Basics of Biostatistics

Hypothesis testing

topics

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

Recommended readings:

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

Further readings:

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

- Nature Collection: Statistics for Biologists

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?

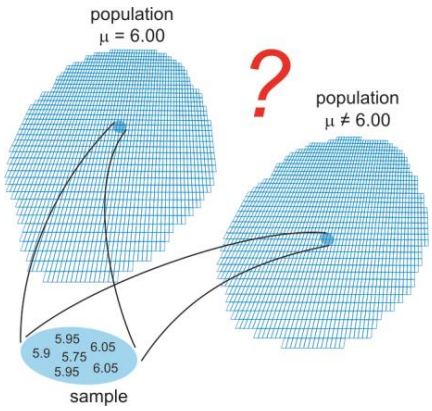

Fig. 17. *Possibilities of the origin of the chosen sample. Values are in mg units.*

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.

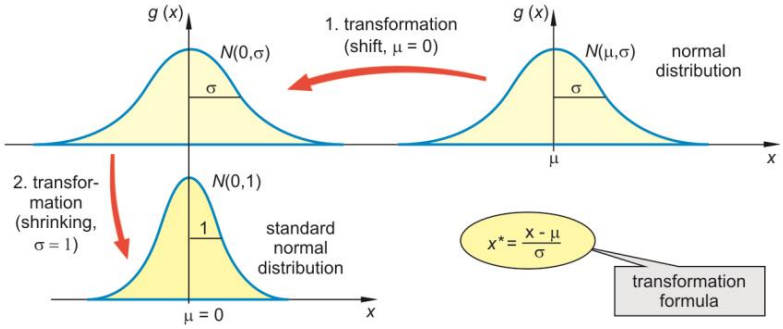

Fig. 18. *Transformation of a normal distribution of general position and width into standard normal distribution ($N(\mu,\sigma) \rightarrow N(0,1)$)*

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.

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.

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.

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.

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

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

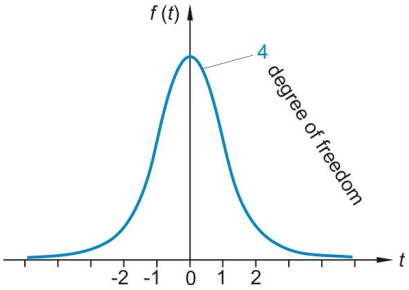

Fig. 19. *Student's t distribution for 4 degrees of freedom. The curve resembles the N(0,1) distribution, but depending on the degree of freedom it differs more or less.*

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

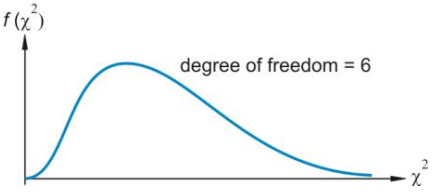

Fig. 20. *The χ^2-distribution of 6 degrees of freedom*

statistical test
statistischer Test
statisztikai próba

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.

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")*?*

t-test
t-Test
t-próba

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).

χ^2-test
χ^2-Test
χ^2-próba

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.

null hypothesis
Nullhypothese
nullhipotézis

An alternative hypothesis $\boldsymbol{H_1}$ must also be specified. $\boldsymbol{H_1}$ is valid if the null hypothesis is rejected. Our first thought might be that defining $\boldsymbol{H_1}$ 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 $\boldsymbol{H_0}$ **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).

alternative hypothesis
Alternativhypothese
alternatív hipotézis

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 $\boldsymbol{H_0}$ **is rejected only if** $\mu - \mu_0 < 0$. This is called the **one-tailed test** (and a directional alternative hypothesis).

two-tailed test
zweiseitiger Test
kétoldalú próba

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

one-tailed test
einseitiger Test
egyoldalú próba

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

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 $\boldsymbol{t = 0}$ **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.

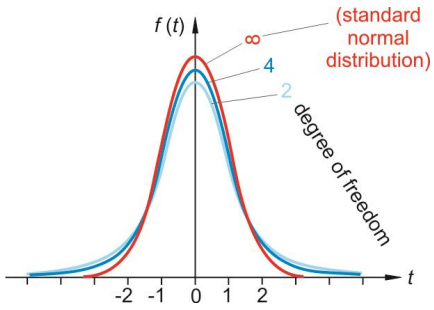

Fig. 21. *The Student's t-distribution for three different degrees of freedom.*

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.

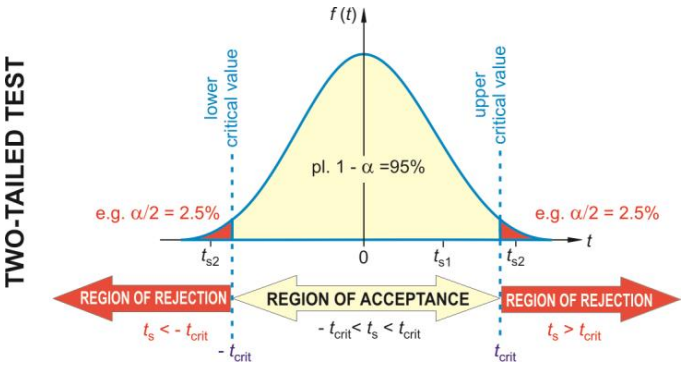

Fig. 22. *Regions of acceptance and rejection in case of a two-tailed t-test.*

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).

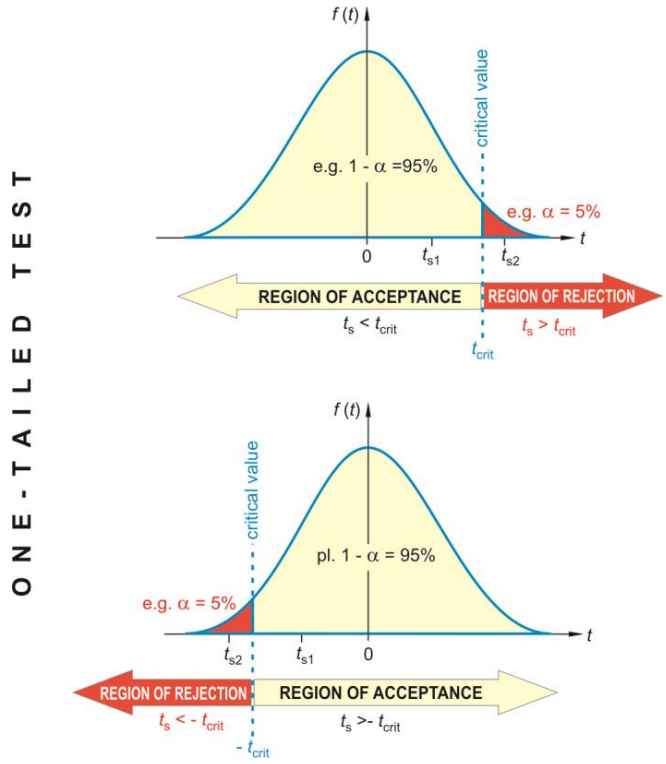

***Fig. 23**. Regions of acceptance and rejection in case of one-tailed t-tests.*

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

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 I error
Fehler 1. Art
elsőfajú 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.

type II error
Fehler 2. Art
másodfajú hiba

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.

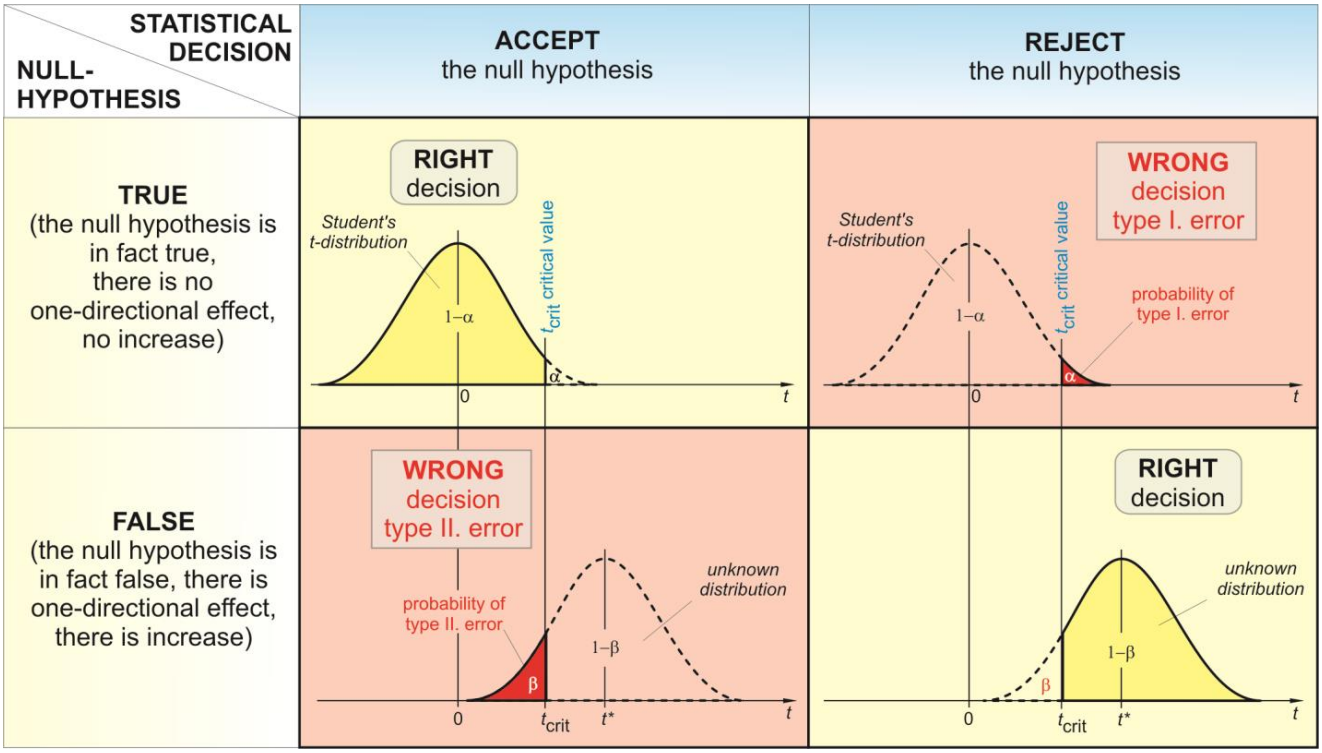

Fig. 24. *Type I and type II error.*

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.)

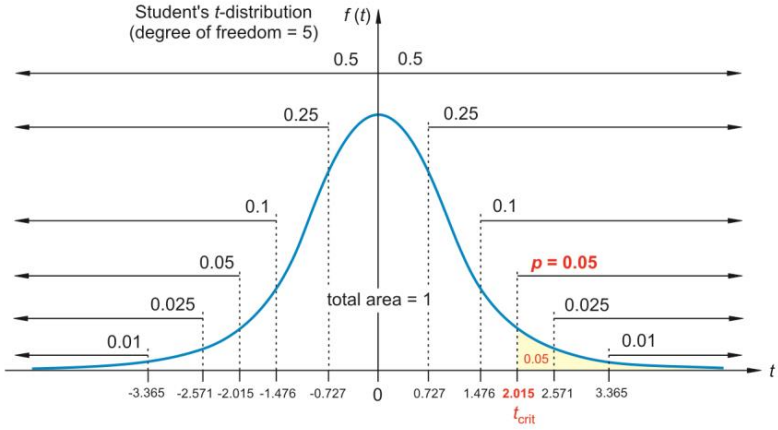

Fig. 25. *Student's t-distribution. Critical values and probabilities that correspond to the one-tailed t-test.*

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: $t_s = \frac{\overline{x_1} - \overline{x_2}}{s^*} \cdot \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$, 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)?

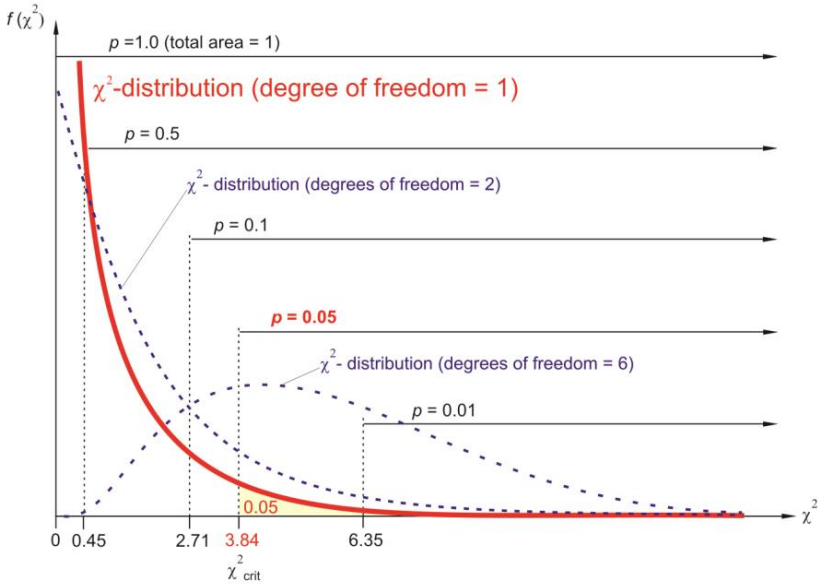

Fig. 26. *The χ^2-distribution for different degrees of freedom, and some critical values of the χ^2-distribution for 1 degree of freedom.*

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