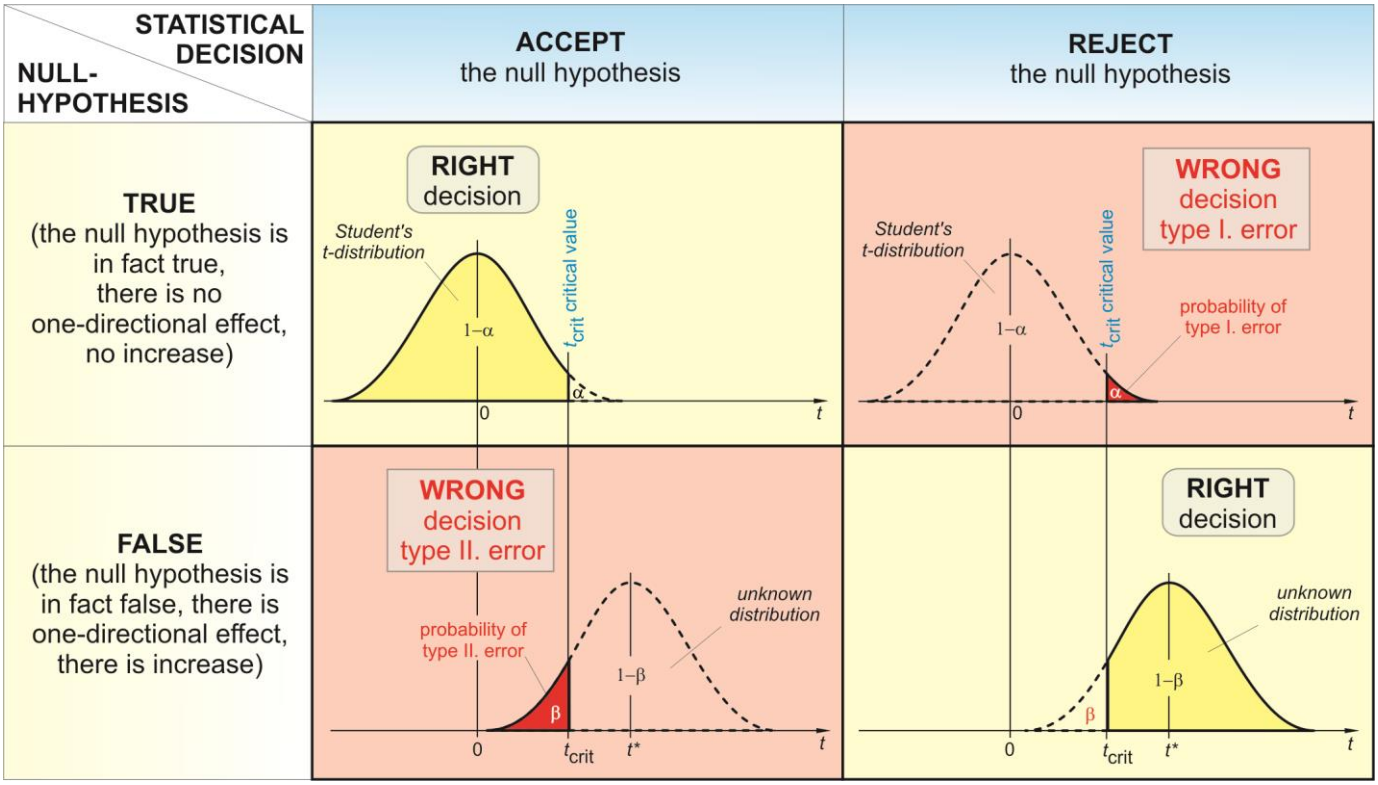


Fig. 24. *Type I and type II error.*

The usual level of significance in medical and biological studies is (p =) $\alpha = 0.05$ (that is, 5 %). It means that, on average, in five out of 100 cases the null hypothesis is rejected although in reality it is true. Often even the 5% probability of making wrong decisions is not allowed. In these cases the significance level can be lowered to $\alpha = 0.01$, $\alpha = 0.001$, or even further. Note, however the probability of making a type II error increases in the meantime.

Let us now see how to perform a *t*-test in practice. For this we need to know that although the Student's *t* distribution could be given by a complicated formula, it is simpler if its values are listed in tables (just like for the trigonometric functions). However, there is a substantial difference between the tables of the Student's *t* distribution and that of the sine function. In case of the sine function, for every x the $f(x) = \sin x$ value is given. The tables of the Student's t distribution have a special structure for an easier use (see Fig. 25 and Table 7).

First of all, the **Student's t-distribution table** contains **many distributions** corresponding to different degrees of freedom. **Degrees of freedom** are listed **in the left column**, and in every corresponding row there are data for the particular distribution. These values are not function values but *t* values; that is, the special values of the independent variable.

In order to explain the meaning of individual numbers of the table, let us compare the graph of the Student's t-distribution for 5 degrees of freedom with the distribution in the 5th row of the table. The table has two **header rows, with probabilities *p*** (for one- and two-tailed tests): these correspond to the area under the curve (in one and two tails) on the graph. The table gives the absolute value of *t*, at which we have to cut off one or two (symmetrical) tails of the distribution to make the cut area equal to the *p*-value in the header. In other words, the significance levels that we can choose in advance are listed in the headers of the table and the corresponding critical values t_{crit} can be found in the row of the given degrees of freedom.

To make an unambiguous decision the order of the steps of the method is very important. The **level of significance** of our future decision **is stated first** and the comparison with the calculated value is done afterwards (see Comment 12).

Comment 12.
Nowadays, in the world of computers the table of Student's t distribution is rarely used, because the computer can calculate the *p*-value for any critical value. (In most of the statistical programs the *t* value is not even calculated.) Hence the result of the test is a *p*-value, and the decision has to be made based on this. If *p* is small enough (smaller than the level of significance α stated in advance), then H_0 is rejected (and H_1 becomes valid), as the probability of rejecting a true hypothesis is small. (In other words, the probability that only the random sampling variation of data causes a value so different from 0 is small.) If *p* is large, then H_0 is accepted for the same reason. (It is our decision whether the *p*-value is large or small.)

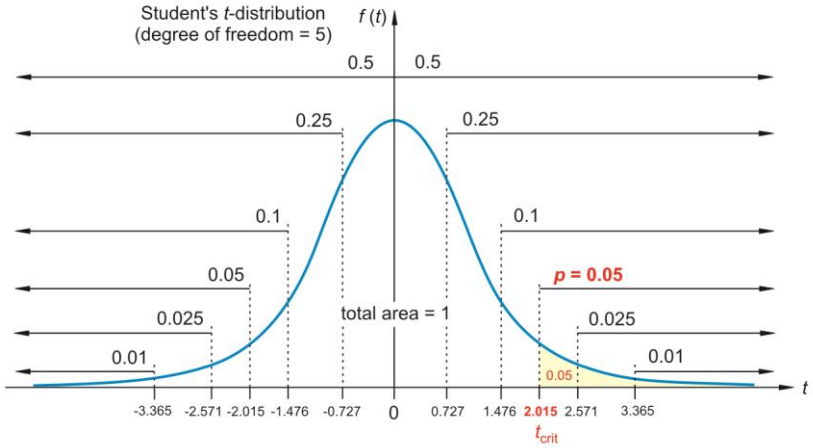


Fig. 25. *Student's t-distribution. Critical values and probabilities that correspond to the one-tailed t-test.*

Comment 13.
t-test for a single sample
(Detailed solution for a problem: DOES THE PULSE RATE CHANGE after holding ones breath for one minute?)

H_0: the pulse rate does not change,
that is $\bar{x} - \mu = 0$, where $\mu_0 = 0$ or $\bar{x} \cong 0$,
where $\bar{x}$ stands for the mean of the changes.
H_1: $\bar{x} - \mu_0 \neq 0$, (two-tailed test).
The first two columns of the table contain the measured pulse rate data before (x_b) and after (x_a) 1 minute of holding the breath, for $n = 6$ participants. The corresponding pulse rate differences (changes) and their squares are listed in the third and fourth columns (the squares are needed for the calculation of mean and the standard deviation).

x_b	x_a	$x'_i = x_a - x_b$	x'^2_i
69	71	2	4
60	63	3	9
68	70	2	4
75	76	1	1
71	70	-1	1
66	69	3	9
		$\Sigma x'_i = 10$	$\Sigma x'^2_i = 28$

Calculate the mean of the differences:

$$\bar{x}' = \frac{\Sigma(x_a - x_b)}{n} = \frac{10}{6} = 1.67$$

Find the empirical standard deviation of the differences from (4) and (6)

$$s' = \sqrt{\frac{\Sigma x'^2_i - \frac{(\Sigma x'_i)^2}{n}}{n-1}} = \sqrt{\frac{28 - \frac{10^2}{6}}{6-1}} = 1.51$$

Calculate the *t*-value of the sample using (20) and substituting $\mu = 0$:

$$t_s = \frac{\bar{x}'}{s_{\bar{x}'}} = \frac{\bar{x}'}{s'} \cdot \sqrt{n} = \frac{1.67}{1.51}\sqrt{6} = 2.72$$

Choose the level of significance as $\alpha = 0.05 \rightarrow 5\%$.
The degree of freedom: $(n-1) = (6-1) = 5$.
Find the critical t_s-value from the Table 7, in the $p = 0.05$ column of the two-tailed test and row that corresponds to the degrees of freedom of 5:
$t_{crit} = 2.571$.
Compare: $t_s = 2.72 > t_{crit} = 2.571$,
that means t_s falls in the region of rejection, thus the **null hypothesis is rejected**. The conclusion is that the one-minute holding of breath caused a **significant** pulse rate change (at a level of significance of 5 %).

	p (one-tailed test)								
	0.45	0.35	0.25	0.15	0.10	0.05	0.025	0.01	0.005
	p (two-tailed test)								
degree of freedom	0.90	0.70	0.50	0,30	0.20	0.10	0.05	0.02	0.01
1	0.158	0.510	1.000	1.963	3.078	6.314	12.706	31.821	63.657
2	0.142	0.445	0.816	1.386	1.886	2.920	4.303	6.965	9.925
3	0.137	0.424	0.765	1.250	1.638	2.35	3.182	4.541	5.841
4	0.134	0.414	0.741	1.190	1.533	2.132	2.776	3.74	4.604
5	0.132	0.408	0.727	1.156	1.476	2.015	2.571	3.365	4.032

Table 7. *Critical values and probabilities of the Student's t distribution for one- and two-tailed tests.*

As a next step, the calculated t_s value and the critical value t_{crit} obtained from the table are compared. **If $t_s \leq t_{crit}$ then the null hypothesis is accepted; if $t_s > t_{crit}$, then the null hypothesis is rejected**, and we say that the difference is significant at the given level of significance.

APPLICATION OF THE *t*-TEST, SOLUTION TO THE PROBLEM 1

We have already formulated two different forms of the null hypothesis:

1. *The further sales of the medication* ***should not be banned****, because the active ingredient content* ***does not differ*** *from the specified value.*
2. H_0: $\mu - \mu_0 = 0$, ($\bar{x} - \mu_0 \cong 0$).

The alternative hypothesis was formulated as H_1: $\mu - \mu_0 \neq 0$. This is a non-directional alternative hypothesis, thus we will use the two-tailed *t*-test. Next steps are:

- Calculate the *t*-value from the sample: $t_s = 1.28$
- Select the level of significance: $\alpha = 0.05$
- Determine the degree of freedom: $n-1 = 5$
- Identify the critical value from the table: $t_{crit} = 2.571$
- As $1.28 < 2.571$ (thus $t_s < t_{crit}$), we accept the null hypothesis.
- Conclusion: *The further sales of the medication* ***should not be banned****, because the active ingredient content* ***does not differ*** *from the specified value.*

More precisely it means that **our data do not provide enough evidence** (the parameter t_s, measuring the strength of the evidences is not high enough) **for the rejection of the null hypothesis** and therefore for the *banning of the sales*.

Further comments: As we accepted the null hypothesis, we would rather be interested in type II error, the value of β, the probability, that **we accepted a false hypothesis**. Earlier we said that it is not possible to determine the β. The only thing we can do is to increase the value of α (which implies the decrease of β) until $t_s = t_{crit}$. Now, if we choose only a bit larger critical value, the null hypothesis will still be accepted, but the corresponding α is much greater than 0.05, that is p = 0.26. As we accepted the null hypothesis, it does not make much sense by itself, but it means that the β, probability of the Type II error decreased. Hence in such cases it is reasonable to give this α value.

Conditions for applying the *t*-test are:

1. the variable is normally distributed,
2. the elements of the sample are independent,
3. in case of two samples their standard deviations are similar "enough".

The first and third conditions are not very strict, but the second is. The *t*-test is used mostly to test a mean and difference of means, but it can be applied for testing the correlation coefficient as well.

Now let us see the most common variations of the *t*-tests, and the corresponding formulas for calculating t_s values.

t-TEST FOR A SINGLE SAMPLE

We have formulated earlier a question as: DOES THE PULSE RATE CHANGE after holding one's breath for one minute (see Comment 13)?

Formulated in general: **Does the expected value of the population change** as a result of an intervention? Or, in other words: Does the intervention have any effect? Formally: **Does the expected value of the population distribution differ from a previously given value?**

Calculation of the t_s:
$$t_s = \frac{\bar{x} - \mu_0}{s_{\bar{x}}} = \frac{\bar{x} - \mu_0}{s}\sqrt{n} \qquad (20)$$

Degrees of freedom: n–1

t-TEST FOR TWO SAMPLES

We have formulated **earlier a question** as: IS THERE A DIFFERENCE between the pulse rates of girls and boys (see Comment 14)?

Formulated **in general: Do the expected values of two independent populations differ**?

Calculation of the t_s: $t_s = \frac{\bar{x}_1 - \bar{x}_2}{s^*} \cdot \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$, where $s^* = \sqrt{\frac{Q_1 + Q_2}{n_1 + n_2 - 2}}$ (21)

Degrees of freedom: $n_1 + n_2 - 2$,

where Q is equivalent to the notation in formula (6) calculated for the first and second samples. (Note that this expression is very similar to equation (20).)

t-TEST FOR CORRELATION

We have formulated **earlier a question as: DOES the accommodation power of the eye DEPEND on the age of the person**?

Formulated **in general**: taking into account the correlation coefficient and the number of data, **can we say about the two quantities that they depend on each other** (based on their changes)**?**

Calculation of t_s:
$$t_s = r \cdot \sqrt{\frac{n-2}{1-r^2}}, \qquad (22)$$

Degrees of freedom: n –2,
where n is the number of measured data (x, y pairs), r is the correlation coefficient

Comment 14.
t-test for two samples (Detailed solution for a problem: IS THERE A DIFFERENCE between the pulse rate of girls and boys?)

- H_0: there is no difference between the pulse rate of girls and boys, that is $\mu_{girls} - \mu_{boys} = 0$.
- H_1: $\mu_{girls} - \mu_{boys} \neq 0$, (two-tailed test).

The table below contains measured pulse rate data of 6 girls (x_{girls}) and 9 boys (x_{boys}).

x_{girls}	x_{boys}
74	71
87	63
62	70
79	74
71	71
77	69
	82
	56
	78
$\bar{x}_{girls} = 75$	$\bar{x}_{boys} = 70$

From the calculated means it seems that girls have higher pulse rates. Is this difference significant or just a result of a random sampling?
Let us suppose that every condition of the *t*-test is fulfilled (even the one about identical standard deviations, which can be checked by calculation).
We calculate by computer the *p*-value for our data, which is p = 0.296. (Parameters of the *t*-test function of the program must be set to two-tailed test, and equal standard deviations are taken into account.)
The *p*-value is rather large, much larger than 0.05, the usual level of significance, hence the null hypothesis cannot be rejected. If we rejected the null hypothesis, then the probability making a type I error, that is, of rejecting a true hypothesis would be almost 30 %. Therefore, we accept the null hypothesis H_0, and our conclusion is that the difference between the two samples is not significant. Thus, there is no significant difference between the pulse rate of girls and boys.

introduced by formula (19). The steps of the further part of the process are the same as in the previous examples (see Comments 13 and 14).

The discussed variations of the *t*-test are summarized in the Table 8. Which form of the test to be used is determined by the initial question. Knowing the experiment and the available data help us to pose the question unambiguously.

	t-test for a single sample	*t*-test for two samples	*t*-test for correlation
a typical question in the field of medicine	Is the treatment effective? (Is there a change in the supposed direction?)	Is there a difference between the effects of the two treatments?	Is there a correlation between two quantities?
the corresponding null hypothesis	The treatment is not effective.	The two treatments have the same effect.	There is no correlation.
the question in general form	Does the sample belong to a distribution of μ_0 expected value?	Do the two samples belong to the same distribution?	Is there a correlation between the two (continuous) variables, even if r is small?
the exact form of the null hypothesis	$\mu - \mu_0 = 0$	$\mu_1 - \mu_2 = 0$	There is no correlation between the two variables.
the experiment	One physical quantity is measured on one sample.	The same physical quantity is measured on two independent samples.	Two physical quantities are measured on the same sample.
t	$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}$	see (21)	$t = r \cdot \sqrt{\frac{n-2}{1-r^2}}$
degree of freedom	$n - 1$	$n_1 + n_2 - 2$	$n - 2$

Table 8. *Summary of the most frequently used t-tests.*

Comment 15.
Chi square (χ^2) test for two samples
Detailed solution for a problem: Is the frequency of pulmonary cancer higher among smoker than among non-smoker patients?

• H_0: The frequency of pulmonary cancer is the same among smoker and among non-smoker patients, thus $\chi^2 \approx 0$.
• H_1: frequency of pulmonary cancer is different among smoker and among non-smoker patients, thus $\chi^2 \neq 0$.
The following table summarizes a study done in the Pulmonology Clinic. The frequencies of pulmonary cancer cases in the two examined groups (and the subtotals, n = 61) are shown.

	Pulmonary cancer	No cancer	
Smoker	14	13	27
Non-smoker	9	25	34
	23	38	61

As 23·27 = 621 > 5·61 = 305, the test can be performed.
From the formula (23) we get the χ_s^2-value:

$$\chi^2{}_m = \frac{61 \cdot (14 \cdot 25 - 9 \cdot 13)^2}{23 \cdot 38 \cdot 34 \cdot 27} = 4.13$$

We can see that $\chi_s^2 \neq 0$, but is this a significant difference, or just a result of random sampling?
Let us choose the level of significance
$\alpha = 0.05 \rightarrow 5\ \%$.
the degree of freedom is :1.
Find the critical value in the p = 0.05 column and first (degree of freedom 1) row of the χ^2-distribution table (see Fig. 26.):
$\chi_{crit}^2 = 3.84$
Because: $\chi_s^2 = 4.13 > \chi_{crit}^2 = 3.84$,
the χ_s^2 value falls in the region of rejection, thus our decision is, that **the null hypothesis is rejected**. The conclusion is that the observed difference between the occurrence of pulmonary cancer among smokers and non-smokers is significant (at 5 % level of significance).

χ^2-TESTS (CHI SQUARE TESTS)

Naturally, *t*-tests can be applied only for numerical (continuous) variables. What shall we do with categorical data? In these cases we make inferences based on the **frequencies of data falling into categories**. For this we will use the so called Chi square (χ^2) tests.

The question is: is the frequency of occurrence of an attribute (symptom) different for two different populations? For example, is the frequency of pulmonary cancer higher among smoker than among non-smoker patients having lung diseases (see Comment 15)?

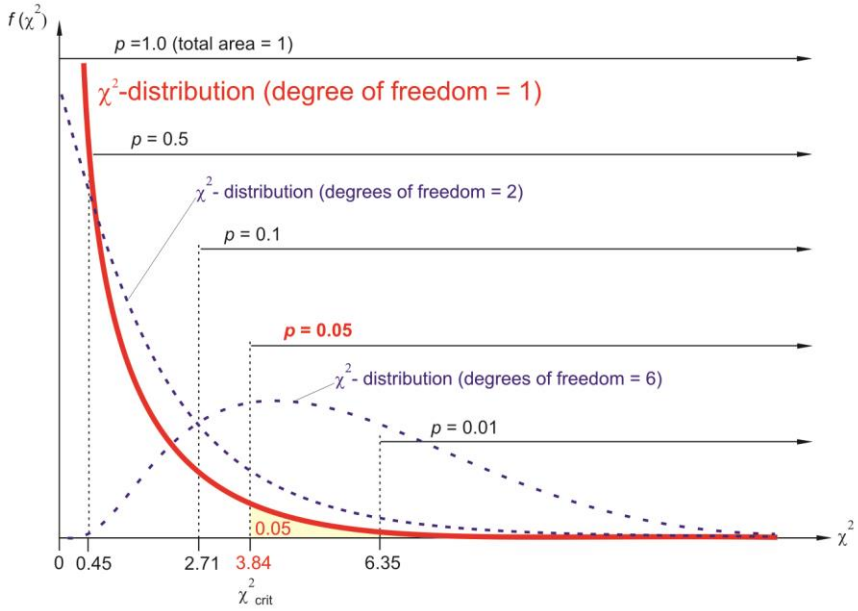


Fig. 26. *The χ^2-distribution for different degrees of freedom, and some critical values of the χ^2-distribution for 1 degree of freedom.*

Measured data are organized in the form of a table, where the two populations are indicated as *A* and *B*. Total number of *n* people was examined. The number of

those who have the examined attribute is a and c, and those who do not is b and d.

The only condition of this test is that the observed frequencies are high enough, or in other words the product of two smallest subtotals has to be greater than $5n$. If this condition is not fulfilled, then the so called Fisher's exact test can be applied (not discussed).

	The examined attribute		
	is present	is not present	total
group A	***a***	***b***	$a+b$
group B	***c***	***d***	$c+d$
total	$a+c$	$b+d$	n

Table 9.

Null hypothesis: There is no difference in the frequency of the examined attribute in the two groups, it is the same in the two samples, thus the χ_s^2-value calculated by the formula (23) is zero, or not significantly different from zero.

Calculation of the χ_s^2-value:

$$\chi_s^2 = \frac{n\cdot(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}. \qquad (23)$$

Degree of freedom: 1.

From the same ratio of frequencies of the examined attribute, thus from $\frac{a}{b}=\frac{c}{d}$ follows, that $ad - bc = 0$, thus the $\chi^2 = 0$ fulfils the null hypothesis. Whether or not the calculated χ_s^2-value is significant, we decide based on the χ^2-distribution table for degree of freedom 1.

χ^2-TESTS (CHI - SQUARE TESTS) IN GENERAL

In the previous section we discussed only the simplest but often applicable version of the χ^2-tests (2 x 2 table). Both the number of populations and the examined attributes may be larger.
In general: we have an (l x m) table, called the **contingency table**, where l is the number of rows and m is the number of columns. It can be used for **independence tests**. In case of independent variables, the relative frequencies are identical.

***H*₀:** independence of the row and column variables, expected value of χ^2 is 0.

Calculation of the value of χ^2:

$$\chi^2 = \sum\left[\frac{(O-E)^2}{E}\right], \qquad (24)$$

where O is the **observed** frequency, E is the **expected** frequency, and the degree of freedom is (l-1)(m-1).

Calculation of the expected frequencies. Besides the contingency table of the observed frequencies, a new contingency table has to be calculated containing the frequencies expected if H_0 is true.
The corresponding subtotals of rows and columns and grand total are equal in the two tables. If H_0 is true, then the frequency expected in any particular cell of the calculated contingency table is equal to the product of the subtotal of the given row with subtotal of the given column divided by the grand total of the sample (rowsubtotal·columnsubtotal/grandtotal) (see Comment 16).

Comment 16.
(The general solution for the problem in Comment 15.)
Question: Is the frequency of pulmonary cancer different among smoker than among nonsmoker patients?

Let us formulate again the statistical hypotheses.
***H*₀**: The frequency of pulmonary cancer is the same among smoker and nonsmoker patients, thus $\chi^2 = 0$.
***H*₁**: The frequency of pulmonary cancer is different among smoker and nonsmoker patients, thus $\chi^2 \neq 0$.

Using a proper statistical software we can easily get the solution. The contingency table of **observed** frequencies was already constructed in the previous comment.

Observed contingency table:

	Pulmonary cancer	No cancer	
smoker	14	13	27
nonsmoker	9	25	34
	23	38	61

After this, we construct the contingency table of the **expected** frequencies under the condition H_0, where the corresponding subtotals of rows and columns and grand total are equal in the two tables. Relative frequencies have to be equal in the expected and observed tables, thus

$$\frac{a}{b}=\frac{c}{d}$$

Expected contingency table:

	Pulmonary cancer	No cancer	
smoker	10.18	16.82	27
nonsmoker	12.82	21.18	34
	23	38	61

The χ^2-test is included in most modern statistical packages. By using such a software one can select the chi-square function and assign the two tables as arguments of the function. The computer calculates the probability p based on the data, in this case $p = 0.042$.
As p value is smaller that the usual significance level (0.05), **the null hypothesis is rejected**. The alternative hypothesis ***H*₁** is accepted as the conclusion, which means that the frequency of pulmonary cancer is different among smoker and among nonsmoker patients at significance level smaller than 5 %.

EXAMPLES FROM THE FIELD OF MEDICAL STATISTICS

It occurs often in medical practice that the studied variable is a continuous quantity, but the conditions of applicability of the *t*-tests are violated (for example, the variable does not have normal distribution, see page 27). In such a situation we can usually apply a *t*-test, but our conclusions will be less valid.

We have already discussed one possible solution to the problem, the transformation of the variable. If this is not possible, then the so-called **non-parametric methods** can be used. (The name indicates that if the type of the distribution of the variable is unknown, the parameters of the distribution are also unknown.) The most widely used non-parametric methods assign ranks to the measured scores according to defined rules. The test then uses the rank numbers instead of the original values. **Ranks** are usually integer numbers that correspond to the place of the original value in the measured series. Sometimes it occurs even in the case of continuous variables that the measurement contains identical scores which lead to ties in rank. A simple and satisfactory way to deal with ties is to assign each of the tied scores the mean of the ranks that would be available to them. For example, if the first two scores are the same, then instead of ranks 1; 2; 3; … we will assign ranks 1.5; 1.5; 3; ...

Non-parametric tests do not specify restrictive conditions on the variables. The only requirement is that the variable is ordinal. One could ask then why not apply these methods in all cases then? It is possible to show that for the same dataset, if applicable, the *t*-tests allow to make decisions at a better significance level. This is why first it should always be determined whether or not a *t*-test can be performed.

MANN-WHITNEY *U*-TEST

The Mann-Whitney *U*-test is an alternative to the two-sample *t*-test. Let us, for example, examine the question **whether aspirin is an effective pain killer for headaches?** We must emphasize that this question can be answered only after an extensive clinical investigation followed by detailed statistical analysis. Here we show, as an example, only one step of this procedure.

H_0: The aspirin does not effectively reduce headache.

The strength of a headache cannot be measured directly, therefore we make the following experiment. Let us sort people who suffer from headache into two groups. The individuals in the first group (n_1) get aspirin. Those in the second group (n_2), the control, receive so-called placebo, a drug without any effective ingredient. Of course the participants do not know which group they belong to; they all believe they received aspirin. After some time they are asked to scale from 1 to 10 the effectiveness of the pain killer so that 0 refers no effect (headache same as before), and 10 means that the drug was fully effective (headache is gone completely). The statistical analysis is based on these data. This dataset corresponds to categories of an ordinal variable (better, little better, etc.), that were quantified for easier calculation, thus the distribution is unknown, and the parameters of the distribution are also unknown. For this reason a non-parametric test is used instead of the *t*-test.

All measured scores from the two groups are combined and ranked from lowest to highest scores, regardless which group they originate from. If the null hypothesis is true, then the scores from the two groups are randomly placed. The sum of the ranks is calculated in the first and second group as T_1 and T_2, respectively. The sum of all ranks (that is sum of integers from 1 to n_1+n_2) can be calculated as $(n_1+n_2+1)\cdot(n_1+n_2)/2$, that equals obviously T_1+T_2. The "average" rank is therefore $(n_1+n_2+1)/2$. Assuming that null hypothesis is true, $T_1 \approx n_1\cdot(n_1+n_2+1)/2$ and $T_2 \approx n_2\cdot(n_1+n_2+1)/2$, or in other form: $T_1 - n_1\cdot(n_1+n_2+1)/2 = 0$.

From this a new variable can be constructed as

$$z = \frac{T_1 - n_1\cdot(n_1+n_2+1)/2}{\sqrt{n_1\cdot n_2\cdot(n_1+n_2+1)/12}}$$

which corresponds to $N(0,1)$, the standard normal distribution (it has to be mentioned that for a small number of data this approximation is not valid). Thus, if the null hypothesis is true, $z = 0$. The deviation of z from 0 is caused only by

Comment 17.
Mann-Whitney *U*-test (Detailed solution)
Question: Is aspirin effective as a pain killer for headaches?)

H_0: The aspirin is not effective in reducing headache, $z = 0$.
H_1: The aspirin is effective in reducing headache, $z > 0$ (one-tailed test).

1st group ($n_1 = 8$):

7.5	8.3	9.1	6.2	5.4	8.3	6.5	8.4

2nd group ($n_2 = 9$):

3.1	5.6	4.5	6.2	5.1	5.3	5.5	4.1	4.3

Assign ranks to all scores!

Subject Score		Rank
1	3.1	1
2	4.1	2
3	4.3	3
4	4.5	4
5	5.1	5
6	5.3	6
7	5.4	7
8	5.5	8
9	5.6	9
10	6.2	10.5
11	6.2	10.5
12	6.5	12
13	7.5	13
14	8.3	14.5
15	8.3	14.5
16	8.4	16
17	9.1	17

$T = 7+10.5+12+13+14.5+14.5+16+17 = 104.5$

$$z = \frac{104.5 - 8\cdot(8+9+1)/2}{\sqrt{8\cdot 9\cdot(8+9+1)/12}} = 3.13$$

Let the level of significance be very small, $\alpha = 0.01 \rightarrow 1\ \%$.
Instead of standard normal distribution $N(0,1)$ one can use *t*-distribution of the degree of freedom ∞ (the two are equivalent).
Find the critical value z_{crit} in the table of one-tailed *t*-test at the row corresponding to the degree of freedom ∞, and column of $p = 0.01$

$z_{crit} = 2.33$.

Compare $z_m = 3.13 > z_{crit} = 2.33$,

which means that z_m falls in the region of rejection, thus **the null hypothesis is rejected** (at level of significance of 1 %). (The probability that corresponds to the value z_m of the sample is $p < 0.0009$, therefore the probability that the H_0 is true and z differs from 0 only by chance is very small.)

Based on the above investigation we conclude that aspirin is effective in reducing headache.

chance if H_0 is true, or by the fact that the original null hypothesis is not true. Decision is made by comparing the calculated z value with the critical value from the standard normal distribution $N(0,1)$ at the given p probability level (see comment 17.).

SPEARMAN-CORRELATION

Spearman-correlation (r_s) is the "non-parametric" equivalent of the correlation coefficient (r, also called Pearson coefficient of correlation, see page 17.), which is used for calculating degree of relationship between two ordinal variables.

First, scores are ranked, the sum of the squares of differences of corresponding ranks (d_i) is calculated filled with the following formula:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \qquad (25)$$

where n is the number of corresponding pairs of scores (see comment 18. on page 33).

The Spearman-correlation coefficient calculated this way takes values between +1 and –1 just like the Pearson correlation coefficient. Values close to zero indicate very little association (or no association at all), and values close to one (plus or minus) indicate high degree of association (positive or negative correlation, respectively). In the latter case variables are "linked together" in some way.

OTHER APPLICATIONS OF 2x2 TABLES

Besides the χ^2-tests, 2 x 2 tables occur in other statistical tests as well. It is often of interest whether different risk factors are involved in the development of certain diseases. To answer such questions, frequency data are collected from two groups, and the results are presented in the form of a 2 x 2 table.

Risk factor	Disease status	
	sick	healthy
present	*a*	*b*
absent	*c*	*d*

Table 10.

Letters *a*, *b*, *c*, *d* mean the frequencies of the corresponding category. If the development of the illness is independent of the examined risk factor, then we expect that the ratios of the present *versus* absent frequencies are the same for the healthy and the ill group. Depending on the circumstances there are two basic ways of acquiring and analyzing data: prospective (cohort) study and retrospective study (case-control study).

PROSPECTIVE, COHORT STUDY

In the **prospective, or cohort study**, a group of healthy individuals is divided in two groups based on the presence of a certain risk factor. These groups are followed (observed) for a long time and at the end the contingency table is constructed from the collected data.

An example of a prospective, cohort study:

Question: Is smoking associated with a risk for heart attack (acute myocardial infarction)?

***H*₀**: There is no connection between smoking and heart attack.

Common characteristic (risk factor): smoking / non-smoking

Comment 18.
Spearman-correlation
(Detailed solution)
Question: Is the evaluation of two doctors similar?

Two medical doctors make aptitude tests, and the question is how similar is their evaluation. Both of them examined a group of seven employees and evaluated them on a scale of 1 to 10 (1 means the worst and 10 is the best).

Results are summarized in the following table.

Per-son	Doctor A	Doctor B	Rank A	Rank B	d^2
A	10	9	7	6.5	0.25
B	7	9	6	6.5	0.25
C	1	3	1	2	1
D	4	6	3.5	3	0.25
E	4	7	3.5	4	0.25
F	3	1	2	1	1
G	5	8	5	5	0
				Total:	3

If we calculate the Pearson correlation coefficient r from the first two columns using the formula (19) we get $r = 0.78$.

Calculating the Spearman-correlation (r_s) was from the formula (25)

$$r_s = 1 - (6\cdot3)/(7(7^2 - 1)) = 0.95.$$

In this case we get $r_s = 0.95$.

We get two different numbers for correlation coefficient, but both indicate the same situation: the two physicians judge differently, but if one of them (A) considers an employee more qualified, the second doctor (B) classifies the same employee more qualified as well.

It is important to mention that the ranks actually contain **less information** than original data, but they focus on the part of the information which is **important** for the decision making.

A cohort is a group of subjects who share a common characteristic (e.g., smoking). Cohorts may be tracked over extended periods of time (e.g., 10 years), and the frequency of heart attack is counted both in the cohort and the control group. Data are summarized as usual, in a 2x2 table.

	heart attack		
Risk factor	present	absent	total:
smoker	a	b	$a+b$
nonsmoker	c	d	$c+d$
total:	$a+c$	$b+d$	$n=a+b+c+d$

Table 11.

The **risk** (probability of the condition) is calculated in the two groups separately:

- The risk in the smoker (exposed) group is: $a/(a+b)$.
- The risk in the non-smoker group is: $c/(c+d)$.

Relative risk (*RR*) is the ratio of the probability of the event (heart attack) occurring in the exposed group (smokers) versus the non-exposed group:

$$RR = \frac{a/(a+b)}{c/(c+d)} = \frac{a\cdot(c+d)}{c\cdot(a+b)} \tag{26}$$

In other words, smokers would be *RR* times as likely as non-smokers to develop heart attack. If the null hypothesis H_0 is true, then the expected value for *RR* is 1, which means that smokers have the same risk of developing heart attack as non-smokers.

In the next step we calculate the standard error (*SE*) of the logarithm of the relative risk (ln(*RR*)):

$$SE(\ln RR) = \sqrt{\frac{1-a/(a+b)}{a} + \frac{1-c/(c+d)}{c}} \tag{27}$$

The reliability of the relative risk (*RR*) value can be characterized by calculating the confidence interval of the *RR* value, that is the interval of 95% probability (±1.96 times, or roughly two times the standard error). Notice that *SE* of the logarithm of *RR* is calculated above, thus the confidence interval, is not symmetric for *RR*. If the confidence interval contains the value 1, then the null hypothesis is accepted, otherwise it is rejected at the given confidence level.

A RETROSPECTIVE, CASE-CONTROL STUDY

In the second case, people with a disease (often, a specific diagnosis) are matched with people who do not have the disease (the controls) and the groups are compared to find out if other characteristics (risk factors) are also different between the two groups. This retrospective observational study is called a **case-control study**.

An example of the retrospective study:

Question: Does the risk factor play role in the development of a disease?

***H*$_0$**: There is no connection between the presence of the risk factor and the development of the disease.

Common characteristic: disease (case and control)

Data are summarized in the following table:

	Disease status		
Risk factor	yes (case)	no (control)	Total:
present	a	b	$a+b$
absent	c	d	$c+d$
total:	$a+c$	$b+d$	$n=a+b+c+d$

Table 12.

First the odds of an event (disease) is determined in both groups:

- The odds of the disease in the presence of the risk factor is: a / b.
- The odds of the disease in the absence of the risk factor is: c / d.

The **odds ratio (*OR*)** is the ratio of the odds of the disease occurring in presence of the risk factor to the odds of it occurring in the absence of the risk factor.

$$OR = \frac{a/b}{c/d} = \frac{a \cdot d}{b \cdot c} \tag{28}$$

Similarly to the previous case, the standard error (SE) of the logarithm of the odds ratio (ln(OR)) can be calculated:

$$SE(\ln OR) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \tag{29}$$

Finally, the corresponding confidence interval can be calculated. If it contains the value 1, then the presence of the risk factor does not increase the development of the disease (at the given confidence level).

Comparison of the prospective, cohort and retrospective, case-control studies:
Cohort studies take a long follow-up time to generate useful data and therefore are expensive to conduct. Because of the long follow-up time the number of cases must be relatively high (it is sensitive to attrition, loss of participants). Nevertheless, the results that are obtained from long-term cohort studies are of substantially superior quality to retrospective studies, especially in cases when the occurrence of the risk factor is very low.

Note that these data can be processed with χ^2-tests as well. Why do we still use prospective, cohort and retrospective, case-control study? The advantage of these tests compared to the χ^2-test is that the decision is based on a numerical value that indicates the strength of the risk factor (RR, OR).

STATISTICAL CHARACTERIZATION OF DIAGNOSTIC TESTS

Investigation of the reliability, accuracy, and further characterization of diagnostic tests is an important field of medical statistics. Any positive or negative result of a diagnostic test can be true or false. A therapy based on a false test result can have unpredictable and serious consequences. It is very important to know the reliability of diagnostic test results. Of course the physician's diagnosis is based on several different diagnostic tests, but one has to know what is their value in the diagnosis of the given disorder.

In the following, instead of a detailed discussion of the field we will emphasize the most important characteristics of diagnostic tests.

Question: How reliable is the result of the test?
Let us **perform the test** and **compare** the result with "**reality**".
There are four basic outcomes in this comparison:
true positive (TP): sick people correctly diagnosed as sick,
false positive (FP): healthy people incorrectly identified as sick,
true negative (TN): healthy people correctly identified as healthy,
false negative (FN): sick people incorrectly identified as healthy.
These are summarized in the following table.

Real condition	Test outcome		total:
	negative	positive	
healthy	**TN** (true negative)	**FP** (false positive)	TN +FP
sick	**FN** (false negative)	**TP** (true positive)	FN + TP
total	TN + FN	FP + TP	n = TN + FN + FP + TP

Table 13.

In such a table the "reality"is always questionable. How could we know the truth, the correct values? Sometimes there are very straightforward cases, when the diagnosis is known definitely. Another possibility is to use so called "gold standard" tests which are extremely reliable.

Definition of some statistical characteristics:

- **Prevalence** (occurrence of a disease): relative number of cases of the disease: (TP + FN)/*n*.
- **Sensitivity** measures the proportion of actual positives which are correctly identified as such (e.g. the percentage of sick people who are identified as having the condition): TP/(TP+FN).
- **Specificity** measures the proportion of negatives which are correctly identified as such (e.g. the percentage of healthy people who are identified as not having the condition): TN/(TN+FP)

These characteristics are usually calculated as percentages. For the perfect test, sensitivity and specificity would be 100 %, for reliable tests they are close to this.

In practice, when evaluating the diagnostic tests, the following considerations are made:

- If the therapy is easy to accomplish and does not involve any serious risk for the patient, then high sensitivity is the goal. (Every suspicious case is identified as positive.)
- Otherwise high specificity is usually more important. (Healthy individuals should not be cured!)

Although sensitivity and specificity are important characteristics of a diagnostic test, the predictive value is an even more important parameter of the test. The predictive value indicates the probability of the presence of the disease after a positive test result, or similarly, the probability of the absence of the examined disease after a negative result.

- **Predictive value positive (PVP,** post test probability of the disease): probability of the true positive result, if the test was positive, TP/(TP+FP)

- **Predictive value negative (PVN,** post test probability of the lack of disease): probability of the true positive result, if the test was positive, TN/(TN+FN)

All characteristics depend on the frequency of the occurrence of these diseases. t a high frequencyof occurrence the PVP is much higher than in rare diseases. For PVN the situation is opposite, it is high for rare diseases.

The discussed parameters characterizing the tests are summarized in Table 14.

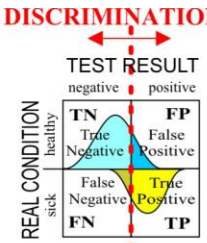
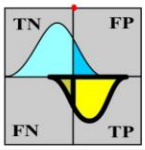
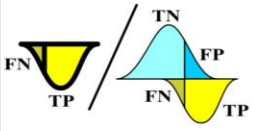
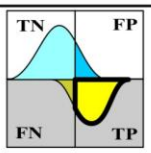
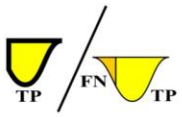
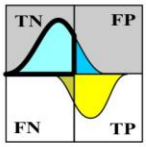
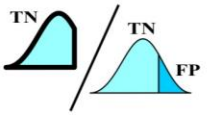
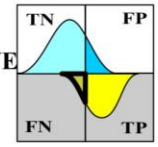
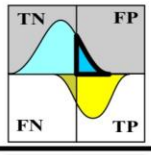
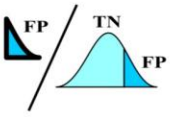
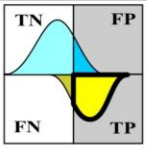
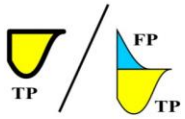
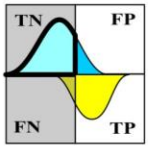
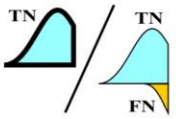
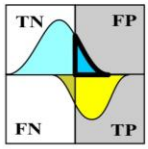
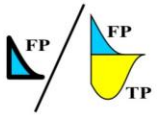
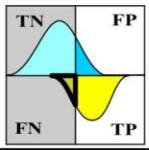
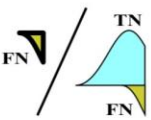
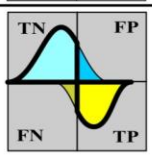
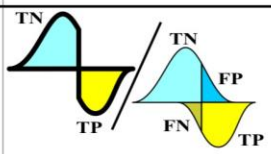
			ALTERNATIVE ENGLISH TERMS
		**DISCRIMINATION VALUE**	cut-off point
PREVALENCE (w)		 **= (TP + FN) / (TP + TN +FN + FP)** w=(de-sp)/(se-sp)	pretest probability occurrence of a disease
SENSITIVITY (se)		 **= TP / (TP + FN)** se	true-positive rate, true positive fraction, recall rate
SPECIFICITY (sp)		 **= TN / (TN + FP)** sp	true-negative rate, true negative fraction
FALSE-NEGATIVE RATE		 **= FN / (TP + FN)** (1-se)	false-negative fraction, type II error
FALSE-POSITIVE RATE		 **= FP / (TN + FP)** (1-sp)	false-positive fraction, type I error
PREDICTIVE VALUE POSITIVE (PVP)		 **= TP / (FP + TP)** PVP= =se*w / [se*w+(1-sp)*(1-w)]	positive predictive accuracy, posttest probability of disease
PREDICTIVE VALUE NEGATIVE (PVN)		 **= TN / (TN + FN)** PVN= =sp*(1-w) / [sp*(1-w)+(1-se)*w]	negative predictive accuracy, posttest probability of the lack of disease
FALSE ALARM RATE		 **= FP / (TP + FP)** (1-PVP)	
FALSE REASSURANCE RATE		 **= FN / (TN + FN)** (1-PVN)	
DIAGNOSTIC EFFECTIVITY (de)		 **= (TP + TN) / (TP + TN +FN + FP)** de=se*w+sp*(1-w)	
DIAGNOSTIC GAIN		(se+sp) / 2	

Table 14.

Comment 19.

Topics that we discussed here constitute only a small part of statistics. More detailed knowledge can be obtained from the literature listed below.

Statistical tables, which are needed for the statistical tests can be found in the 30. APPENDIX of this manual.

It often happens that data collection and the derived conclusion are falsified or "face-lifted". Therefore, always pay attention to the truthfulness, transparency and verifiability of the data and the statistics derived from them. The former British prime minister was probably misinformed several times taking advantage of difficult verifiability of statistics that finally made him to say ironically:

"The only statistics you can trust are those you falsified yourself."
Sir Winston Churchill

Literature:

***Mendenhall, W.*: Introduction to probability and statistics, Duxbury Press, Boston 1987**

***Norman T. J. Bailey*: Statistical methods in biology, Cambridge university press 1993**

***P. Armitage, G. Berry*: Statistical methods in medical research, Oxford 1994**

***Beth Dawson, Robert G. Trapp*: Basic and Clinical Biostatistics, Lange Medical Books/McGraw-Hill 2001**

SUMMARY (see Comment 19)

THE STEPS OF HYPOTHESIS TESTS

- ✓ **Pose the question.**
- ✓ **Choose** the proper **test**.
- ✓ Formulate the **initial statement**: null hypothesis (H_0, usually a characteristic value is zero) and the alternative hypothesis (H_1).
- ✓ **Theory**: knowing the population the question can be answered exactly (zero or non-zero).
- ✓ **Practice**: in case of a finite sample the value can be non-zero as a result of random sampling.

presumptions	
H_0 is true (the difference is caused by random sampling)	H_0 is not true

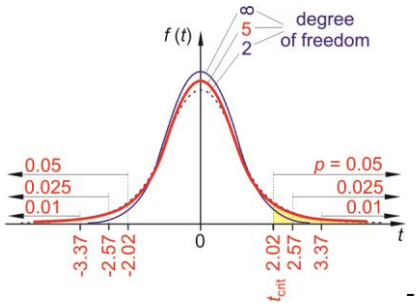
- ✓ **Sample**: a random sample, taken into account special medical requirements (e.g., having certain disease, any types of disqualification).
- ✓ **Choose** the properly low **level of significance** α (usually $\alpha \leq 0.05$).
- ✓ **Calculate the proper characteristic value** (t, χ^2, . . .) from the sample. The theoretical distribution of the value gives the probability of random differences.
- ✓ **Find** the theoretical **critical value** from the table that corresponds to the level of significance α.
- ✓ **Decision**: (in case of positive one-tailed test)
  1. theoretical critical value ≥ calculated value → H_0 is accepted, no reason to reject
  2. theoretical critical value < calculated value → H_0 is rejected, the probability that the difference is due only to random sampling is small.

The level of significance of the conclusion: α that in the 2nd case means the probability of the wrong decision.

1. STATISTICAL TABLES

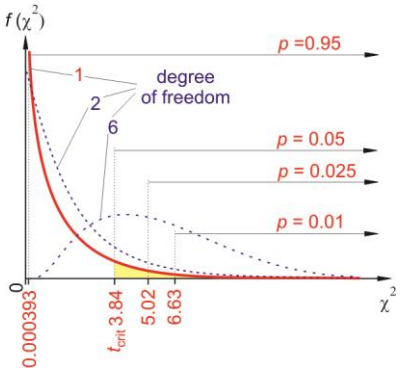
STUDENT'S *t*-DISTRIBUTION

Degree of freedom	*p* (probability, one-tailed test)						
	0.4	**0.25**	**0.1**	**0.05**	**0.025**	**0.01**	**0.005**
	p (probability, two-tailed test)						
	0.8	**0.5**	**0.2**	**0.1**	**0.05**	**0.02**	**0.01**
1	0.325	1.000	3.078	6.314	12.70	31.82	63.65
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704
60	0.255	0.679	1.296	1.671	2.000	2.390	2.66
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617
∞	0.250	0.674	1.282	1.645	1.960	2.326	2.576



χ^2 -DISTRIBUTION (CHI SQUARE)

	p (probability)						
Degree of freedom	**0.99**	**0.975**	**0.95**	**0.05**	**0.025**	**0.01**	**0.001**
1	0.0000157	0.0000982	0.000393	3.84	5.02	6.63	10.83
2	0.0201	0.0506	0.103	5.99	7.88	9.21	13.82
3	0.115	0.216	0.352	7.81	9.35	11.34	16.27
4	0.297	0.484	0.711	9.49	11.14	13.28	18.47
5	0.554	0.831	1.15	11.07	12.83	15.09	20.51
6	0.872	1.24	1.64	12.59	14.45	16.81	22.46
7	1.24	1.69	2.17	14.07	16.01	18.47	24.32
8	1.65	2.18	2.73	15.51	17.53	20.09	26.13
9	2.09	2.70	3.33	16.92	19.02	21.67	27.88
10	2.56	3.25	3.94	18.31	20.48	23.21	29.59
11	3.05	3.61	4.57	19.68	21.92	24.72	31.26
12	3.57	4.40	5.23	21.03	23.34	26.22	32.91
13	4.11	5.01	5.89	22.36	24.74	27.69	34.53
14	4.66	5.63	6.57	23.68	26.12	29.14	36.12
15	5.23	6.26	7.26	25.00	27.49	30.58	37.70
16	5.81	6.91	7.96	26.33	28.85	32.00	39.25
17	6.41	7.56	8.67	27.59	30.19	33.41	40.79
18	7.01	8.23	9.39	28.87	31.53	34.81	42.31
19	7.63	8.91	10.12	30.14	32.85	36.19	43.82
20	8.26	9.59	10.85	31.41	34.17	37.57	45.31
21	8.90	10.28	11.59	32.67	35.48	38.93	46.80
22	9.54	10.98	12.34	33.92	36.78	40.29	48.27
23	10.20	11.69	13.09	35.17	38.08	41.64	49.73
24	10.86	12.40	13.85	36.42	39.36	42.98	51.18
25	11.52	13.12	14.61	37.65	40.65	44.31	52.62
26	12.20	13.84	15.38	38.89	41.92	45.64	54.05
27	12.88	14.57	16.15	40.11	43.19	46.96	55.48
28	13.56	15.31	16.93	41.34	44.46	48.28	56.89
29	14.26	16.05	17.71	42.56	45.72	49.59	58.30
30	14.95	16.79	18.49	43.77	46.98	50.89	59.70
31	15.66	17.54	19.28	44.99	48.23	52.19	61.10
32	16.36	18.29	20.07	46.19	49.48	53.49	62.49
33	17.07	19.05	20.87	47.40	50.73	54.78	63.87
34	17.79	19.81	21.66	48.60	51.97	56.06	65.25
35	18.51	20.57	22.47	49.80	53.20	57.34	66.62
40	22.16	24.43	26.51	55.76	59.34	63.69	73.40
50	29.71	32.36	34.76	67.51	71.42	76.15	86.66
60	37.48	40.48	43.19	79.08	83.30	88.38	99.61
100	70.06	74.22	77.93	124.3	129.5	135.8	149.4



2. PROBLEMS

1. The lengths of four different frog red blood cells are: 18, 17, 21 and 18 μm. Calculate the mean, the standard deviation and the standard error. ($\bar{x}$ = 18.5 μm, Q_x = 9 μm^2, s_x = 1.73 μm, $s_{\bar{x}}$ = 0.87 μm)
2. The mean of the body heights of 25 students is 170 cm, and the standard deviation is 8 cm. Estimate the expected value at a 95 % confidence level. ($\bar{x} - 2s_{\bar{x}}$ = 166.8 cm, $\bar{x} + 2s_{\bar{x}}$ = 173.2 cm)
3. In the first series of measurements we have 5 measurement points, and in a second one we have 20. A correlation coefficient of 0.6 was calculated from both datasets. Can we claim that the variables are correlated if the level of significance is 5 %? (r = 0.6, n_1 = 5, n_2 = 20, therefore f_1 = 3, f_2 = 18, t_1 = 1.299 does not confirm correlation, but t_2 = 3.18 which indicates correlation even at 1 % significance level)
4. A drug preparation contains 20 % active ingredient (mass ratio). Before commercial introduction of the drug the stability of the preparation should be confirmed. For this stability test, six preparations were stored in the refrigerator for a time period, then the remaining amount of the active ingredient was measured. The result of the analysis showed 20.1, 19.8, 18.9, 19.7, 19.9 and 20.2 % active ingredient content. Is there a significant difference between the measured and nominal values of the active ingredient content of the drug, or the deviation is only by chance? (p = 27 %, not significant)
5. In parallel to the measurement described in the previous problem, there were six samples stored at room temperature. After the same time as in the previous case the active ingredient contents were 19.6, 18.9, 19.5, 20.1, 19.3 and 19.4 %. Should it be prescribed to keep the preparation in the refrigerator? (p = 2.1 %, yes it should be stored in the refrigerator)
6. In a small town 424 individuals were vaccinated prior to an influenza epidemic, while 425 persons in a control group received placebo only. 105 infections occurred later among the vaccinated individuals and 140 in the control group. There were 31 serious cases in the vaccinated group and 55 in the control group. Decide whether the vaccination against influenza was effective or not. (Both the ratio of infections and that of serious infections were significantly smaller among the vaccinated persons: p < 1 %)
7. Successful surgical correction of a special eye disease (non-artherial ischemic optic neuropathia) appeared in a scientific report in 1989. There was not known any successful treatment before, hence this surgical method was applied at many places. However, short time after, some reported that the treatment was not effective. A surway was made in 25 clinical centers involving 244 such patients, 119 of them went through the surgical procedure and 125 was not treated. The data of the surway is in the following table:

	better condition	same condition	worse condition	total
went through the surgery	39	52	28	119
not treated	53	56	16	125

Answer the questions using statistical analysis!

a) is the number of patients in better condition higher after the surgery? (no, p>0.1)
b) is the worse condition more often among those who went through the surgery? (yes, p<0.05)

8. The osteocalcin (OC) level of growth-hormone-deficient children was determined before and after 2 years of growth-hormone therapy. As a control, parallel measurements were done on a group of healthy children. The OC concentration values are given in the table below:

OC concentration (μg/l)

In patients before treatment	In patients after treatment	In healthy children
8.5	14.5	17.3
7.2	10.2	16.9
8.3	12.7	17.7
15.5	16.5	18.6

Two questions were asked in the study:

a) Is the OC level significantly lower in the growth-hormone-deficient children before treatment than in healthy children? (yes, significantly lower)
b) Did the OC level change as a consequence of the treatment? (yes, significantly)

9. Women after menopausa often suffer from calcium defficiency. This might support the observation, that frequency of bone fractures is higher among women of this age. Can we conclude that the decrease of oestrogen level after menopause results calcium defficiency? To investigate this relationship in three different groups of women, a mineral content of bones was measured in g/cm^3 units, that is proportional to the calcium content. First group was from women of 25-50 age, after oophorectomy, and had oestrogen sufficiency prooved. Healthy women of the same age group formed the second group, before menopausa. Women after menopausa formed the third group. There were 14 women in each group. The results are summarized in the following table:

	first group	second group	third group
mean (g/cm^3)	0.93	1.21	0.92
standard deviation (g/cm^3)	0.04	0.03	0.04

How would you answer the questions based on the data, that we have ? (yes there is a connection between menopausa and calcium defficiency)