Probability and Statistics

(my email address: borbely.jozsef@uni-obuda.hu)

Probability and some elementary problems

First we will introduce some basic terms.

Relative frequency: Let n be a positive integer. We have the goal to observe a given event. We make n experiments, and we count the number of experiments where our given event appears. Let us denote the number of these experiments by k. Using this labelling, we define the relative frequency of our given event by the fraction k/n.

Probability of a given event: Using the labelling above, we define the probability of given event by the limit of the fractions k/n, where n tends to infinity.

Remark 1: These limit does not always exist. There are events that do not have a probability.

Remark 2: It is an easy consequence of the definition of the probability that the value of the probability has to lie between 0 and 1 (if it exists, of course).

Below you will find some problems on probability that you can solve by simple elementary methods.

Problem 1: In a country the number plates are labelled by five digit numbers from 00000 to 99999. We randomly pick one number plate. What is the probability of the event that

a, there is a six among the digits on the plate

b, the digits on the plate are different

c, three of the digits are the same?

Problem 2: At a soccer training session participate 20 students, two of them are Simon and Garfunkel. We divide the participants into two groups of 10 persons. With which probability will play Simon and Garfunkel against each other?

Problem 3: We put 8 rooks randomly on an empty chess table (the rooks have the same color). What is the probaility of the event that none of them attacks another one?

Problem 4: We want to create a password containing 6 characters. We can randomly choose from six digits and the twenty-six letters of the English alphabet. Generating our password randomly with which probability will it contain exactly three digits?

Random variables

Random variables are variables whose values are randomly chosen.

There are two types of random variables: **discrete** and **continuous** random variables.

The difference between them: discrete random variables take on countable many values, continuous random variables take on uncountable many values.

Every random variable defines a so called **distribution**.

Example (a discrete distribution, unfair die): P(X=1)=1/21, P(X=2)=2/21, P(X=3)=3/21, P(X=4)=4/21, P(X=5)=5/21, P(X=6)=6/21

Main properties of discrete variables :

Let X_1, X_2, ... be the possible values of the discrete variable X.

Then, using the labelling P(X=X_i) = p_i for each possible i, we have

1 = p_1 + p_2+ ...

Complete system of events: A set of events that exclude each other, none of them is the impossible event, and at least one of the occurs.

Remark: It is the same phenomenon that we have seen in part „Main properties of discrete variables". There the events X=X_i determine a complete system of events, where i runs through all of its possible values.

Some well known discrete distributions

Binomial distribution: for a given positive integer n and the probability p we have

P(X=k) = $\binom{n}{k}\cdot p^k\cdot(1-p)^{n-k}$, where k=0, 1, 2, ..., n.

Geometric distribution: for a given probability p we have P(X=k) = $p^{k-1} \cdot (1-p)$, where k runs through the set of the positive integers.

Poisson distribution: P(X=k) = $\frac{\lambda^k}{k!} \cdot e^{-\lambda}$, where λ is a positive parameter, and k runs through the set of the nonnegative integers..

Hipergeometric distribution: for positive integer parameters r ≤ n, d, we have

P(X=k) = $\frac{\binom{r}{k} \cdot \binom{n-r}{d-k}}{\binom{n}{d}}$.

Below are some problems concerning discrete random variables.

Problem 5: In a family every new child will be a girl or a boy with 0,5-0,5 probability. Which probability distribution describes the number of girls in such a family if we take in account the (different) birth dates of the children?

Problem 6: We are tossing a fair coin until we get a head, and then we will stop. With which probability do we have to toss the coin at least 2021 times?

Problem 7: In the Hungarian game „Lotto" you have to choose five different numbers from the set

{1,2, ..., 90}. After you choose your five numbers, exactly five winning numbers of the given 90-element set will be declared. With which probbability will you have

a, exactly two winning numbers

b, at most two winning numbers among your chosen ones?

Problem 8: We are tossing an unfair coin (the probability of throwing a head is p). We count the number of the first identic throws (for example, if we have HHHT..., then this number is three). What is the distribution of this random number?

Problem 9: In a forest the number of seedlings per square meter can be described with a Poisson distribution having parameter 2,5. What is the probability of the event that a sample of one square meter in this forest

a, contains at most one seedling

b, contains more than three seedlings?

Problem 10: We write the letters A, A, A, A, B, L, M randomly in a row. With which probability will we get the word ALABAMA?

Problem 11: We know that ten of hundred given devices are faulty. We choose randomly five ones from the set of the hundred devices in question. With which probability will be at least one of the chosen ones faulty?

Problem 12: Take the set of five digit positive integers that contain each of the digits 1, 2, 3, 4, 5 exactly once. We pick randomly one number from this set. With which probability will be this number divisible by eight?

Problem 13: Determine wheteher the following variables are discrete random variables:

a, P(0)=p^3, P(1)=3p^2(1-p), P(2)=3p$(1-p)^2$, P(3)=$(1-p)^3$

b, P(k)=$\frac{1}{k(k+1)}$, where k runs through the positive integers.

c, P(k)=$\binom{17}{k}\left(\frac{1}{2}\right)^k\left(\frac{1}{3}\right)^{17-k}$, where k =0, 1, 2, ..., 17.

Continuous distributions:

We will generalize the properties of the discrete distributions for uncountable many cases (where we cannot sum up the probabilities).

Instead of summing up, we will integrate.

Density function: Let f be a real, nonnegative, integrable, everywhere defined function. If

$\int_{-\infty}^{\infty} f(x)$dx = 1, then we call f a density function.

Remark: the above condition with the integral means that the area below the curve equals one.

For continuous distributions, the probability will be defined using the density function.

Cumulative distribution function: the cumulative distribution function F(t) is generated by the density function f(x), using the formula F(t)= $\int_{-\infty}^{t} f(x)$dx.

Remark: Here F(t)=P(X≤ t). Moreover, for every two real numbers t and t' we have P(t≤X≤t')= $\int_{t}^{t'} f(x)$dx

Some well known continuous distributions

Uniform distribution: for two fixed real values a < b its density function takes on the value $\frac{1}{b-a}$ on the whole interval (a,b), on every other point of the real line the density function will be zero.

Exponential distribution: for a given positive parameter λ its density function is f(x)=$\lambda \cdot e^{-\lambda x}$ for every nonnegative number x, and the density function is zero otherwise.

Normal distribution: For a given positive parameter σ and a fixed real parameter m, its density function is

f(x) = $\frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{(x-m)^2}{2\sigma^2}}$ for every real number x.

Standard normal distribution (a special case of the normal distribution): here m=0 and σ=1. Its cumulative distribution function will be labelled with Φ.

Here there are some problems for practice:

Problem 14: Let 0<Y<3 be a random variable. Its cumulative distribution function on the given interval is F(x)=cx^3. Determine c and the probability P(-1<Y<1).

Problem 15: Let X be a continuous random variable on the interval [0,c] with the following density function:

f(x)=$\frac{1}{9}x^2$ for every number x on the interval [0,c], otherwise f(x)=0.

Determine the value of c and the cumulative distribution function of X.

Problem 16: We randomly pick a point of the circular disc characterised by the condition x^2+y^2<1.

Let Z be the distance of the chosen point from the center of the circle. Determine the density function and the cumulative distribution function of the random variable Z.

Independence of random variables, operations with random variables

We will need the following terms and definitions to be able to continue the study of properties of random variables.

1. By the **independence** of some random variables we simply mean that the realization of one of them does not affect the probability distribution of the other ones (we do not want to give here more details, in our praxis the independence/dependence will be pretty straightforward).

2. If m(x) is a strictly increasing (and therefore invertable) function, then for a given random variable X we can interprete the random variable m(X) in the following way:

F_m(t) = P(m(x) < t) = P(x < m^{-1}(t)) = F(m^{-1}(t)), where F_m(t) and F(t) are the cumulative distribution functions of m(X) and X.

3. The sum of two continuous random variables X and Y is described by the formula

F_{X+Y} (z) = P(X+Y < z).

Analogously, the sum of two discrete random variables Z and T can be defined using

P(Z+T=w)=$\sum_{r+s=w} P\,(Z = r)$P(T=s), i.e. we sum up regarding the pairs (r,s) whose sum equals w.

The product of random variables can be interpreted similarly.

Expected value, variance and their properties

If the discrete variable X takes the values x_1, x_2, x_3, ... with the probabilities p_1, p_2, p_3,..., then the expected value of X is defined by E(X) = $\sum_{i=1}^{\infty} p_i\, x_i$.

The expected value of a continuous random variable X is defined by E(X)=$\int_{-\infty}^{+\infty} x \cdot f(x)$ dx, where f(x) is the density function of X.

For every random variable X we get the variance by the formula

$D^2(X)$ = E$\left[\left(X - E(X)\right)^2\right]$= E$(X^2)$ – E^2(X).

Here we will give the expected values and the variances of the most important distributions (we use our standard labellings):

The expected value of the binomial distribution is np, its variance is np(1-p).

The expected value of the geometric distribution is $\frac{1}{1-p}$, its variance is $\frac{p}{(1-p)^2}$.

The expected value and the variance of the Poisson distribution is λ.

The expected value of the uniform distribution is$\frac{a+b}{2}$, its variance is $\frac{(b-a)^2}{12}$.

The expected value of the exponential distribution is $\frac{1}{\lambda}$, its variance is $\frac{1}{\lambda^2}$.

The expected value of the normal distribution is m, its variance is σ^2.

Here you can find some problems concerning the expected value and the variance:

Problem 17: We have two fair dies. In all of our throws, we toss our two fair dies simultanously.

We repeat the throw exactly n times. Let us call a throw succesful iff we get at least one six.

What is the expected value of succesful throws among our n throws?

Problem 18: The density function of the random variable X is f(x)=$\frac{c}{x^4}$ for every x>1, and zero otherwise. Determine the value of c and the variance of X.

Problem 19: Let X be a uniform distribution on the interval [1,4]. Determine the expected value of the variable $(X-1)^2$.

Properties of the expected value and the variance

For constants a and b we have D^2(aX+b) = $a^2 \cdot D^2$(X).

If X and Y are random variables, and their sum exists, then E(X+Y)=E(X)+E(Y).

If X and Y are random variables, and their sum exists, then E(XY)=E(X)E(Y).

If X and Y are independent random variables and their sum exists, then D^2(X+Y)=D^2(X)+D^2(Y).

Some properties of the well known random variables

The following facts will be useful in a lot of applications:

1.The sum of independent normal variables is a variable with normal distribution

2. If c and d are real numbers and X is a random variable with normal distribution then (cX+d) is also a random variable with normal distribution

2. The sum of independent Poisson variables is a variable with Poisson distribution

3. The sum of independent binomial variables with the same probability parameter p is a variable with binomial distribution

Remark: using property 2. and our former notations, we can easily check that for every normally distributed random variable X the variable $\frac{X-m}{\sigma}$ is standard normally distributed. We say that by taking the variable $\frac{X-m}{\sigma}$ we standardize the variable X.

Problem 20: The daily milk production of a cow can be described by a normally distributed random variable which has the expected value m=22,1 litres and the variance σ^2=2,25 litres^2. With which probability lies the daily milk production of a given cow between 23 and 25 litres?

Problem 21: In a society the IQ level can be described by a normally distributed random variable with the expected value 110 and the variance 100. If we pick one random person of this society, with which probability will be his IQ level higher than 120?

Problem 22: Let Y and Z be independent, normally distributed random variables, Y has the expected value 2 and the variance 9, Z has the expected value 4 and the variance 16. Determine P(1≤Y<3) and P$\left(\frac{Y-Z}{2} > 0\right)$.

Statistical functions

We defined operations with random variables, thus we are able to interprete **statistical functions**: let X_1, X_2, ..., X_n be random variables. For an elementary function T with n variables we can define T$(X_1, X_2, \ldots, X_n)$.

Unbiased estimators

Among statistical functions, the so called **unbiased estimators** play a key role. An unbiased estimator of a given parameter p is the statistical function T$(X_1, X_2, \ldots, X_n)$, if for every possible value of the parameter p we have $E_p\big(\mathrm{T}(X_1, X_2, \ldots, X_n)\big)$ =p.

Let us see an example: the sample mean is an unbiased estimator of the expected value. To see this let us define the independent random variables $X_1, X_2, \ldots, X_n$ of the same distribution. Using the additive properties of the expected value, we can easily see that E$\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right)$ equals the common expected value of the variables $X_1, X_2, \ldots, X_n$, which proves our claim.

An unbiased estimator of the variance

Let us consider the independent random variables $X_1, X_2, \dots, X_n$ of the same distribution. Using the fact that the sample mean $\overline{\mathrm{X}} = \frac{X_1 + X_2 + \dots + X_n}{n}$ statistical function is an unbiased estimator of the common expected value of the random variables $X_1, X_2, \dots, X_n$, we can find an unbiased estimator for their variance σ^2:

$$\mathrm{E}\left(\sum_{i=1}^{n} (\mathrm{X_i} - \overline{\mathrm{X}})^2\right) = \sum_{i=1}^{n} \mathrm{E}((\mathrm{X_i} - \overline{\mathrm{X}})^2) = \sum_{i=1}^{n} \mathrm{E}\left(\mathrm{X_i}^2 - 2\, \overline{\mathrm{X}} \cdot \mathrm{X_i} + \overline{\mathrm{X}}^2\right) =$$

$$= \sum_{i=1}^{n} \mathrm{E}\left(\mathrm{X_i}^2\right) + \mathrm{n} \cdot \mathrm{E}(\overline{\mathrm{X}}^2) - 2 \cdot \sum_{i=1}^{n} \mathrm{E}(\overline{\mathrm{X}} \cdot \mathrm{X_i}) =$$

$$= \sum_{i=1}^{n} \mathrm{E}\left(\mathrm{X_i}^2\right) + \mathrm{n} \cdot \mathrm{E}\left(\left(\frac{X_1 + X_2 + \dots + X_n}{\mathrm{n}}\right)^2\right) - 2 \cdot \sum_{i=1}^{n} \mathrm{E}\left(\frac{X_1 + X_2 + \dots + X_n}{n} \cdot \mathrm{X_i}\right) =$$

$$= \sum_{i=1}^{n} \mathrm{E}\left(\mathrm{X_i}^2\right) + \frac{1}{n} \cdot \mathrm{E}((X_1 + X_2 + \dots + X_n)^2) - \frac{2}{n} \cdot \sum_{i=1}^{n} \mathrm{E}((X_1 + X_2 + \dots + X_n) \cdot \mathrm{X_i}) =$$

$$= \sum_{i=1}^{n} \mathrm{E}\left(\mathrm{X_i}^2\right) + \frac{1}{n} \cdot \mathrm{E}\left(\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{X_i}^2 + 2 \cdot \sum_{\mathrm{i} \neq \mathrm{j}} \mathrm{X_i X_j}\right) - \frac{4}{n} \cdot \mathrm{E}\left(\sum_{\mathrm{i} \neq \mathrm{j}} \mathrm{X_i X_j}\right) - \frac{2}{n} \cdot \sum_{i=1}^{n} \mathrm{E}\left(\mathrm{X_i}^2\right) =$$

$$= \sum_{i=1}^{n} \mathrm{E}\left(\mathrm{X_i}^2\right) + \frac{1}{n} \cdot \mathrm{E}\left(\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{X_i}^2\right) + \frac{2}{n} \cdot \mathrm{E}\left(\sum_{\mathrm{i} \neq \mathrm{j}} \mathrm{X_i X_j}\right) - \frac{4}{n} \cdot \mathrm{E}\left(\sum_{\mathrm{i} \neq \mathrm{j}} \mathrm{X_i X_j}\right) - \frac{2}{n} \cdot \sum_{i=1}^{n} \mathrm{E}\left(\mathrm{X_i}^2\right) =$$

$$= \frac{n-1}{n} \cdot \sum_{i=1}^{n} \mathrm{E}\left(\mathrm{X_i}^2\right) - \frac{2}{n} \cdot \mathrm{E}\left(\sum_{\mathrm{i} \neq \mathrm{j}} \mathrm{X_i X_j}\right) = \frac{n-1}{n} \cdot \sum_{i=1}^{n} \mathrm{E}\left(\mathrm{X_i}^2\right) - \frac{2}{n} \cdot \binom{n}{2} \cdot m^2 =$$

$$= (n-1) \cdot \mathrm{E}\left(\mathrm{X_1}^2\right) - (n-1) \cdot m^2 = (n-1) \cdot \left(\mathrm{E}\left(\mathrm{X_1}^2\right) - \mathrm{E}^2(\mathrm{X_1})\right) = (n-1) \cdot \sigma^2.$$

Thus $\frac{\sum_{i=1}^{n} (\mathrm{X_i} - \overline{\mathrm{X}})^2}{n-1}$ is an unbiased estimator of the variance σ^2.

Method of moments

Let X be a random variable. For any positive integer number k, let $\mathrm{E}\left(\mathrm{X}^{\mathrm{k}}\right)$ be **the k-th moment of the variable X.**

If $X_1, X_2, \dots, X_n$ are random variables of the same distribution, then for every positive

integer k the statistical function is an unbiased estimator $\frac{{X_1}^k + {X_2}^k + \cdots + {X_n}^k}{n}$ of the common k-

th moment of the variables $X_1, X_2, \dots, X_n$.

Using this observation, we can estimate the unknown parameters of some distributions with the help of our samples.

Example:

Let X be a random variable of normal distribution with unknown parameters. We have a sample of five elements of this distribution: 11, 12, 13, 14, 20.

We will use the method of moments to estimate the two unknown parameters.

E(X)=m, E(X^2) = σ^2 + m^2

Using our data, we get

$\frac{11+12+13+14+20}{5}$ = m

and

$\frac{11^2+12^2+13^2+14^2+20^2}{5}$ = σ^2 + m^2.

The solutions will be m=14 and σ = $\sqrt{10}$.

Joint density function

Joint density functions are functions h of n variables which take on only nonnegative values and fulfill the property $\int_{-\infty}^{\infty}$ ... $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}$ h$(t_1, t_2, \dots, t_n)$ $dt_1 dt_2$ ... dt_n= 1.

The term joint density function is a natural expansion of the traditional density function.

Using joint density functions, we will be able to use the maximum likelihood method.

Maximum likelihood method

We have the samples x_1, x_2, ..., x_n from a given probability distribution. If we do not know a parameter λ of this distribution, we can estimate it in the following way: we take the function $f_\lambda(x_1, x_2, \dots, x_n)$ which depends only on λ, and we determine its maximum point. In this case it is worth to take the logarithm of $f_\lambda(x_1, x_2, \dots, x_n)$, because generally it will be easier to take the derivative of this function (we derivate by λ).

We call $f_\lambda(x_1, x_2, \dots, x_n)$ the maximum likelihood function, and we call ln$f_\lambda(x_1, x_2, \dots, x_n)$ the log likelihood function.

Example:

We have a normal distribution with a given variance σ^2 and an unknown expected value m. Let $x_1, x_2, \dots, x_n$ be independent samples from this distribution. Let us define the function

$f_m(x_1, x_2, \ldots, x_n)$ = $\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}}$·exp$\left(-\frac{1}{2}\cdot\left(\frac{x_i-m}{\sigma}\right)^2\right)$, and the log likelihood function

ln$\left[\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}}\cdot \exp\left(-\frac{1}{2}\cdot\left(\frac{x-m}{\sigma}\right)^2\right)\right]$ = n·ln$\frac{1}{\sigma\sqrt{2\pi}}$ + $\sum_{i=1}^{n} -\frac{1}{2}\cdot\left(\frac{x_i-m}{\sigma}\right)^2$.

We derivate by m:

$\sum_{i=1}^{n} -\frac{1}{2}\cdot\frac{2\cdot(x_i-m)\cdot(-1)}{\sigma^2} = \sum_{i=1}^{n}\frac{(x_i-m)}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - m).$

We get the condition $\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - m)$ = 0, which means that $\sum_{i=1}^{n}(x_i - m)$=0,

thus $\frac{x_1 + x_2 + \ldots + x_n}{n}$ = m. It is easy to see that this is a maximum point.

Thus using the maximum likelihood method we can estimate the first parameter of a normal distribution by the sample mean.

Problem 23: Let $X_1, X_2, \ldots, X_n$ be independent random variables of the same distribution. Let us denote its common unknown expected value by m. Which of the following statistical functions are unbiased estimnators of m?

a, T($X_1, X_2, \ldots, X_n$)=X_8

b, T($X_1, X_2, \ldots, X_n$)=$\frac{X_8+X_{18}}{9}$

c, T($X_1, X_2, \ldots, X_n$)=$\bar{X}$v

Problem 24: Give an unbiased estimator for the unknown parameter v of the uniform distribution on the interval (0,v) using the sample mean.

Problem 25: The possible values of a discrete random variable are -1, 1, 2, which can be taken by the probabilities c, 3c, 1-4c, respectively. Give an estimate for the value of c using the maximum likelihood method and the method of moments as well.

Problem 26: We have a sample with n elements from the uniform distribution on the interval

(a, b), where a and b are unknown parameters. Estimate a and b using the moment of methods.

Problem 27: We have a sample with n independent elements from a binomial distribution having an unknown probability parameter p (the other parameter is given). Using the maximum likelihood method, give an estimate for p.

Problem 28: We have a sample with n independent elements from an exponential distribution having an unknown parameter. Using the maximum likelihood method, give an estimate for this unknown parameter.

Interval estimates

The estimates we have seen so far are so-called were pointwise estimates, since we approximated the value of the unknown parameter with a single value. Now, however, we will use estimates that are quite likely close to the value of the unknown parameter in a closed interval (hence we call them interval estimates).

We will use so-called statistical tests. We proceed from a given sample of finitely many independent elements, where the elements are taken from the same distribution. We want to estimate one (or more) unknown parameters of the distribution. We will have a null hypothesis concerning the value of the unknown parameter(s). If the null hypothesis is satisfied with a sufficiently large, fixed probability, we accept the statement of the null hypothesis (in practice, we usually expect a probability of 95%). If the null hypothesis is fulfilled with less probability than expected, then the null hypothesis will be rejected at the given level (depending on the value of the fixed probability).

Of course, this is not an exact proof, but we can only tell with a good chance whether the null hypothesis is satisfied or not.

The Z-test

Let us take n independent samples from the same probability distribution, which we know to have a normal distribution, and even their standard deviation (the square root of the variance) is known (let it be σ).

We want to estimate the first parameter (the expected value) based on the sample. Denote this unknown parameter by m.

Let our null hypothesis be that the value of m is equal to the constant m_0. The significance level of the test should be α, which means that we want to ensure with 1-α probability that the null hypothesis is true. In other words, this is also expressed as deciding at the (1-α) confidence level.

Let X_1, X_2, ..., X_n be independent probability variables that are normally distributed with the parameters m and σ (so we took the sample elements from these ones).

Construct the statistical function T (X_1, X_2, ..., X_n) = $\sqrt{n}\cdot \frac{\bar{X}-m_0}{\sigma}$

Due to our result on the sum of normal distributions, if the null hypothesis is fulfilled, then $\sqrt{n}\cdot \frac{\bar{X}-m_0}{\sigma}$ has a standard normal distribution, i.e. it follows a normal distribution with an expected value of 0 and a standard deviation of 1. By convention, we perform the estimation “symmetrically”, i.e. we take an interval that is symmetric to zero and occupies just α units under the density function of the standard normal distribution.

So we look for a positive value of r for which

P$\left(-r < \sqrt{n} \cdot \frac{\bar{X}-m_0}{\sigma} < r\right)$ = 1-α, an this is equivalent to looking for an r for which the standard normal distribution assumes a value between (-r) and r with exactly (1-α) probability.

Denoting the distribution function of the standard normal distribution by Φ, this means that ($\Phi(r)$ - $\Phi(-r)$) must be equal to (1-α). Since the density function of the standard normal distribution is axially symmetric to the y-axis, $\Phi(0)$ = ½, and is located in the middle between $\Phi(r)$ and $\Phi(-r)$. It follows that

$\Phi(r)$ - $\frac{1}{2}$ = $\frac{1-\alpha}{2}$, hence $\Phi(r)$ = 1 - $\frac{\alpha}{2}$.

Φ is strictly monotonic, thus it can be inverted. So r = $\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$. This value is also denoted by $u_{1-\frac{\alpha}{2}}$.

Thus, at a given level of significance α, we accept the null hypothesis if, substituting the sample elements into the statistical function

T(X_1, X_2, ..., X_n) = $\sqrt{n}\cdot \frac{\bar{X}-m_0}{\sigma}$, the value obtained (called test statistics in this case) lies between -$u_{1-\frac{\alpha}{2}}$ és $u_{1-\frac{\alpha}{2}}$.

Otherwise, we reject the null hypothesis and we accept its denial, the so-called counter-hypothesis (although we cannot know certainly which one is true!).

The interval in question is called confidence interval, while the set of real numbers that do not belong to the confidence interval is called critical range.$\left(-u_{1-\frac{\alpha}{2}}, u_{1-\frac{\alpha}{2}}\right)$

Note that if the null hypothesis had been that m≤m_0 then it can be calculated in the same way that the critical range is a set of numbers greater than $u_{1-\alpha}$ (for the same statistical function).

Similarly, if the null hypothesis is m≥m_0, then the critical range is the set of numbers less than u_{α} (for the same statistical function).

Example of using the Z-test:

The salaries of the leaders of a party (expressed in HUF million) can be well approximated by a normal distribution with the unknown expected value μ_1 and the standard deviation 2. We found out the salaries of 12 party members (also expressed in HUF million):

$$20.47, 21.10, 18.67, 16.67, 18.00, 20.40, 22.17, 20.05, 24.85, 19.93, 19.73, 20.39$$

Let us decide whether hypothesis μ_1= 20 is acceptable at a confidence level of 1-α = 0.95.

The standard deviation (which is 2) is known, so a one-sample Z-test is carried out. Let be α = 0.05. The sample is

20.47, 4.56, 21.10, 6.67, 18.67, 4.10, 16.67, 11.91, 18.00, 3.89, 20.40, 5.48,

hence the sample mean is

$$\frac{20.47+21.10+18.67+16.67+18.00+20.40+22.17+20.05+24.85+19.93+19.73+20.39}{12} = 20.2025,$$

and the null hypothesis is that the first parameter μ_1 is equal to 20.

We establish the test statistics:

u = $\frac{20.2025-20}{\frac{2}{\sqrt{12}}}$ = 0.2025·$\sqrt{3}$, the critical range is the set of u's with an absolute value greater than $u_{0,975}$.

1,96 = $u_{0,975}$ > 0.2025·$\sqrt{3}$,, so we accept the null hypothesis at the given level.

Student's t-test

The basic problem is similar to that of the Z-test, the only difference being that the standard deviation of the normal distribution is not known, and thus the expected value must be estimated based on n independent samples (we have to decide whether the null hypothesis is acceptable).

Let us first have the null hypothesis that the value of m is equal to the constant m_0. The significance level of the test is α, which means that we want to ensure with 1-α probability that the null hypothesis is true.

We used the standard deviation in the statistical function for the Z-test, so we cannot use exactly that function. In such cases, it is a good idea to replace the unknown value there (i.e. the standard deviation) with a value that depends only on the sample, which gives a good point estimate for it.

We have previously seen that $\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$ is an unbiased estimator for the variance, so we will use a statistical function

$$T(X_1, X_2, ..., X_n) = \sqrt{n}\cdot \frac{\bar{X}-m_0}{\sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}}}.$$

If the null hypothesis is satisfied then the variable $\sqrt{n}\cdot\frac{\bar{X}-m_0}{\sqrt{\frac{\sum_{i=1}^{n}(\mathrm{x_i}-\bar{\mathrm{x}})^2}{n-1}}}$ follows a so-called Student t-distribution of degree (n-1). All that is used is that its density function (as well as the standard normal distribution) is axially symmetric to the y-axis and has a positive value everywhere. Denote by S_{n-1} the distribution function of the t-distribution of degree

(n-1).

We are looking for a positive number r for which

P = 1-α.

$$P\left(-r < \sqrt{n}\cdot\frac{\bar{X}-m_0}{\sqrt{\frac{\sum_{i=1}^{n}(\mathrm{x_i}-\bar{\mathrm{x}})^2}{n-1}}} < r\right) = 1\text{-}\alpha.$$

Using the same ideas as in the case of the Z-test we get

r = ${S_{n-1}}^{-1}\left(1-\frac{\alpha}{2}\right)$, which is also denoted by $t_{n-1,1-\frac{\alpha}{2}}$.

So the confidence interval is here $\left(-{S_{n-1}}^{-1}\left(1-\frac{\alpha}{2}\right), {S_{n-1}}^{-1}\left(1-\frac{\alpha}{2}\right)\right)$, which is the same as $\left(-t_{n-1,1-\frac{\alpha}{2}}, t_{n-1,1-\frac{\alpha}{2}}\right)$, and the critical range consists of the real numbers outside lying this interval.

If the null hypothesis were m≤m_0, we would have to choose a set of real numbers greater than $t_{n-1,1-\alpha}$.

If the null hypothesis were m≥m_0, then we would have to choose a set of real numbers less than $t_{n-1,\alpha}$.

Example of using the t-test

A company selling orange drinks has been reported because, according to customer complaints, their soft drinks sold as 500 ml contain less beverages than indicated. We have a sample of ten soft drinks in which the amounts of soft drinks were the following ones (expressed in ml):

483, 502, 498, 496, 502, 483, 494, 491, 505, 486

Decide whether the hypothesis that the expected volume of soft drink in the vials is 500 ml is acceptable (i.e. μ = 500 hypothesis). We can assume normal distribution concerning the amount of soft drink in the vials, we have to decide with a confidence level of 1-α = 0,95.

We assume a normal distribution, but now the standard deviation is unknown. In this case, a t-test is used. The null hypothesis is that the first parameter of the distribution (which is its expected value) is equal to 500, i.e. μ = 500. Let α = 0.05.

The number of elements in the sample is n = 10, the sample mean is 494, the corrected empirical standard deviation

$\frac{11^2+8^2+4^2+2^2+8^2+11^2+0^2+3^2+11^2+8^2}{9}$ = 64,888 ....

The test statistics is

t = $\sqrt{10}. \frac{494-500}{\frac{\sqrt{58,4}}{\sqrt{9}}}$, and the critical range is a set of numbers with absolute values greater than

$t_{9;\ 975/1000}$.

In the present case, t falls in the critical range (in the negative direction), so the null hypothesis must be rejected at this level. This means that it is acceptable to assume that the manufacturer wants to mislead the customers.

Two-sample Z-test

Let us consider an n-element, independent sample taken from a normal distribution with unknown (say m_1) expected value but known (say ${\sigma_1}^2$) variance.

Independently of the above, let us give an m-element, independent sample derived from a normal distribution of unknown (say m_2) expected value but known (say ${\sigma_2}^2$) variance. At a given 1-α confidence level, we want to decide whether it can be assumed that

m_1=m_2 (so this will be the null hypothesis).

In this case we will use the statistical function $\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{{\sigma_1}^2}{n}+\frac{{\sigma_2}^2}{m}}}$.

If m_1=m_2, then $\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{{\sigma_1}^2}{n}+\frac{{\sigma_2}^2}{m}}}$ follows the standard normal distribution, so we can proceed in the

same way as for the original, one-sample Z-test.

Two-sample t-test

We have the same basic problem as for the two-sample Z-test, except that the standard deviations are assumed to be unknown but identical.

In this case we use a statistical function

$$\sqrt{\frac{mn}{m+n}}\cdot\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{(n-1)\cdot\frac{\sum_{i=1}^{n}(\mathrm{X_i}-\bar{\mathrm{X}})^2}{n-1}+(m-1)\cdot\frac{\sum_{i=1}^{m}(\mathrm{Y_i}-\bar{\mathrm{Y}})^2}{m-1}}{n+m-2}}}$$

(essentially the variances are replaced by their unbiased estimators). The resulting variable follows a t-distribution of degree (n + m-2), so we can proceed so as we have seen during the t-test.

Welch test

Our conditions are the same as for the two-sample t-test, but we now assume that the unknown standard deviations are different. In this case we use the statistical function

$$\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\frac{\sum_{i=1}^{n}(\mathrm{X_i}-\bar{\mathrm{X}})^2}{n-1}}{n}+\frac{\frac{\sum_{i=1}^{m}(\mathrm{Y_i}-\bar{\mathrm{Y}})^2}{m-1}}{m}}}$$

The corresponding probability variable then follows a t-distribution of degree f, where f is the integer value of

$$\frac{\left(\sqrt{\frac{\frac{\sum_{i=1}^{n}(\mathrm{X_i}-\bar{\mathrm{X}})^2}{n-1}}{n}+\frac{\frac{\sum_{i=1}^{m}(\mathrm{Y_i}-\bar{\mathrm{Y}})^2}{m-1}}{m}}\right)^2}{\frac{\left(\frac{\frac{\sum_{i=1}^{n}(\mathrm{X_i}-\bar{\mathrm{X}})^2}{n-1}}{n}\right)^2}{n-1}+\frac{\left(\frac{\frac{\sum_{i=1}^{m}(\mathrm{Y_i}-\bar{\mathrm{Y}})^2}{m-1}}{m}\right)^2}{m-1}}.$$

F-test

The two-sample t-test and the Welch test are very similar to each other: the only difference between them is that the former one was used for normal distributions with the same standard deviation, while the latter one was used when the two unknown standard deviations were different. But how could we decide which of the two we should apply if we know nothing about the standard deviations? This is what the F-test is for: we want to decide at the (1-α) confidence level whether the standard deviation of the two distributions can be assumed to be the same.

Calculate the corrected empirical standard deviation of the n- and m-element samples taken from samples X and Y, respectively. We can assume that the sample set for X is at least as large as that for the set Y, i.e.

$\frac{\sum_{i=1}^{n}(\mathrm{X_i}-\bar{\mathrm{X}})^2}{n-1} \geq \frac{\sum_{i=1}^{m}(\mathrm{Y_i}-\bar{\mathrm{Y}})^2}{m-1}$, which is equivalent to $\frac{\frac{\sum_{i=1}^{n}(\mathrm{X_i}-\bar{\mathrm{X}})^2}{n-1}}{\frac{\sum_{i=1}^{m}(\mathrm{Y_i}-\bar{\mathrm{Y}})^2}{m-1}} \geq 1$.

The variable $\frac{\frac{\sum_{i=1}^{n}(\mathrm{X_i}-\bar{\mathrm{X}})^2}{n-1}}{\frac{\sum_{i=1}^{m}(\mathrm{Y_i}-\bar{\mathrm{Y}})^2}{m-1}}$ is called an F-distribution with parameters (n-1) and (m-1). Here we have to pay attention that the roles of n and m are not interchangeable. The cumulative distribution function of the corresponding F-distribution is denoted by $F_{n-1,m-1}$.

We accept that the standard deviations equal to each other if the value of the test statistics (which is at least 1) is equal to at most ${F_{n-1,m-1}}^{-1}$ (1- $\frac{\alpha}{2}$).

The specific value of this must also be found in a table, with a given confidence level and given n and m values.

In this way, it can be clearly decided whether we have to use a two-sample t-test or a Welch test (after the F-test).

Two paired, non-independent, normally distributed samples

Let be given two sets of samples of the same size, in which there are independent samples that can be approximated by a normal distribution, but in each sample the elements with the corresponding sequence number are no longer independent. If the standard deviations are not known, but we want to decide what relation the expected values are in, we can apply the traditional t-test to the set of differences obtained from the samples.

Example:

A farmer's cows consume only grass. Someone tells him that if he fed the cows with silage, the fat percentage of their milk would increase. The farmer tried this diet on six cows for a month, and their fat percentage changed as follows:

3.84 → 3.90, 3.79 → 4.05, 3.78 → 3.80, 4.00 → 4.01, 3.83 → 3.81, 3.84 → 3.90.

We will assume a normal distribution concerning fat percentages.

Based on the data, we try to decide whether we can assume that feeding with silage increases the fat percentage of milk.

We have two sets of samples, each with 6 samples, but each sample is not independent because we always get two values for each cow. The standard deviation of the assumed normal distribution is unknown, so a one-sample t-test should be applied to the differences.

For grass-fed cows, the expected value of the normal distribution (first parameter) will be denoted by μ_g, and for cows fed also by silage by μ_s. The null hypothesis is that μ_s-μ_g≥ 0, as this expresses that if cows are also fed by silage, their milk fat percentage will increase.

Subtract the old data from the data after feeding the silage and then take the average of the differences (this is the average of the differences between the two samples):

$\frac{0.06+0.26+0.02+0.01-0.02+0.06}{6}$= 0.065. The null hypothesis is now of type $\mu \geq \mu_0$, so its critical range is a set of values less than $t_{5;\,5/100}$. The corrected empirical standard deviation

$\frac{0.05^2+0.195^2+0.45^2+0.55^2+0.45^2+0.005^2}{5}$ = 0.14961.

The test statistics is

t = $\frac{0,065-0}{\frac{\sqrt{0,14961}}{\sqrt{6}}} = \frac{0,065}{\sqrt{0,024935}}$ = 0,411..., which lies in the critical range. So at that level, we reject the hypothesis that silage increases the fat percentage.

Problem 29: We want to analyse the change of the average temperature in Budapest on the 15th October. In the last four years we had the following averages (in grade Celsius): 14.8, 12.2, 16.8, 11.1.

We assume that our data have been taken from a normal distribution. Decide whether we can accept the hypothesis that the average temperature lies under 15 grade Celsius on the 15th October.

Problem 30: We have the following samples from factories A and B concerning the number of faulty products (given per mille):

A: 11.9, 12.1, 12.8, 12.2, 12.5, 11.9, 12.5, 11.8, 12.4, 12.9

B: 12.1, 12.0, 12.9, 12.2, 12.7, 12.6, 12.6, 12.8, 12.0, 13.1

Can we assume that factory A is more effective than factory B (we can assume that our samples are independent and come from normal distributions)?

Problem 31: We want to compare two servers with each other. On the first one the average speed of 30 runnings was 6.7 sec, and on the other one that of 20 runnings was 7.2 sec. Decide whether there is a significant difference between the speeds of the two servers, if the standard deviation of both computers equals 0.5 sec?

Problem 32: We measured the toxic gas emission on the same ten places (independently) in November and in December. We got the following results:

November: 20.9, 17.1, 15.8, 18.8, 20.1, 15.6, 14.8, 24.1, 18.9, 12.5

December: 21.1, 16.7, 16.4, 19.2, 19.9, 16.6, 15.0, 24.0, 19.2, 13.2

Was the change of the toxic gas emission significant?